Neural Computing and Applications
https://doi.org/10.1007/s00521-025-10999-9

ORIGINAL ARTICLE

Machine learning to predict gut microbiomes of agricultural pests

Md Jobayer[1] · Alexander Taylor[2] · Md Rakibul Hasan[1,3] · Khandaker Asif Ahmed[4] · Md Zakir Hossain[2,3]

Received: 15 June 2024 / Accepted: 10 January 2025

Abstract
While current efforts to control agricultural insect pests largely focus on the widespread use of insecticides, predicting microbiome composition can provide important data for creating more efficient and long-lasting pest control methods by analysing the pest's food-digesting capacity and resistance to bacteria or viruses. We aim to develop a machine learning model to predict the microbiome composition in agricultural pests and investigate the dynamics of these microbiome compositions using metagenomic samples taken from fruit flies. In this paper, we propose three machine learning-based biological models. Firstly, we propose an intrafamilial model that predicts the relative abundance of bacterial families within themselves using their past generations. Next, we propose two interfamilial models following quantitative and qualitative approaches. The quantitative model predicts the number of bacterial families in a given sample based on the presence of other families in that sample. The qualitative model predicts the relative abundance using binary information of all bacterial families. All three models were tested against least angle regression, random forest, elastic-net, and Lasso. The third approach exhibits promising results by applying a random forest with the lowest mean coefficient of variance of 1.25. The overall results of this study highlight how complex these dynamic systems are and demonstrate that more computationally efficient methods can characterise them quickly. The results of this study are intended to be used as a tool to identify vital taxological families, genera and species of the potential microbiome for better pest control.

Keywords Gut microbiome · Agricultural pests · Biological model · Machine learning

1 Introduction

As the global population continues to grow and food security issues increase in importance, future efforts to improve productivity and, ultimately, the yield of crops must also intensify [55]. A critical element of this effort is the management of and response to biophysical threats that can potentially cause devastating crop losses [47]. Among crops, one of the most significant threats is insect pests, which reduce the amount of helpful produce and account for losses ranging from 10% to 40% worldwide [51]. Current efforts to reduce the impact that these insect pests have on crops are typically focused on the widespread use of insecticides [42] due to their relatively low cost, their ease of application to crops, and the diversity of insect pests they can control [3]. Despite these advantages, the unchecked application of insecticides presents a wide range of consequential issues, such as the non-specific killing of beneficial insects [21], the emergence of resistance within target pests [61], heavy environmental pollution [67] and, in some cases, negative impacts on human health [23]. As a result, many emerging management strategies have started to focus on using more targeted, cheaper insecticide alternatives that can sustain higher efficacy over much longer time frames [17].

Some common alternatives include the integration of transgenic genes into plants [50], the introduction of beneficial predators and parasites [11], the attraction and trapping of insect pests using pheromones [26], and the application of natural, horticultural oils [39]. While these alternatives have been proven to be efficacious in varying

✉ Md Jobayer
md.jobayer@bracu.ac.bd

1 BSRM School of Engineering, BRAC University, Dhaka 1212, Bangladesh

2 School of Engineering, Australian National University, Canberra, ACT 2600, Australia

3 School of Electrical Engineering, Computing and Mathematical Sciences, Curtin University, Perth, WA 6102, Australia

4 CSIRO Australian Centre for Disease Preparedness, Geelong, VIC, Australia

Published online: 21 February 2025

degrees, various financial, societal and environmental drawbacks limit their use to specific applications [2]. One relatively novel and interesting alternate strategy is the development of a highly specific bio-agent (i.e. microbiome) that can modulate pest behaviour and has the potential to be utilised as biopesticides. Microbiome composition can influence insect control by influencing the insect's capacity to digest food, resist bacteria and viruses, and interact with the environment [5, 24]. Researchers can learn more about how they might be able to modify the microbiome to control the behaviour or population of the insect by predicting the microbiome composition of a particular insect species [48]. Studies investigating the microbes present in gut samples usually originate from flies caught in the wild or reared in a laboratory. In either case, metagenomic samples are typically collected from whole flies or dissected guts of surface-sterilised flies. After DNA extraction and amplification of bacterial-specific sequences, the resulting samples were sequenced in a high-throughput sequencing platform. Within the literature, it is apparent that the most common bacterial families present in the gut microbiome of wild fruit flies are *Enterobacterales* and *Lactobacillales*. These bacterial families' ubiquity is mainly due to the probiotic effects they confer on the host flies [25, 32]. An interesting study from [6] demonstrated a strong link between the microbiome and developing larvae, indicating that a reduction in the amount of *Enterobacterales* present can significantly increase the larval period. These observations were followed up by [46] where they noted that a reduction in the amount of *Acetobacter* also increased the larval period, but also significantly reduced the metabolic rate of flies, altered their carbohydrate allocation and increased glucose levels within the blood.

The gut microbiome can be predicted using various methods, including abundance-dependent methods such as analysis of composition of microbiomes (ANCOM) [35] and distance-based methods like Similarity Percentage Analysis (SIMPER) [49]. However, these methods have their limitations. For instance, ANCOM performs poorly when the abundance of Operational Taxonomic Units (OTUs) exceeds 25% or is very low [28, 35]. On the other hand, SIMPER only allows pairwise comparison, which can lead to misleading statistical output due to variance in the number of OTUs [35]. Additionally, mathematical models [22] are highly specific to the host type of the microbiome [27]. For example, if a model is built for a human host-based microbiome, we need to make a hypothesis and test it in-vitro and in species other than humans to make it a generic model. High dimensionality also poses a problem by increasing computational complexity. In this paper, we implement models that overcome the limitations of context-specific models and high dimensionality by focusing on the most valuable features of genomic data.

While metagenomic tools can accurately identify the exact composition of a small number of gut microbiomes, the computational complexity required to process the large datasets that describe vast numbers of samples is still too high [59]. One potential method of characterising and predicting the composition of a large number of gut microbiomes in a more computationally affordable yet coarse-grained way involves the use of machine learning (ML) techniques to make several simplifying assumptions [20] due to their ability to identify patterns in large sets of data without being explicitly programmed to do so [20]. From these patterns, simplified assumptions can be made to construct models that can estimate the microbiome of a sample based on several input parameters. ML models have become increasingly popular for predicting gut microbiome-based outcomes, evidenced by a significant increase in number of studies on the gut microbiome's role in colorectal cancer, which has risen 200 times from 2000 to 2021 [66]. Regarding ML-based microbiome prediction, [65] conducted a study to investigate how the microbial community of black soldier fly changes over time due to host starvation. They observed a significant decrease in community diversity during the experiment. Using the Random Forest (RF) algorithm, they identified the most influential features that predict changes in the community. Notably, features such as 'Cell growth and death,' 'Transport and Catabolism,' and 'Cancers' were among the top predictors.

Also, an experiment was conducted by [58] to determine the correlation between wild fish and their gut microbiota. Two species of teleost wild fish were collected from multiple lakes and islands, and the locations were categorised into two groups based on anthropogenic aquatic impacts: compromised aquatic environments (CAE) and non-compromised aquatic environments (non-CAE). Using ML models, the researchers attempted to classify new fish of those species into either CAE or non-CAE categories. Apart from that, [44] describes the first ML data analysis of Parkinson's disease microbiota dysbiosis using RF, neural network, and support vector machine models. However, the discrepancies in their outcome were attributed to a lack of standardised experimental and bioinformatic protocols, and the authors recommended creating a unique standard to ensure reliable comparisons in future studies. Other ML-based microbiome predictions include an experiment conducted on 2,320 individuals from Prince of Wales Hospital in Hong Kong. The author utilised their faecal microbiome sequence data to predict nine well-characterised phenotypes [53]. Another study by [41] created a learning framework called DeepMicro that utilises autoencoders and other ML algorithms to classify diseases such as

inflammatory bowel disease, type-2 diabetes and colorectal cancer. Their proposed framework outperforms the current state-of-the-art approaches conducted on humans while significantly reducing the input data's dimensions.

While there are many studies on these fruit flies gut microbiome and microbiome-regulated behaviours, no ML model has been developed so far to smoothen the microbiome study and assist in agricultural pest management. This article has several contributions as shown below, to fill these gaps.

1. We use metagenomic samples taken from fruit flies (tephritidae) to analyse their microbiome composition. Using this knowledge, we observe how these microbiome compositions change with the developmental stage of the fly, as well as how they change across flies successively bred within a laboratory
2. We analyse whether the complete microbiome composition of fruit flies or elements within this microbiome can be predicted by different ML models, using data obtained from experiments as predictors
3. We evaluate which ML models are best suited to predict systems such as gut microbiomes. The results of this research are intended to be used as tools to identify vital taxological families, genera and species of the microbiome that could be selected as targets for future insect pest control methods and to provide preliminary indications of the microbiome composition of a given sample.

2 Methodology

2.1 Data description

We used data from [34] to construct the ML models. The raw data are presented in a tabular form that contains the total measured value of specific OTUs, which can then be mapped to particular bacterial phyla, classes, orders, families and genera. Within the raw data, there were 452 OTUs and 115 samples. After concatenating the OTUs to their corresponding bacteria, there were 6 unique phyla, 13 unique classes, 13 unique orders, 21 unique families and 47 unique genera.

Each sample represents a fly mapped to a developmental stage and generation using a provided mapping file. The stages include larvae, pupae, and adult flies, with adults further divided by male or female sexes. Samples are categorised into one of five generations, with most sets consisting of six samples. Sample size details can be found in the supplementary material (Table 5).

To simplify ML models, OTUs were grouped by taxonomic families instead of genera, considering that variation is prominent at the family level [34]. The raw data lacked units for bacterial measurements, so normalisation was performed. The relative abundance of each OTU was calculated as a fraction of the bacterial family's amount relative to the total bacteria within each sample. This normalisation created composition vectors for each sample, summing up to 100%.

