









Infection profiles in a wild rat–protozoan network are shaped by host traits and environmental factors

Matan Markfeld ^{1,†}, Itamar Talpaz ^{1,†}, Barry Biton ¹, Toky Maheriniaina Randriamoria², Voahangy Soarimalala ^{2,3}, Steven M. Goodman ^{2,4}, Charles L. Nunn ^{5,6}, Georgia Titcomb ^{7,*}, and Shai Pilosof ^{1,8,*}

¹Department of Life Sciences, Ben-Gurion University of the Negev, Be'er-Sheva, Israel

²Association Vahatra, Antananarivo, Madagascar

³Institut des Sciences et Techniques de l'Environnement, Université de Fianarantsoa, Fianarantsoa, Madagascar

⁴Field Museum of Natural History, Chicago, IL, USA

⁵Department of Evolutionary Anthropology, Duke University, Durham, NC, USA

⁶Duke Global Health Institute, Durham, NC, USA

⁷Department of Fish, Wildlife, and Conservation Biology, Colorado State University, Fort Collins, CO, USA

⁸The Goldman Sonnenfeldt School of Sustainability and Climate Change, Ben-Gurion University of the Negev, Be'er Sheva, Israel

[†]Equal contribution

*Corresponding authors: pilos@post.bgu.ac.il, georgia.titcomb@colostate.edu

Abstract

Heterogeneity in parasite infection among hosts shapes transmission dynamics and spillover risk to other host species but remains poorly understood in natural systems. We applied network-based stochastic block modeling and machine learning to a uniquely rich dataset to identify and predict protozoan infection profiles in introduced black rats (*Rattus rattus*) sampled along an environmental gradient in Madagascar. Three host infection profiles emerged, differing in parasite richness and composition, revealing distinct host roles in transmission. Predictive models incorporating host traits (e.g., body mass, microbiome composition) and environmental variables (e.g., population density, habitat structure) accurately classified hosts into profiles, with host traits contributing to predictions 40% more than environmental features. Our findings show how intrinsic and extrinsic factors jointly structure individual-level infection heterogeneity and underscore the value of infection profiles for understanding host–parasite dynamics. Our integrative approach offers a framework for predicting infection risk at human–animal interfaces where zoonotic pathogens circulate.

Introduction

Heterogeneity in infection is a defining feature of host–parasite systems, with parasite distributions typically skewed: most hosts harbor few parasites, while a small subset carries heavy burdens (1,2). This variation, driven by differences in host physiology and environmental context, has major consequences for transmission dynamics, infection persistence, and spillover risk (3–7). Further complexity arises from co-infections involving multiple parasite species or strains, each with distinct natural history traits and transmission strategies (8–10). Understanding these interacting sources of heterogeneity is essential for predicting and mitigating parasite spread and impact (11). However, the combined influence of host traits, co-infections, and environmental factors on patterns of parasite infection heterogeneity remains poorly understood, particularly at the individual host level.

Bipartite networks, where parasites are linked to the hosts they infect, offer a powerful framework for studying heterogeneity in host–parasite interactions (12). However, studies to date have largely focused on species-level networks to explore ecological and evolutionary processes underlying heterogeneity (13–16), with only a limited use of individual-level networks (17). Individual-level networks can be used to identify groups of individual hosts with similar parasite associations and, conversely, groups of parasites infecting hosts with similar characteristics. These emergent group structures, which we term *infection profiles*, can provide insights into the ecological and epidemiological roles of hosts and parasites (18,19). For example, distinguishing generalist parasites from rare or highly specialized ones (20), or identifying host groups with distinct parasite assemblages.

Detecting infection profiles involves clustering nodes with similar interaction patterns, which can be done using stochastic block modeling (SBM) (19). Unlike clustering methods that emphasize dense intra-group links (e.g., modularity), SBMs identify latent group structures based on connection probabilities within and between blocks, allowing detection of nodes that have similar probabilities of connecting to nodes in other groups (18,19). Despite their potential (21), SBMs are rarely used in ecological analysis (22–25). However, such previous studies confirmed their usefulness. For example, in the human gut microbiome, SBMs identified generalist and specialist microbes (23). To date, however, SBMs have not, to our knowledge, been applied to host–parasite networks.