To assess the quality of the processed data, the variance in relative abundances of bacterial families was measured for each developmental stage. Correlation matrices were generated to understand the pair-wise correlation between bacterial families within each stage and across all samples. An average matrix was constructed from the previous correlation matrices to identify trends in these correlations.

2.2 Model development

2.2.1 Model 1—intrafamilial successor prediction

This model, as shown in Fig. 1, aims to predict the relative abundance of each bacterial family in Generation 5 using the past four generations as predictors. The 'n' refers to the number of samples in each generation. The first four generations of each developmental stage were grouped to form the predictor set, while the generation 5 samples formed the outcome set. As each developmental stage has a different number of samples in Generations 1 to 4, each was reduced to the smallest size, resulting in sizes of n = 3 for Generation 1, n = 5 for Generation 2, n = 6 for Generation 3, n = 6 for Generation 4 and n = 6 for Generation 5. Consequently, this resulted in training data with a predictor set of size n = 20 and an outcome set size of n = 6 for each bacterial family in each developmental stage.

Four different ML methods were employed to construct models: least angle regression (LAR), RF, Lasso and elastic-net (EN). The model was then tested on the adult Male developmental stage samples to measure the performance. The root mean square error (RMSE) was selected as the evaluation metric for the model due to its low computational cost and simplicity of implementation [63]. Leave-one-out cross-validation was employed to obtain a more robust RMSE estimation and detect and prevent data over-fitting. From this cross-validation process, the standard error of measurement for the test scores was calculated so that it could be used to compare the performance between the different ML methods.

2.2.2 Model 2—interfamilial quantitative prediction

This model aims to predict the amount of a selected bacterial family in each sample using the amount of all other bacteria present in the sample as predictors (Fig. 2). Because this model uses data containing the amount of

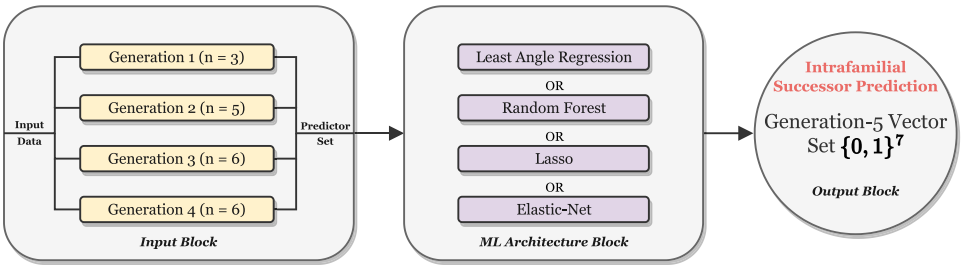


Fig. 1 Intrafamilial successor prediction model architectural overview. Here, 'n' refers to the number of samples in each generation

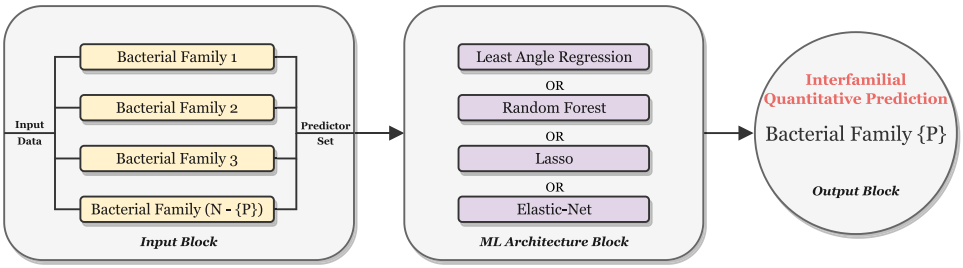


Fig. 2 Interfamilial quantitative prediction model architectural overview. We use the variable 'N' to denote the total number of bacterial families to represent a generic figure. Additionally, we assign the label 'P' to the particular bacterial family that the model will predict, and we represent it as an element of a set enclosed in curly braces

other bacterial families as predictors, the relative abundances cannot be used, as this would reduce the model to a simple subtraction of all relative abundances from 100%. Consequently, the raw measurements provided in the data were used to construct and test the ML models.

The dataset was iterated over, whereby one bacterial family was selected as the outcome set for each iteration, while all other bacterial families formed the predictor set. Furthermore, the same four ML methods employed for constructing Model 1 were used again. The model was then tested to measure the performance, with the coefficient of variance (CV) selected as the test score with leave-one-out cross-validation. It was done, as CV is a standardised form of RMSE, being normalised by the mean of the actual values, which allows greater comparison between different bacterial families [9]. The mean CV and standard error were calculated from this set of CV values obtained through the leave-one-out cross-validation.

Families of several orders, such as the Lactobacillales [64], Enterobacterales [52], and Bacillales [18], are abundant in the gut microbiome of fruit flies and are linked to certain roles in the microbial communities. Such behaviour facilitates the formation of co-occurrence patterns among the families, therefore projecting the interfamilial quantitative predictive model's capability.

2.2.3 Model 3—interfamilial qualitative prediction

The construction of this model was adapted from [36] and was developed in three phases. The first phase aimed to predict the relative abundance of each bacterial family using information on the presence or absence of all bacterial families within a sample as predictors. To create the predictor set for this model, each corresponding *composition vector* for the samples was reduced to a family vector, $P \in \{0,1\}^N$. These vectors contain binary elements, where the *ith* bacterial family, P_i, is assigned to 1 if present and 0 if absent. The original *composition vectors* were used as the outcome set. The model was then tested to measure the performance, with RMSE selected as the test score and leave-one-out cross-validation employed. From this set of RMSE values obtained through the leave-one-out cross-validation, the mean RMSE and standard error were calculated.

The second phase individually introduced two new predictor sets into the model. The first set contained *generation vectors*, $G \in \{0,1\}^5$, describing which of the five generations the sample belonged to. The second set contained *developmental stage* vectors, $D \in \{0,1\}^4$, indicating whether the sample was from a larvae, pupae, adult female or adult male fly. These vectors were individually appended to the family vectors, and the data were split into a training and testing set in the ratio of 70:30. The model was then tested to measure the performance, with RMSE selected and leave-one-out cross-validation employed. The third phase appended all three predictor vectors with the same method of training and testing applied. Finally, we obtain our desired set of bacterial family vectors by following the cNODE deep learning architecture [36]. A pictorial representation is shown in Fig. 3.

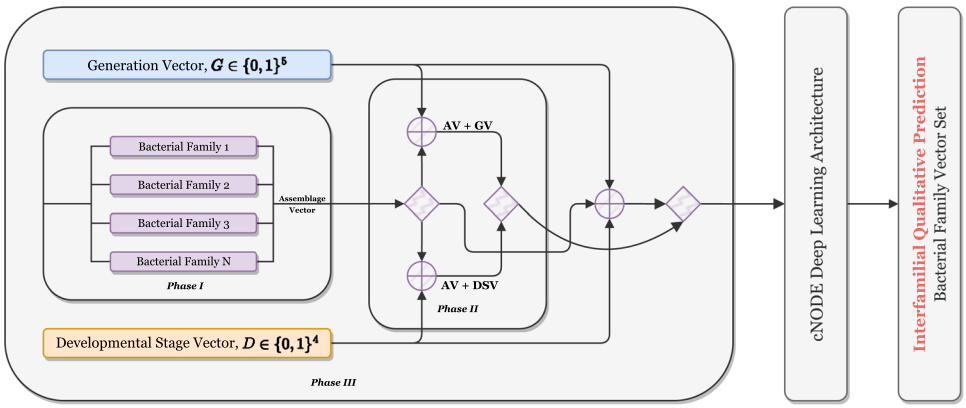


Fig. 3 Interfamilial qualitative prediction model architectural overview. This model is comprised of three distinct phases. Phase I's input is comprised of bacterial families of four generations. In phase II, the input is comprised of the input of phase I cascaded with generation vector G and developmental stage vector D separately. Unlike phase II, the input data in phase III comprise the input values of phase I and both generation vector G and developmental stage vector D simultaneously. The output or target column remains the same in all phases: the Generation 5 vector. Here, cNODE refers to the 'compositional neural ordinary differential equation' deep learning architecture proposed in [36]

This qualitative predictive ML model has an advantage over the traditional approach in using quantitative microbiome profiles like species-level relative abundances and strain-specific markers to generate accurate predictions[43]. Moreover, this kind of ML model can forecast complex relationships between the microbiota and their host, providing insights into the interactions. [19].

3 Results

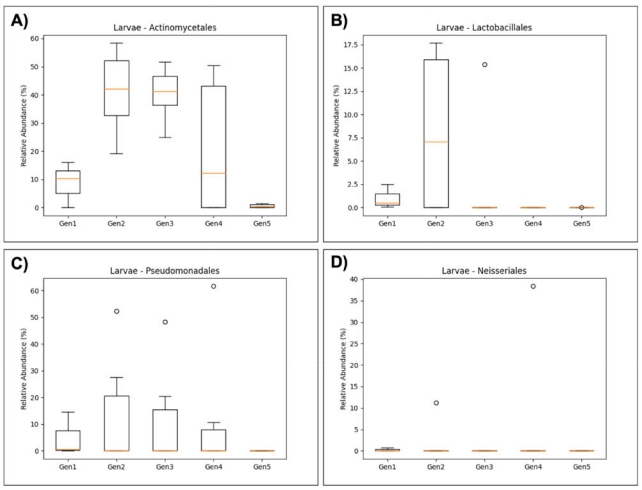
3.1 Data variance and correlation

To inform the development of the ML models, the underlying data were characterised to understand the levels of variance amongst the number of bacteria present and identify any high-level correlation that may exist between the relative abundance of different bacterial families. The larvae samples exhibited the highest diversity in the measured number of families, with 16 different families present in this developmental stage, consistent with the findings in [34]. As a result, the variance in the relative abundances of families was observed to be the largest in larvae samples. Different types of variance in the larvae samples are highlighted in Fig. 4a, describing the relative abundance of a bacterial family.

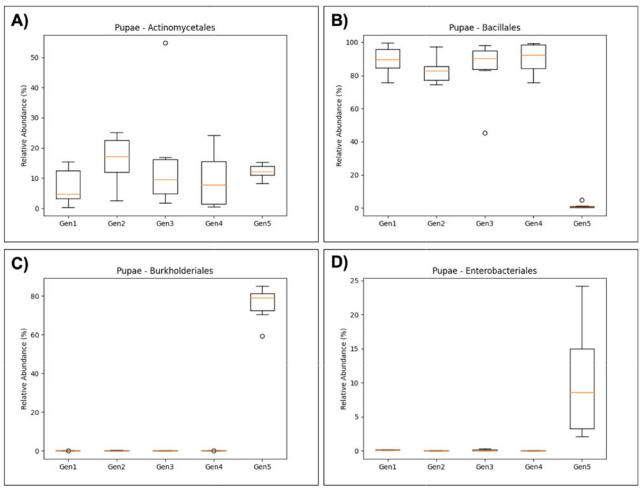
Figure 4a (A) provides an example of the large variance in the abundance of a family observed across different generations, with the amount of *Actinomycetales* sharply increasing between Generations 1 and 2, remaining steady across Generation 2 and 3, and then sharply decreasing between Generations 3 to 5. Conversely, Fig. 4a (B) demonstrates how a bacterial family, such as *Lactobacillales*, can be almost entirely absent from every generation but present in a relatively significant proportion in a single generation. For some families, such as *Pseudomonadales*, the mean abundance across all generations is almost zero; however, a handful of samples skew the spread of measured abundance towards large values, as highlighted in Fig. 4a (C). Finally, several bacterial families are almost entirely absent from every generation in the larvae developmental stage. Yet, one or two samples record a large amount of the bacteria, as observed with *Neisseriales* in Fig. 4a (D).

The lowest diversity among all the families was observed in pupae samples, with 12 different families present in this developmental stage. Consequently, the lowest variance was observed in pupae samples, as evidenced in Fig. 4b, describing the relative abundance of four bacterial families. The majority of the bacteria present in pupae samples from Generations 1 to 4 are *Actinomycetales* and *Bacillales*, as shown in Fig. 4b (A) and 4b (B). In Generation 5, the amount of *Bacillales* drops significantly, being largely replaced by *Burkholderiales* and *Enterobacterales*, as shown in Fig. 4b (C) and b (D). The other eight families observed in the pupae developmental stage are present in almost insignificant amounts (<3%) for every sample of each generation. Similarly, the changes in the adult female and male developmental stages are shown in Fig. 4c and d, respectively.

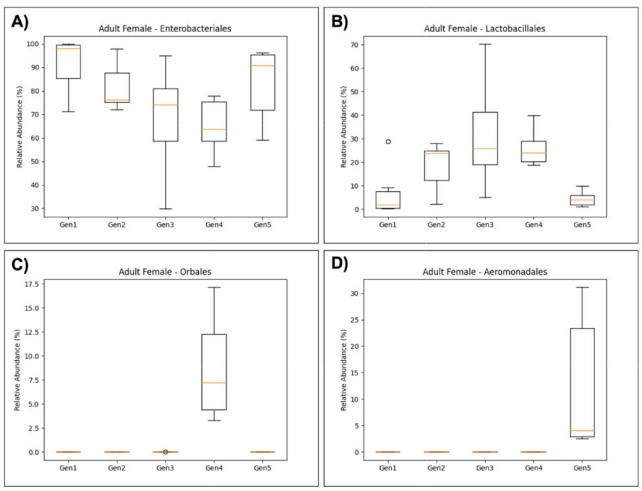
Figure 5 shows that most bacterial families exhibit significant variations and low correlation among them. Handling such variations requires careful feature engineering and preprocessing to manage differences in scales, distributions, and missing values. A correlation heatmap was generated to identify high-level relationships between the families. Notably, *Bacillales*, *Lactobacillales*, *Enterobacterales*, and *Actinomycetales* show significant



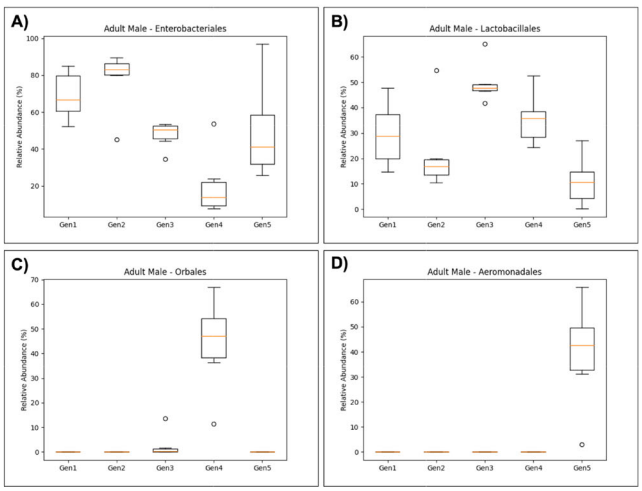
(a) Larvae – (A) *Actinomycetales*, (B) *Lactobacillales*, (C) *Pseudomonadales*, (D) *Neisseriales*



(b) Pupae – (A) *Actinomycetales*, (B) *Bacillales*, (C) *Burkholderiales*, (D) *Enterobacteriales*



(c) Adult female – (A) *Enterobacteriales*, (B) *Lactobacillales*, (C) *Orbales*, (D) *Aeromonadales*



(d) Adult male – (A) *Enterobacteriales*, (B) *Lactobacillales*, (C) *Orbales*, (D) *Aeromonadales*

Fig. 4 Relative abundances of some selected bacterial families ranging from Generation 1–5

relationships (Fig. 5). However, clusters of high correlation, particularly among *Erysipelotrichales*, *Bacteroidales*, and *Clostridiales*, are often due to their low relative abundance. The strongest relationship is observed between *Bacillales* and *Enterobacterales*, with a correlation coefficient of −0.66, indicating antagonistic relative abundances. Similarly, *Actinomycetales* and *Enterobacterales*, as well as *Bacillales* and *Lactobacillales*, exhibit negative correlations of −0.55 and −0.42, respectively.

Some notable, cooperative relationships exist amongst the data, particularly between *Orbales* and *Lactobacillales*, *Flavobacteriales* and *Lactobacillales*, and *Actinomycetales* and *Pseudomonadales*, having correlation coefficients of 0.33, 0.34 and 0.30, respectively. Interestingly, the correlation coefficient between *Lactobacillales* and *Enterobacterales*, which were observed to be the dominant bacterial families present in the adult developmental stages, was only measured to be 0.16. Likewise, the correlation coefficient between *Actinomycetales* and *Bacillales*, which were observed to be dominant in the pupae developmental stage, was only measured to be 0.24.

Our result on relative abundance aligns with previous studies. During the 5th generation developmental stage, bacterial families consistently clustered together [33, 34]. One such family, *Enterobacterales*, was found to have increased abundance in all developmental stages [34]. We also observed that the pupae stage exhibited the lowest relative abundance levels, corresponding with findings from [33].