Here, we identify infection profiles in individual hosts of the introduced black rat (*Rattus rattus*) and their protozoan parasites in northeastern Madagascar. Protozoa commonly inhabit the mammalian gut, yet few studies have explored their diversity in wild mammals or the factors influencing their community structure (26). Detecting infection profiles is particularly important in rural, low-income regions like Madagascar, where introduced rats, living both in the wild and near human settlements, serve as reservoirs for zoonotic pathogens, increasing the risk of spillover to humans (27,28). For example, in Madagascar *R. rattus* was previously found to have an infection rate of ~20–50% for protozoa of the zoonotic genera *Trypanosoma*, *Cryptosporidium*, and *Giardia*, which have also been detected in humans on the island (28,29).

While infection profiles can reveal structural heterogeneity in host–parasite interactions, they do not explain how it emerges. To gain mechanistic insight to these interactions, our second goal is to identify the intrinsic and extrinsic factors shaping infection profiles by focusing on the two main phases of infection: exposure (the likelihood of encountering a parasite) and susceptibility (the likelihood of infection post-exposure) (30–32). Intrinsic host traits such as sex, age, body mass, immune function, and co-infections can influence both stages by affecting host

behavior, immunity, and survival (33–35). The host microbiome is also a key modulator, shaping susceptibility via its effects on immunity, metabolism, and overall health (36,37). Extrinsic factors, including land-use change, habitat fragmentation, and environmental reservoirs, alter host movement, contact rates, and exposure risk (38–41). For example, gut protozoa, transmitted via fecal-oral routes, are particularly sensitive to environmental contamination, which varies in different ecological contexts (42,43). These extrinsic pressures often interact with intrinsic traits—for instance, through effects on diet, body condition, or stress. However, how such interactions drive infection heterogeneity remains poorly understood, especially in wild or semi-wild mammal populations.

We combined SBM and machine learning tools with detailed field sampling of ecological and biological data to detect infection profiles and assess the key drivers of infection heterogeneity. Our dataset is unusually rich and detailed, containing diverse host traits (e.g., body mass, sex, nematode co-infection, and gut microbiome composition) and environmental factors (e.g., habitat characteristics, disturbance, and community composition). This allows us to test multiple host and environmental drivers simultaneously, an approach rarely possible in most systems. By linking these factors to SBM-identified infection profiles, we disentangle the relative contributions of host traits and environmental conditions to emerging heterogeneity in parasite infection patterns (Figure 1).