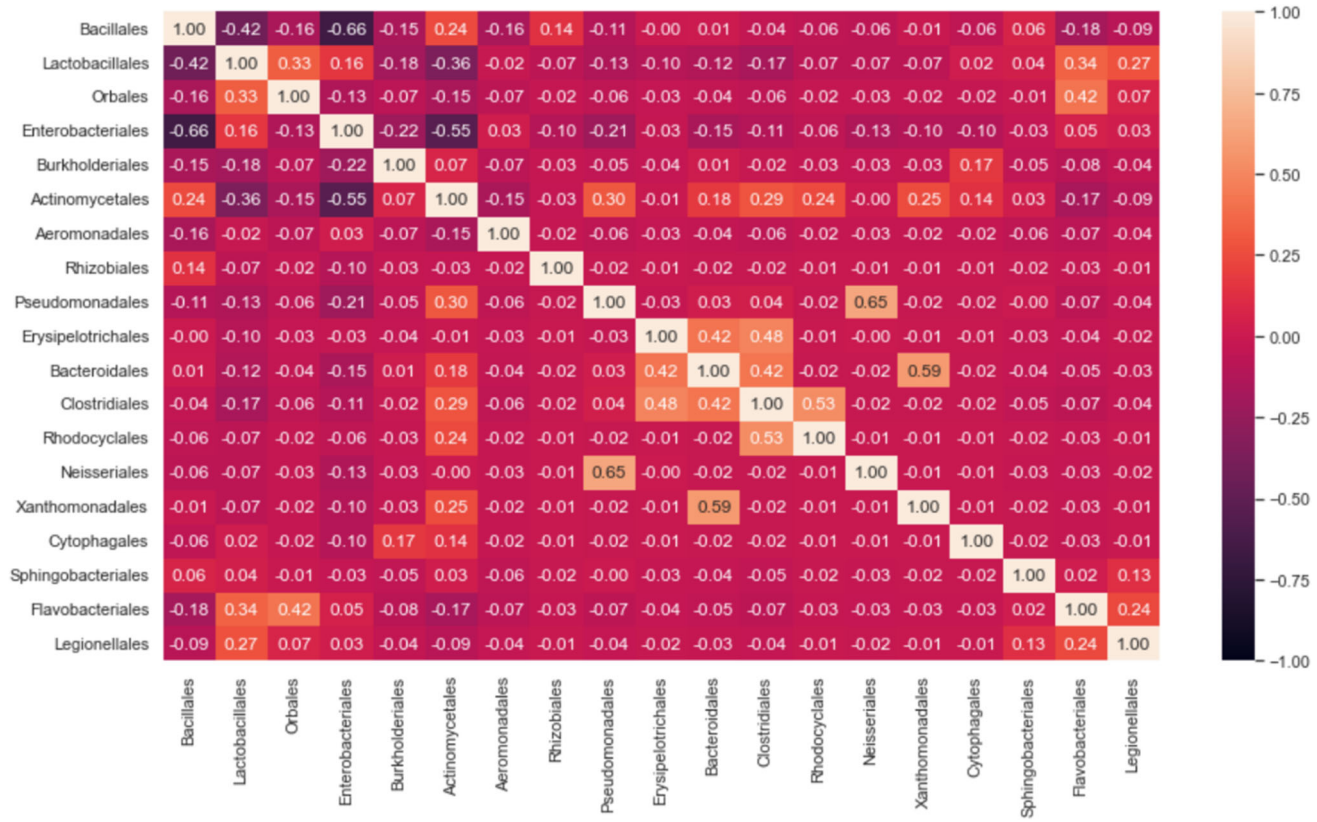


Fig. 5 Correlation heatmap for all the 19 families' samples used in the experiment, which reveals a strong correlation among *Bacillales*, *Lactobacillales*, *Enterobacterales* and *Actinomycetales*

3.2 Performance of intrafamilial successor prediction machine learning model

This model aimed to predict the relative abundance of each bacterial family in Generation 5 using the past four generations as predictors. The output of the developed model only predicted that seven bacterial families should be present in Generation 5 of each developmental stage. Obtained from the leave-one-out cross-validation, the mean RMSE and normalised RMSE (nRMSE) [45] between the actual and predicted relative abundances of each bacterial family in Generation 5 are presented in Tables 1 and 2, respectively. The most significant errors were observed in the *Enterobacterales* and *Burkholderiales* families, with the model getting the relative abundance of these families wrong by roughly 40%, on average. While not large, significant error was also observed in the prediction of *Aeromonodales* relative abundances, which was roughly 23% incorrect on average. Far less error was observed in the *Bacillales*, *Lactobacillales*, *Actinomycetales* and *Pseudomonadales* families with an average error of approximately 1%, 5%, 5% and 0.02%, respectively. *Pseudomonadales* are known to play a role in terpenoid biosynthesis in organic samples and fluorobenzoate degradation in conventional sites [7, 15, 37]. They are sometimes considered the primary symbiont group in many terrestrial isopod populations [8, 14]. However, in the case of nRMSE, the least relative error was observed in the case

Table 1 RMSE ± standard error for intrafamilial successor prediction model of the relative abundances of bacterial families

Family	LAR	RF	EN	Lasso	Mean
Bacillales	1.06 ± 0.31	0.67 ± 0.43	0.60 ± 0.44	**0.58 ± 0.45**	0.73 ± 0.41
Lactobacillales	6.89 ± 1.84	5.05 ± 1.70	**3.64 ± 1.44**	3.80 ± 1.56	4.85 ± 1.64
Enterobacterales	45.59 ± 8.14	48.74 ± 9.25	**38.89 ± 7.78**	41.16 ± 8.77	43.60 ± 8.49
Burkholderiales	38.14 ± 11.01	**36.42 ± 11.82**	46.64 ± 9.36	42.84 ± 15.75	41.01 ± 11.99
Actinomycetales	6.10 ± 1.76	5.49 ± 2.17	**4.36 ± 2.52**	**4.36 ± 2.52**	5.08 ± 2.24
Aeromonadales	**22.85 ± 5.09**	23.0 ± 5.01	**22.85 ± 5.09**	**22.85 ± 5.09**	22.89 ± 5.07
Pseudomonadales	**0.02 ± 0.01**	**0.02 ± 0.01**	**0.02 ± 0.01**	**0.02 ±0.01**	**0.02 ± 0.01**

Table 2 nRMSE ± standard error for intrafamilial successor prediction model of the relative abundances of bacterial families

Family	LAR	RF	EN	Lasso
Bacillales	1.46 ± 0.59	0.92 ± 0.65	0.82 ± 0.65	**0.80 ± 0.66**
Lactobacillales	1.42 ± 0.45	1.04 ± 0.39	**0.75 ± 0.32**	0.78 ± 0.35
Enterobacterales	1.05 ± 0.21	1.12 ± 0.24	**0.89 ± 0.20**	0.94 ± 0.22
Burkholderiales	0.93 ± 0.30	**0.89 ± 0.32**	1.14 ± 0.28	1.04 ± 0.41
Actinomycetales	1.20 ± 0.44	1.08 ± 0.49	**0.86 ± 0.53**	**0.86 ± 0.53**
Aeromonadales	**1.00 ± 0.25**	**1.00 ± 0.25**	**1.00 ± 0.25**	**1.00 ± 0.25**
Pseudomonadales	**1.00 ± 0.56**	**1.00 ± 0.56**	**1.00 ± 0.56**	**1.00 ± 0.56**

of *Burkholderiales* and *Enterobacterales* families, respectively, followed by *Aeromonadales* and *Pseudomonadales* families with minimal equal errors. The mean RMSE was consistent across all methods, with the best-performing method, Lasso, making an average error of 16.52%, and the worst-performing method, LAR, making an average error of 17.24%. Similarly, the EN algorithm made an average error of 16.71%, while the RF method made an average error of 17.05%.

The predicted relative abundances can be graphed against the actual values to better characterise this model's performance and understand the source of these errors. A figure of predicted versus the actual relative abundance of the *Lactobacillales* family within the adult male developmental stage for each of the four ML methods has been shown in the supplementary material (Fig. 7). The best-performing method for this particular case was Lasso, with an RMSE of 8.19. Across all methods, while it appears that the model correctly predicts higher values when the actual values increase, a much lower magnitude in the predicted relative abundances is observed compared to the actual values. This is much more noticeable in the LAR output, where the predicted relative abundance does not exceed 10% when the actual values can get as large as 25%.

A similar comparison is presented in a figure in the supplementary material (Fig. 8), showing the predicted versus actual relative abundance of the *Enterobacterales* family within the adult male developmental stage for each of the four ML methods. Lasso was the best-performing method for this particular case, with an RMSE of 22.22. For all of the methods used, the model could not successfully attribute higher predicted relative abundances with higher actual relative abundances, resulting in no clear relationships between predicted and actual values. The EN and Lasso methods output an almost identical prediction, while the RF method outputs a much higher range of relative abundances.

3.3 Performance of interfamilial quantitative prediction machine learning model

This model aimed to predict the amount of a given bacterial family in each sample using the amount of all other bacteria present in the sample as predictors. Obtained from the leave-one-out cross-validation, the mean coefficient of variance (CV) between the actual and predicted relative abundances of each bacterial family in Generation 5 is presented in Table 3. Across all ML methods, the model consistently predicted bacterial families with the least error were *Lactobacillales* and *Enterobacterales*.

The presence of *Lactobacillales* is a common factor that drives divergence in dietary treatment groups [29]. This group has been shown to improve pesticide resistance and help control gastrointestinal pathogens [16, 57]. Organic waste that has been inoculated with *Lactobacillales* tends to yield higher biomass output and a better nutritional spectrum in black soldier fly larvae compared to waste amended with artificial feed [54]. *Lactobacillales* also play a role in vitamin and cofactor metabolism [30]. On the other hand, the *Enterobacterales* family has been identified as an important biomarker for schizophrenia [38, 56, 68]. This family is known to produce short-chain fatty acids [68], which likely play a central role in the microbiota-host crosstalk that regulates brain function and behaviour [12, 60]. The families that the model consistently predicted with the highest error were *Rhizobiales*, *Rhodocyclales*, *Erysipelotrichales* and *Neisseriales*.

The RF method was by far the most accurate in its predictions, with a mean CV of 1.25. The LAR and EN methods performed similarly, with a mean CV of 1.76 and 1.61, respectively. Due to the large inaccuracies in predicting the *Bacillales* family, the Lasso method performed worst, with a mean CV of 3.06.

3.4 Performance of interfamilial qualitative prediction machine learning model

This model aimed to predict the relative abundance of each bacterial family within a sample using binary information on the presence or absence of all bacterial families as

Table 3 Coefficient of variance ± standard error for interfamilial quantitative prediction model of the relative abundances of bacterial families

Family	LAR	RF	EN	Lasso
Bacillales	1.58 ± 0.12	**0.54 ± 0.11**	1.35 ± 0.19	33.11 ± 32.09
Lactobacillales	1.20 ± 0.08	**0.35 ± 0.05**	0.94 ± 0.10	1.14 ± 0.31
Orbales	1.79 ± 0.32	**0.85 ± 0.24**	2.06 ± 0.64	1.31 ± 0.33
Enterobacterales	1.02 ± 0.04	**0.21 ± 0.04**	0.90 ± 0.05	0.82 ± 0.10
Burkholderiales	1.89 ± 0.36	**0.31 ± 0.11**	1.82 ± 0.37	1.54 ± 0.38
Actinomycetales	1.47 ± 0.18	**0.59 ± 0.15**	0.86 ± 0.18	2.20 ± 1.45
Aeromonadales	1.81 ± 0.33	1.45 ± 0.27	1.80 ± 0.33	**1.37 ± 0.35**
Rhizobiales	2.00 ± 0.99	2.22 ± 1.11	2.00 ± 1.00	**1.92 ± 1.00**
Pseudomonadales	1.74 ± 0.49	**0.95 ± 0.37**	1.80 ± 0.53	2.22 ± 0.87
Erysipelotrichales	1.96 ± 0.66	2.06 ± 0.76	1.95 ± 0.71	**1.69 ± 0.72**
Bacteroidales	1.93 ± 0.53	**1.35 ± 0.55**	1.63 ± 0.57	1.41 ± 0.57
Clostridiales	1.82 ± 0.33	1.31 ± 0.34	1.40 ± 0.32	**1.07 ± 0.35**
Rhodocyclales	2.00 ± 0.99	**1.69 ± 1.18**	2.03 ± 1.06	1.80 ± 1.04
Neisseriales	1.95 ± 0.76	1.82 ± 0.86	2.06 ± 0.77	**1.30 ± 0.78**
Xanthomonadales	1.82 ± 0.80	1.83 ± 1.00	1.87 ± 0.90	**1.62 ± 0.87**
Cytophagales	2.00 ± 0.99	1.73 ± 1.05	2.00 ± 0.99	**1.03 ± 1.00**
Sphingobacteriales	1.66 ± 0.41	**1.27 ± 0.48**	1.45 ± 0.41	1.41 ± 0.40
Flavobacteriales	1.74 ± 0.29	**0.88 ± 0.28**	1.23 ± 0.28	1.23 ± 0.28
Unknown	1.84 ± 0.38	1.49 ± 0.4	1.28 ± 0.37	**1.26 ± 0.37**
Legionellales	1.91 ± 0.59	2.09 ± 0.69	1.85 ± 0.59	**1.76 ± 0.59**
Mean	1.47 ± 0.48	**1.25 ± 0.50**	1.61 ± 0.52	3.06 ± 2.19

predictors. Furthermore, this model aimed to build in further information detailing the generation and developmental stage from which the sample originated. A comparative analysis of the overall performance between the model with information limited strictly to the presence or absence of bacterial families and successive models with more information introduced is presented in Fig. 6, which presents the mean RMSE for each model. From this comparison, it can be seen that the mean RMSE for the model limited to information on the presence and absence of bacterial families is 7.14, while the introduction of information detailing which generation the sample originates from results in a slight increase in the mean RMSE to 7.30. Adding information about the developmental stage of the sample to the model leads to the biggest improvement in performance, which reduces the mean RMSE to 5.27. Only a slight reduction in the RMSE to 5.12 is observed when the model is fed with all available information.

4 Discussion

4.1 Variance between different generations of laboratory-reared fruit flies leads to significant errors in intrafamilial models

The overarching aim of this research was to use metagenomic samples taken from fruit flies to analyse their microbiome composition and, consequently, understand whether different ML models could make predictions of the microbiome composition. Regarding the intrafamilial model developed in this research, the goal was to use this tool to observe how these microbiome compositions may change across different generations within a given developmental stage of flies. Using four different ML models, the relative abundances of seven different bacterial families were predicted for Generation 5 of each developmental stage, using the past four generations as predictors. By performing a comparative analysis of the mean RMSE between the actual and predicted relative abundances of the seven different bacterial families, no discernible difference was observed between the ML methods. This lack of diversity in performance amongst the methods is largely attributed to the poor performance of the overall model, with the relative abundance of some bacterial families being predicted in errors of close to 50% for all methods.

While similarly large errors are observed across several bacterial families, the reason for these errors can be attributed to a range of different sources. For example, the *Enterobacterales* family had a mean RMSE between 39% and 48%; however, the average relative abundance of *Enterobacterales* in Generation 5 amongst all developmental stages was approximately 60%. Furthermore, the range of *Enterobacterales* relative abundance was the largest observed, with almost 100% relative abundance in Generation 5 of larvae but only 10% in Generation 5 of

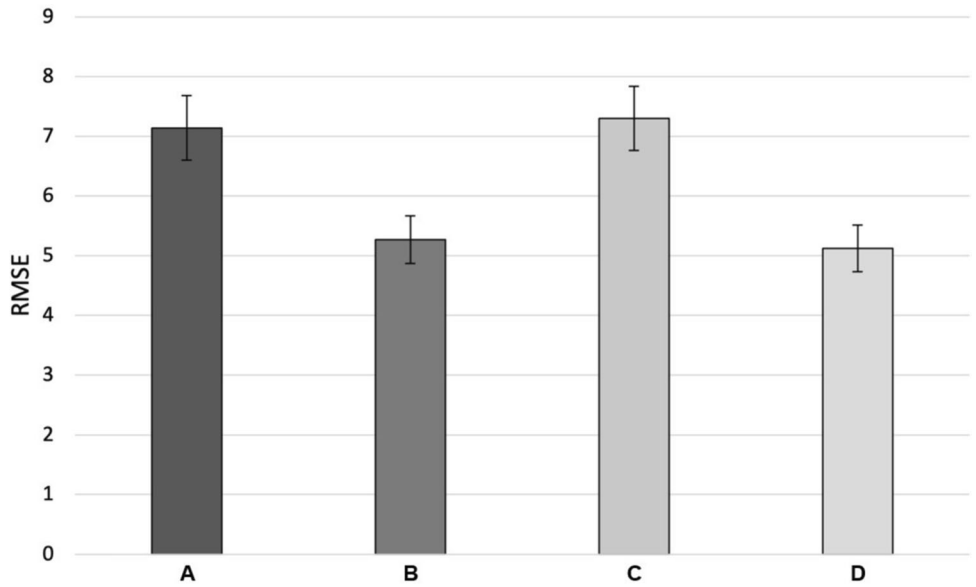


Fig. 6 Mean RMSE between actual and predicted relative abundances for the different ML methods used in constructing the interfamilial qualitative prediction model with error bars representing the standard error of the mean. Here, the first column "A" contains information about the presence of bacterial families, indicating whether a particular bacterial family is present in a given metagenomic sample. In the second column "B", information regarding the developmental stage has been added along with the presence status. Similarly, in the third column "C", additional generational information has been added along with information similar to the first column. Finally, the fourth column "D" contains all the information about a microbial family

pupae. This high average and the large range of relative abundances lead to a significant error in prediction by the model, as no discernible pattern can be resolved amongst the highly variable outcomes. While this was the case for *Enterobacterales*, the *Burkholderiales* family, which had a mean RMSE between 36% and 47%, was predicted in error due to its relative absence in all developmental stages, except the pupae stage.

Although the poor performance of this model is disappointing, the results reiterate previous observations made by [34], where they describe significant shifts in the overall microbiome composition between Generations 1 and 5. This large variance between generations inherently leads to considerable uncertainty in predictions made by the ML model, particularly noting that the changes between generations are completely different across developmental stages. Furthermore, for the data used in the development of this model, the maximum sample size within a generation was six, which limits the ability of the ML model to identify and build on any significant patterns that may exist within or between generations. As a result of the high variability in the microbiome composition between generations, coupled with the low sample size used to characterise these populations, a novel conclusion cannot be drawn from the results of this model, particularly regarding how the microbiome compositions may change across different generations. As described consistently throughout the literature, these microbiomes are highly complex systems that often display no discernible patterns in how they develop, change and evolve over time [4, 69]. Moreover, within a given microbiome, there is a high amount of functional redundancy between bacterial families, whereby some bacteria can easily be replaced by others with similar functions without causing any changes to the fly's physiology [31]. As a result of this high complexity and plasticity within the gut microbiomes of fruit flies, models that can accurately predict the dynamics of these communities across different generations would require far more detail in the development of the ML algorithms used, alongside a much deeper pool of data to train these models with.

4.2 Random forest algorithm is more appropriate in characterising interfamilial models

Building on the observations made within the first model and towards achieving the overall goal of using metagenomic samples to analyse and predict the microbiome composition of fruit flies, the interfamilial model developed in this research was to characterise the relationships between different bacterial families. Across all ML methods, vastly different prediction patterns were observed, indicating that these relationships may be harder to decipher than anticipated. A comparative analysis of the mean CV between the actual and predicted amounts of each

bacterial family within all samples demonstrated that the RF ML method was the most capable of producing predictions resembling the actual measurements. When the output of each method was compared to the actual number of bacterial families present, it was clear that LAR, EN and Lasso methods produced inaccurate predictions. Specifically, there was no resolvable relationship between the predictions generated by these methods and the actual observations in the samples.

The LAR model predictions aligned almost perfectly linearly with the actual values; however, the relationship was inverted. Furthermore, the range of values in which the LAR model predicted the amount of bacteria should be present was far smaller than what was actually measured. The EN and Lasso methods could output predictions that somewhat more accurately resembled the actual amounts of the bacterial families present within the samples; however, the performance of these methods remained very poor. While these three ML methods yielded no useful predictions, the RF method produced some surprising results. This method yielded a somewhat proportional, linear relationship between predicted and actual amounts for most bacterial families. This is highlighted by RF having the lowest mean CV, significantly outperforming the other three models. For some rarer bacterial families, this method tended to make more errors in prediction; however, more importantly, for almost all of the bacterial families, this method successfully predicted the correct range of the amounts of bacteria. While the model's results signify that RF methods are best suited to microbiome evaluation, the model's accuracy still significantly limits its usefulness as a reliable predictive model in its current form. To accurately pinpoint the source of these errors, some rarer bacterial families could be removed from the dataset to allow the model to hone in on the more prevalent families.

Despite the positive results of this model, its usefulness in a wider context should be considered. The fundamental basis of this model implies that there is some existing information on the composition of a gut microbiome that can be used to predict an otherwise unknown amount of a selected bacterial family. To influence the physiology and behaviour of fruit flies and their impact on important crops, it may be possible to introduce certain bacterial families to create significant shifts in microbiome composition. As a result, an ML model such as that generated in this study may be able to be fed information on the existing composition of a gut microbiome and provide indications as to how bacterial family candidates may propagate amongst the system. Obviously, this would require the model to be trained on systems where the candidate bacterial family is already present to obtain any existing relationships.

4.3 Predictions of the overall microbiome composition are largely driven by the developmental stage

The final model within this study aimed to understand whether the complete microbiome composition of fruit flies could be predicted, only utilising information that detailed which bacterial families were present and which generation and developmental stage the samples originated from. A clear comparison between how such information informs the model was drawn by systematically integrating these elements into the predictive model. The most basic of these models, in which the presence or absence of bacteria was the only information fed in, was able to predict the overall composition of samples with a relatively low amount of error. The performance of this basic model was largely limited to predicting the correct types of bacteria present rather than the correct relative abundances. Upon the introduction of information detailing which generation the samples originated from, no discernible difference in the model's performance was observed. These results can be interpreted in two ways: (*i*) the generation to which a fruit fly belongs does not significantly contribute to its overall microbiome composition; or (*ii*) the variance amongst the gut microbiomes within and between each generation is so large that no reasonable pattern can be drawn.

The introduction of indicators for which developmental stage the samples belonged yielded the largest increase in performance of the model, with a reduction to the lowest levels of prediction error observed throughout any of the models developed in this study. This increase in performance from the base model was largely attributed to the readjustment of the correctly predicted bacterial family relative abundances towards the true values. Furthermore, introducing this information often resulted in the model correctly removing bacterial families that the base model predicted should be present but were, in fact, absent from the actual samples.

Despite these promising results, there were instances where the iterations of the model were unable to accurately predict the overall composition of some microbe samples. They often overestimated the amounts of particular bacterial families and included bacterial families that were entirely absent from the actual measurements. While the source of these errors is hidden within the cryptic patterns between the relative abundances of different bacterial families, the results further signify and reiterate the highly complex nature of these systems.

To enhance the proposed method further, one promising approach could be transfer learning, yielding more positive results in classifying microbial communities [13]. Moreover, more curated data normalisation and other feature

engineering techniques should be applied to maintain a balance between the range of the data. This would help reduce the weight biases of the ML model. Additionally, establishing a uniform dataset based on lab-grown species as a standard would facilitate better comparison of the models' performance. Lastly, introducing Explainable AI (XAI) can make the models much more readable, ultimately aiding in translating our findings for other gut microbiome studies.

4.4 Comparison with existing literature

In the context of animal gut microbiome comparison, the results from [58] were surprising, with an accuracy rate of over 90% (equivalent to an error percentage of 10%). Experiments conducted by [44] on the microbiota dysbiosis in Parkinson's disease yielded an AUC score of 0.8 and an accuracy score of 71% using the RF algorithm. Another study conducted by [53] in Hong Kong produced an accuracy of 83.3% and AUC scores ranging from 0.90 to 0.99. Although it is challenging to compare all of these studies on a common ground due to variations in their methodologies, on average, we can say that each of these models performs with an error rate somewhere between 10% and 30%. This indicates a high variance in the models' prediction capability compared to our proposed solution. A summary of the outcomes from existing literature compared to ours is shown in Table 4.

4.5 Implications for pest control

The implications of our proposed machine learning-based gut microbiome composition prediction for fruit flies are significant for pest control strategies in agriculture. By employing the proposed three different models in practice, we can gain insights into the microbiome dynamics that influence the fruit flies' food-digesting capacity and their resistance to pathogens. This understanding can lead to the identification of key microbial taxa that play a crucial role in pest management, potentially allowing for the development of biocontrol strategies that leverage these microbial communities to suppress pest populations more sustainably than traditional insecticides [10, 62]. Moreover, the results of our study highlight the complexity and variability of these microbiomes, suggesting that a one-size-fits-all approach to pest management may be ineffective. Instead, targeted interventions considering the specific microbiome profiles of different fruit fly populations could enhance the efficacy of integrated pest management practices.

Furthermore, understanding the microbiome's role in pest behaviour, such as attraction to host plants through volatile organic compounds emitted by associated bacteria, can inform the development of more effective trapping and monitoring systems [1]. This could reduce reliance on chemical pesticides, aligning with sustainable agricultural practices that prioritise ecological balance and long-term pest management solutions [40]. Overall, our findings advocate for a paradigm shift in pest control, emphasising the potential of microbiome manipulation as a viable strategy for enhancing agricultural resilience against insect pests.

5 Conclusion

This study used metagenomic samples from fruit flies (tephritidae) to characterise their microbiome composition. Furthermore, this study aimed to use this knowledge to observe how these microbiome compositions change with the development of flies across their lifespan and between generations. The second key aim of this research was to understand whether the complete microbiome composition of fruit flies or elements within this microbiome could be predicted using simple ML models. Extending on this, a further aim was to understand which ML models best predict systems such as gut microbiomes. The first of these models did not perform well and could not accurately predict the relative abundances of bacterial families in a final generation. This was partly due to the large variance between the gut microbiome composition of different generations of flies within the data used. The second model's performance was promising and demonstrated that simple ML models can be built around the RF method to predict gut microbiomes with reasonable accuracy. The

Table 4 ML model performance comparison with existing literature

Host	Best ML model	Performance score	Ref
Fish	Neural network	Accuracy = 0.96	[58]
Human	Random forest	Area under the curve (AUC) = 0.90 ~ 0.99	[53]
Human	Random forest	Accuracy = 0.71, AUC = 0.80	[44]
Human	DeepMicro	Mean AUC = 0.84[1]	[41]
Fruit Flies	Random forest	CV = 1.25 ± 0.50	Us

[1]Mean AUC score of 5 public datasets

final model was the most successful, able to predict the overall composition of gut microbiomes with significant accuracy. The overall results of this study firstly demonstrate how complex these dynamic systems are but also signify that more computationally efficient methods can be developed to determine microbiome compositions in a more practical way. This method can be used as a biological tool for better pest management by modifying the microbiome dynamics and understanding their role in pest behaviour. The absence of simple ML models performing these types of analyses within the literature demonstrates the importance of the results within this study. It indicates that further parameter tuning within these models could yield greater accuracy. Although our model explored the fruit fly data as an example with our three proposed ML models, the methods are easily transferable to any other fruit fly or microbiome study.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s00521-025-10999-9.

Acknowledgements We would like to thank Professor Salman Durrani and Professor Dan MacDonald from the Australian National University for their support and time throughout this research project.

Funding The author(s) received no financial support for the research, authorship, and/or publication of this article.

Data availability The datasets used during and/or analysed during the current study are publicly available in the NCBI database under Bioproject PRJNA717989 and are available at the following URL: https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJNA717989.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

References

1. Awad M, Ben Gharsa H, ElKraly OA, Leclerque A, Elnagdy SM (2022) COI haplotyping and comparative microbiomics of the peach fruit fly, an emerging pest of Egyptian olive orchards. Biology. https://doi.org/10.3390/biology12010027
2. Aidley DJ (1976) Alternatives to insecticides. Sci Prog 63(250):293–303
3. Aktar W, Sengupta D, Chowdhury A (2009) Impact of pesticides use in agriculture: their benefits and hazards. Interdiscip Toxicol 2(1):1–12. https://doi.org/10.2478/v10102-009-0001-7
4. Australian Bureau of Agricultural Resource Economics and Sciences: Snapshot of Australian Agriculture 2022. https://www.agriculture.gov.au/abares/products/insights/snapshot-of-australian-agriculture-2022
5. Ami EB, Yuval B, Jurkevitch E (2010) Manipulation of the microbiota of mass-reared Mediterranean fruit flies Ceratitis capitata (Diptera: Tephritidae) improves sterile male sexual performance. ISME J 4(1):28–37. https://doi.org/10.1038/ismej.2009.82
6. Bakula M (1969) The persistence of a microbial flora during postembryogenesis of Drosophila melanogaster. J Invertebr Pathol 14(3):365–374. https://doi.org/10.1016/0022-2011(69)90163-3
7. Bartuv R, Berihu M, Medina S, Salim S, Feygenberg O, Faigenboim-Doron A, Zhimo VY, Abdelfattah A, Piombo E, Wisniewski M, Freilich S, Droby S (2023) Functional analysis of the apple fruit microbiome based on shotgun metagenomic sequencing of conventional and organic orchard samples. Environ Microbiol 25(8):1728–1746. https://doi.org/10.1111/1462-2920.16353
8. Bredon M, Herran B, Lheraud B, Bertaux J, Grève P, Moumen B, Bouchon D (2019) Lignocellulose degradation in isopods: new insights into the adaptation to terrestrial life. BMC Genom 20(1):462. https://doi.org/10.1186/s12864-019-5825-8
9. Brown CE (1998) Coefficient of variation. In: Brown CE (ed) Applied multivariate statistics in geohydrology and related sciences. Springer, Berlin, Heidelberg, pp 155–157. https://doi.org/10.1007/978-3-642-80328-4_13
10. Bigiotti G, Sacchetti P, Pastorelli R, Lauzon CR, Belcari A (2021) Bacterial symbiosis in Bactrocera oleae, an Achilles' heel for its pest control. Insect Sci. https://doi.org/10.1111/1744-7917.12835
11. Caltagirone LE, Doutt RL (1989) The history of the Vedalia beetle importation to California and its impact on the development of biological control. Annu Rev Entomol 34(1):1–16. https://doi.org/10.1146/annurev.en.34.010189.000245
12. Caspani G, Swann J (2019) Small talk: microbial metabolites involved in the signaling from microbiota to brain. Curr Opin Pharmacol 48:99–106. https://doi.org/10.1016/j.coph.2019.08.001
13. Chong H, Zha Y, Yu Q, Cheng M, Xiong G, Wang N, Huang X, Huang S, Sun C, Wu S, Chen W-H, Coelho LP, Ning K (2022) EXPERT: transfer learning-enabled context-aware microbial community classification. Brief Bioinform 23(6):396. https://doi.org/10.1093/bib/bbac396
14. Dittmer J, Bouchon D (2018) Feminizing Wolbachia influence microbiota composition in the terrestrial isopod Armadillidium vulgare. Sci Rep 8(1):6998. https://doi.org/10.1038/s41598-018-25450-4
15. Devpura N, Jain K, Patel A, Joshi CG, Madamwar D (2017) Metabolic potential and taxonomic assessment of bacterial community of an environment to chronic industrial discharge. Int Biodeterior Biodegrad 123:216–227. https://doi.org/10.1016/j.ibiod.2017.06.011
16. Daisley BA, Trinder M, McDowell TW, Welle H, Dube JS, Ali SN, Leong HS, Sumarah MW, Reid G (2017) Neonicotinoid-induced pathogen susceptibility is mitigated by Lactobacillus plantarum immune stimulation in a Drosophila melanogaster model. Sci Rep 7(1):2703. https://doi.org/10.1038/s41598-017-02806-w
17. Furlan L, Pozzebon A, Duso C, Simon-Delso N, Sánchez-Bayo F, Marchand PA, Codato F, Lexmond M, Bonmatin J-M (2021) An update of the Worldwide Integrated Assessment (WIA) on systemic insecticides. Part 3: alternatives to systemic insecticides. Environ Sci Pollut Res 28(10):11798–11820. https://doi.org/10.1007/s11356-017-1052-5
18. Glasl B, Bourne DG, Frade PR, Thomas T, Schaffelke B, Webster NS (2019) Microbial indicators of environmental perturbations in coral reef ecosystems. Microbiome 7(1):94. https://doi.org/10.1186/s40168-019-0705-7
19. Gould AL, Zhang V, Lamberti L, Jones EW, Obadia B, Korasidis N, Gavryushkin A, Carlson JM, Beerenwinkel N, Ludington WB (2018) Microbiome interactions shape host fitness. Proc Natl Acad Sci. https://doi.org/10.1073/pnas.1809349115

20. Hernández Medina R, Kutuzova S, Nielsen KN, Johansen J, Hansen LH, Nielsen M, Rasmussen S (2022) Machine learning and deep learning applications in microbiome research. ISME Commun 2(1):1–7. https://doi.org/10.1038/s43705-022-00182-9
21. Hill MP, Macfadyen S, Nash MA (2017) Broad spectrum pesticide application alters natural enemy communities and may facilitate secondary pest outbreaks. PeerJ 5:4179. https://doi.org/10.7717/peerj.4179
22. Inamine H, Ellner SP, Newell PD, Luo Y, Buchon N, Douglas AE (2018) Spatiotemporally Heterogeneous Population Dynamics of Gut Bacteria Inferred from Fecal Time Series Data. mBio 9(1):01453–17. https://doi.org/10.1128/mBio.01453-17
23. Igbedioh SO (1991) Effects of agricultural pesticides on humans, animals, and higher plants in developing countries. Arch Environ Health Int J 46(4):218–224. https://doi.org/10.1080/00039896.1991.9937452
24. Kyritsis GA, Augustinos AA, Cáceres C, Bourtzis K (2017) Medfly gut microbiota and enhancement of the sterile insect technique: similarities and differences of klebsiella oxytoca and enterobacter sp. AA26 probiotics during the larval and adult stages of the VIENNA 8D53+ genetic Sexing Strain. Front Microbiol 8:2064. https://doi.org/10.3389/fmicb.2017.02064
25. Kyritsis GA, Augustinos AA, Ntougias S, Papadopoulos NT, Bourtzis K, Cáceres C (2019) Enterobacter sp. AA26 gut symbiont as a protein source for Mediterranean fruit fly mass-rearing and sterile insect technique applications. BMC Microbiol 19(S1):288. https://doi.org/10.1186/s12866-019-1651-z
26. Kirsch P (1988) Pheromones: their potential role in control of agricultural insect pests. Am J Altern Agric 3(2–3):83–97. https://doi.org/10.1017/S0889189300002241
27. Kumar M, Ji B, Zengler K, Nielsen J (2019) Modelling approaches for studying the microbiome. Nat Microbiol 4(8):1253–1267. https://doi.org/10.1038/s41564-019-0491-9
28. Khomich M, Måge I, Rud I, Berget I (2021) Analysing microbiome intervention design studies: comparison of alternative multivariate statistical methods. PLoS ONE 16(11):0259973. https://doi.org/10.1371/journal.pone.0259973
29. Klammsteiner T, Walter A, Bogataj T, Heussler CD, Stres B, Steiner FM, Schlick-Steiner BC, Insam H (2021) Impact of processed food (canteen and oil wastes) on the development of black soldier fly (Hermetia Illucens) larvae and their gut microbiome functions. Front Microbiol 12:619112. https://doi.org/10.3389/fmicb.2021.619112
30. McMullen JG, Bueno E, Blow F, Douglas AE (2021) Genome-inferred correspondence between phylogeny and metabolic traits in the wild drosophila gut microbiome. Genome Biol Evol 13(8):127. https://doi.org/10.1093/gbe/evab127
31. Moya A, Ferrer M (2016) Functional redundancy-induced stability of gut microbiota subjected to disturbance. Trends Microbiol 24(5):402–413. https://doi.org/10.1016/j.tim.2016.02.002. (**Accessed 2023-02-15**)
32. Matos RC, Leulier F (2014) Lactobacilli-host mutualism: ''learning on the fly''. Microb Cell Fact 13(Suppl 1):6. https://doi.org/10.1186/1475-2859-13-S1-S6
33. Majumder R, Sutcliffe B, Adnan SM, Mainali B, Dominiak BC, Taylor PW, Chapman TA (2020) Artificial larval diet mediates the microbiome of queensland fruit fly. Front Microbiol 11:576156. https://doi.org/10.3389/fmicb.2020.576156
34. Majumder R, Taylor PW, Chapman TA (2022) Dynamics of the queensland fruit fly microbiome through the transition from nature to an established laboratory colony. Microorganisms 10(2):291. https://doi.org/10.3390/microorganisms10020291
35. Mandal S, Van Treuren W, White RA, Eggesbø M, Knight R, Peddada SD (2015) Analysis of composition of microbiomes: a novel method for studying microbial composition. Microb Ecol Health Disease. https://doi.org/10.3402/mehd.v26.27663
36. Michel-Mata S, Wang X, Liu Y, Angulo MT (2022) Predicting microbiome compositions from species assemblages through deep learning. iMeta. https://doi.org/10.1002/imt2.3
37. McDonald RC, Watts JEM, Schreier HJ (2019) Effect of diet on the enteric microbiome of the wood-eating catfish Panaque nigrolineatus. Front Microbiol 10:2687. https://doi.org/10.3389/fmicb.2019.02687
38. Nguyen TT, Kosciolek T, Maldonado Y, Daly RE, Martin AS, McDonald D, Knight R, Jeste DV (2019) Differences in gut microbiome composition between persons with chronic schizophrenia and healthy comparison subjects. Schizophr Res 204:23–29. https://doi.org/10.1016/j.schres.2018.09.014
39. Nile AS, Kwon YD, Nile SH (2019) Horticultural oils: possible alternatives to chemical pesticides and insecticides. Environ Sci Pollut Res 26(21):21127–21139. https://doi.org/10.1007/s11356-019-05509-z
40. Nobre T (2019) Symbiosis in sustainable agriculture: can olive fruit fly bacterial microbiome be useful in pest management? Microorganisms. https://doi.org/10.3390/microorganisms7080238
41. Oh M, Zhang L (2020) DeepMicro: deep representation learning for disease prediction based on microbiome data. Sci Rep 10(1):6026. https://doi.org/10.1038/s41598-020-63159-5
42. Pavlidi N, Kampouraki A, Tseliou V, Wybouw N, Dermauw W, Roditakis E, Nauen R, Van Leeuwen T, Vontas J (2018) Molecular characterization of pyrethroid resistance in the olive fruit fly Bactrocera oleae. Pestic Biochem Physiol 148:1–7. https://doi.org/10.1016/j.pestbp.2018.03.011
43. Pasolli E, Truong DT, Malik F, Waldron L, Segata N (2016) Machine learning meta-analysis of large metagenomic datasets: tools and biological insights. PLoS Comput Biol 12(7):1004977. https://doi.org/10.1371/journal.pcbi.1004977
44. Pietrucci D, Teofani A, Unida V, Cerroni R, Biocca S, Stefani A, Desideri A (2020) Can gut microbiota be a good predictor for Parkinson's disease? A machine learning approach. Brain Sci 10(4):242. https://doi.org/10.3390/brainsci10040242
45. Radmanesh M, Ahmadi SH, Sepaskhah AR (2023) Measurement and simulation of irrigation performance in continuous and surge furrow irrigation using WinSRFR and SIRMOD models. Sci Rep. https://doi.org/10.1038/s41598-023-32842-8
46. Ridley EV, Wong AC-N, Westmiller S, Douglas AE (2012) Impact of the Resident Microbiota on the Nutritional Phenotype of Drosophila melanogaster. PLoS ONE 7(5):36765. https://doi.org/10.1371/journal.pone.0036765
47. Sundström JF, Albihn A, Boqvist S, Ljungvall K, Marstorp H, Martiin C, Nyberg K, Vågsholm I, Yuen J, Magnusson U (2014) Future threats to agricultural food production posed by environmental degradation, climate change, and animal and plant diseases - a risk analysis in three economic and climate settings. Food Secur 6(2):201–215. https://doi.org/10.1007/s12571-014-0331-y
48. Steinigeweg C, Alkassab AT, Erler S, Beims H, Wirtz IP, Richter D, Pistorius J (2022) Impact of a Microbial Pest Control Product Containing Bacillus thuringiensis on Brood Development and Gut Microbiota of Apis mellifera Worker Honey Bees. Microb Ecol. https://doi.org/10.1007/s00248-022-02004-w
49. Schulz N, Belheouane M, Dahmen B, Ruan VA, Specht HE, Dempfle A, Herpertz-Dahlmann B, Baines JF, Seitz J (2021) Gut microbiota alteration in adolescent anorexia nervosa does not normalize with short-term weight restoration. Int J Eat Disord 54(6):969–980. https://doi.org/10.1002/eat.23435
50. Schnepf E, Crickmore N, Van Rie J, Lereclus D, Baum J, Feitelson J, Zeigler DR, Dean DH (1998) Bacillus thuringiensis and Its Pesticidal Crystal Proteins. Microbiol Mol Biol Rev 62(3):775–806. https://doi.org/10.1128/MMBR.62.3.775-806.1998

51. Savary S, Ficke A, Aubertot J-N, Hollier C (2012) Crop losses due to diseases and their implications for global food production losses and food security. Food Secur 4(4):519–537. https://doi.org/10.1007/s12571-012-0200-5
52. Sivakala KK, Jose PA, Matan O, Zohar-Perez C, Nussinovitch A, Jurkevitch E (2021) In vivo predation and modification of the Mediterranean fruit fly Ceratitis capitata (Wiedemann) gut microbiome by the bacterial predator Bdellovibrio bacteriovorus. J Appl Microbiol 131(6):2971–2980. https://doi.org/10.1111/jam.15170
53. Su Q, Liu Q, Lau RI, Zhang J, Xu Z, Yeoh YK, Leung TWH, Tang W, Zhang L, Liang JQY, Yau YK, Zheng J, Liu C, Zhang M, Cheung CP, Ching JYL, Tun HM, Yu J, Chan FKL, Ng SC (2022) Faecal microbiome-based machine learning for multi-class disease diagnosis. Nat Commun 13(1):6818. https://doi.org/10.1038/s41467-022-34405-3
54. Somroo AA, Ur Rehman K, Zheng L, Cai M, Xiao X, Hu S, Mathys A, Gold M, Yu Z, Zhang J (2019) Influence of Lactobacillus buchneri on soybean curd residue co-conversion by black soldier fly larvae (Hermetia illucens) for food and feedstock production. Waste Manage 86:114–122. https://doi.org/10.1016/j.wasman.2019.01.022
55. Savary S, Willocquet L, Pethybridge SJ, Esker P, McRoberts N, Nelson A (2019) The global burden of pathogens and pests on major food crops. Nature Ecol Evol 3(3):430–439. https://doi.org/10.1038/s41559-018-0793-y
56. Shen Y, Xu J, Li Z, Huang Y, Yuan Y, Wang J, Zhang M, Hu S, Liang Y (2018) Analysis of gut microbiota diversity and auxiliary diagnosis as a biomarker in patients with schizophrenia: a cross-sectional study. Schizophr Res 197:470–477. https://doi.org/10.1016/j.schres.2018.01.002
57. Trinder M, Bisanz JE, Burton JP, Reid G (2015) Probiotic lactobacilli: a potential prophylactic treatment for reducing pesticide absorption in humans and wildlife. Beneficial Microbes 6(6):841–847. https://doi.org/10.3920/BM2015.0022
58. Turner JW, Cheng X, Saferin N, Yeo J-Y, Yang T, Joe B (2022) Gut microbiota of wild fish as reporters of compromised aquatic environments sleuthed through machine learning. Physiol Genom 54(5):177–185. https://doi.org/10.1152/physiolgenomics.00002.2022
59. Tang W, Wilkening J, Desai N, Gerlach W, Wilke A, Meyer F (2013) A scalable data analysis platform for metagenomics. In: 2013 IEEE international conference on big data, pp 21–26. https://doi.org/10.1109/BigData.2013.6691723
60. Van De Wouw M, Boehme M, Lyte JM, Wiley N, Strain C, O'Sullivan O, Clarke G, Stanton C, Dinan TG, Cryan JF (2018) Short-chain fatty acids: microbial metabolites that alleviate stress-induced brain-gut axis alterations: SCFAs alleviate stress-induced brain-gut axis alterations. J Physiol 596(20):4923–4944. https://doi.org/10.1113/JP276431
61. Vontas J, Hernández-Crespo P, Margaritopoulos JT, Ortego F, Feng H-T, Mathiopoulos KD, Hsu J-C (2011) Insecticide resistance in Tephritid flies. Pestic Biochem Physiol 100(3):199–205. https://doi.org/10.1016/j.pestbp.2011.04.004
62. Vargas R, Piñero J, Leblanc L (2015) An overview of pest species of bactrocera fruit flies (Diptera: Tephritidae) and the integration of biopesticides with other biological approaches for their management with a focus on the pacific region. Insects, pp 297–318 https://doi.org/10.3390/insects6020297
63. Wang Z, Bovik AC (2009) Mean squared error: love it or leave it? A new look at signal fidelity measures. IEEE Signal Process Mag 26(1):98–117. https://doi.org/10.1109/MSP.2008.930649
64. Yoon BW, Lim S-H, Shin JH, Lee J-W, Lee Y, Seo JH (2021) Analysis of oral microbiome in glaucoma patients using machine learning prediction models. J Oral Microbiol 13(1):1962125. https://doi.org/10.1080/20002297.2021.1962125
65. Yang F, Tomberlin JK, Jordan HR (2021) Starvation alters gut microbiome in black soldier fly (Diptera: Stratiomyidae) larvae. Front Microbiol 12:601253. https://doi.org/10.3389/fmicb.2021.601253
66. Yu C, Zhou Z, Liu B, Yao D, Huang Y, Wang P, Li Y (2023) Investigation of trends in gut microbiome associated with colorectal cancer using machine learning. Front Oncol 13:1077922. https://doi.org/10.3389/fonc.2023.1077922
67. Zhang W, Jiang F, Ou J (2011) Global pesticide consumption and pollution: with china as a focus. Proc Int Acad Ecol Environ Sci 1(2):125
68. Zhuang Z, Yang R, Wang W, Qi L, Huang T (2020) Associations between gut microbiota and Alzheimer's disease, major depressive disorder, and schizophrenia. J Neuroinflamm 17(1):288. https://doi.org/10.1186/s12974-020-01961-8
69. Zhao X, Zhang X, Chen Z, Wang Z, Lu Y, Cheng D (2018) The divergence in bacterial components associated with bactrocera dorsalis across developmental stages. Front Microbiol 9:114. https://doi.org/10.3389/fmicb.2018.00114

Terms and Conditions

Springer Nature journal content, brought to you courtesy of Springer Nature Customer Service Center GmbH (“Springer Nature”).
Springer Nature supports a reasonable amount of sharing of research papers by authors, subscribers and authorised users (“Users”), for small-scale personal, non-commercial use provided that all copyright, trade and service marks and other proprietary notices are maintained. By accessing, sharing, receiving or otherwise using the Springer Nature journal content you agree to these terms of use (“Terms”). For these purposes, Springer Nature considers academic use (by researchers and students) to be non-commercial.
These Terms are supplementary and will apply in addition to any applicable website terms and conditions, a relevant site licence or a personal subscription. These Terms will prevail over any conflict or ambiguity with regards to the relevant terms, a site licence or a personal subscription (to the extent of the conflict or ambiguity only). For Creative Commons-licensed articles, the terms of the Creative Commons license used will apply.
We collect and use personal data to provide access to the Springer Nature journal content. We may also use these personal data internally within ResearchGate and Springer Nature and as agreed share it, in an anonymised way, for purposes of tracking, analysis and reporting. We will not otherwise disclose your personal data outside the ResearchGate or the Springer Nature group of companies unless we have your permission as detailed in the Privacy Policy.
While Users may use the Springer Nature journal content for small scale, personal non-commercial use, it is important to note that Users may not:

1. use such content for the purpose of providing other users with access on a regular or large scale basis or as a means to circumvent access control;
2. use such content where to do so would be considered a criminal or statutory offence in any jurisdiction, or gives rise to civil liability, or is otherwise unlawful;
3. falsely or misleadingly imply or suggest endorsement, approval , sponsorship, or association unless explicitly agreed to by Springer Nature in writing;
4. use bots or other automated methods to access the content or redirect messages
5. override any security feature or exclusionary protocol; or
6. share the content in order to create substitute for Springer Nature products or services or a systematic database of Springer Nature journal content.

In line with the restriction against commercial use, Springer Nature does not permit the creation of a product or service that creates revenue, royalties, rent or income from our content or its inclusion as part of a paid for service or for other commercial gain. Springer Nature journal content cannot be used for inter-library loans and librarians may not upload Springer Nature journal content on a large scale into their, or any other, institutional repository.
These terms of use are reviewed regularly and may be amended at any time. Springer Nature is not obligated to publish any information or content on this website and may remove it or features or functionality at our sole discretion, at any time with or without notice. Springer Nature may revoke this licence to you at any time and remove access to any copies of the Springer Nature journal content which have been saved.
To the fullest extent permitted by law, Springer Nature makes no warranties, representations or guarantees to Users, either express or implied with respect to the Springer nature journal content and all parties disclaim and waive any implied warranties or warranties imposed by law, including merchantability or fitness for any particular purpose.
Please note that these rights do not automatically extend to content, data or other material published by Springer Nature that may be licensed from third parties.
If you would like to use or distribute our Springer Nature journal content to a wider audience or on a regular basis or in any other manner not expressly permitted by these Terms, please contact Springer Nature at

onlineservice@springernature.com