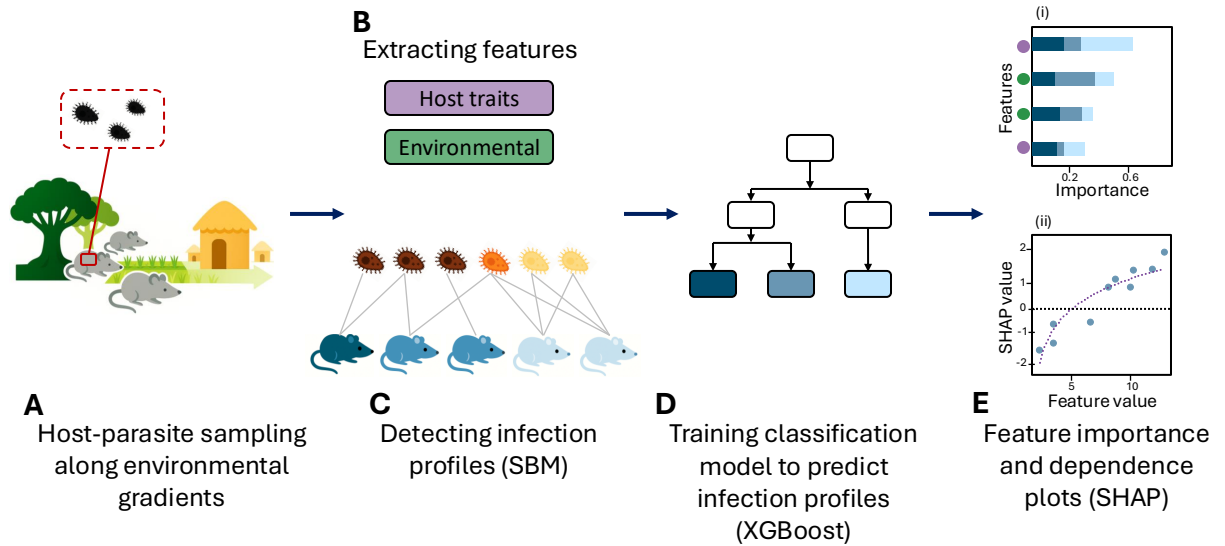


Figure 1: Conceptual scheme of methods. (A) *Rattus rattus* individuals and their protozoan parasites were sampled along a land-use change gradient to investigate infection heterogeneity. (B) Host traits (e.g., mass, sex, nematode co-infection status, and microbiome composition) and environmental features (e.g., habitat attributes, population density) were measured for each host. (C) We used Stochastic Block Modeling (SBM) to identify host and parasite infection profiles based on interaction patterns in the host-parasite network. Colors indicate SBM-assigned group membership of protozoa OTUs and rat hosts. (D) We trained an XGBoost model on the host and environmental features to quantify their relative importance in predicting infection profiles. (E) (i) We used SHAP (SHapley Additive exPlanations) values to estimate the mean absolute contribution of each host-trait (purple circles) and environmental (green circles) feature by measuring changes in model predictions upon feature removal (44). (ii) The relationship between features (e.g., host mass) and SHAP values for each infection profile was assessed. Positive SHAP values indicate a feature that supports classification into an infection profile, while negative values indicate the opposite.

Results

Our study focused on the black rat (*Rattus rattus*), a species introduced to Madagascar probably during the 10th-century and the most abundant small mammal in non-forested, rural areas on the island (45–47). As a large generalist omnivore with a highly adaptable diet (48), *R. rattus* is a primary agricultural pest, a major disruptor of native ecological communities, and a potential vector for zoonotic diseases in the area (27,47). Rats were trapped in three different villages and across seven habitat types, ranging from semi-intact moist evergreen forest, secondary grasslands, zones of agroforestry and agricultural areas, to human settlements, using standardized trapping grids and pitfall lines over three seasonal replicates. We used DNA metabarcoding to detect a range of protozoan operational taxonomic units (OTUs) in each rat, identify nematode infections, and characterize the rat gut microbiome.

We constructed a bipartite network linking individual rats to protozoan OTUs. OTUs were identified using 18S rDNA-based metagenomics, as species-level resolution is often unattainable for protozoa. This approach enables exploration of the distinct functional roles and impacts of different protozoan groups on the host (49). We also assigned the lowest taxonomic identities possible to OTUs to complement the functional perspective with traditional classifications. Another novel aspect of our analysis is the inclusion of hosts without detectable protozoan infections. Host-parasite network studies exclude uninfected individuals because the lack of a link could be a false negative. Our primer set amplifies both parasitic and non-parasitic eukaryotic DNA, allowing us to designate rats with only non-parasitic protozoa as “uninfected”. We still acknowledge that, as in any study, very low-abundance parasite infections might escape detection (see *Materials and methods*). The rat–protozoa bipartite network included 841 host nodes—271 uninfected (singleton nodes with no links) and 570 infected—and 41 protozoan OTU nodes. The network contained 1,557 links, with a connectance of 0.045 (i.e., 4.5% of all possible links were realized).

Stochastic block modeling reveals structured host and protozoan profiles

We used a stochastic block model (SBM) to cluster hosts and parasites into groups with similar interaction patterns. The bipartite SBM estimates the probability of infection between host and parasite groups, identifying blocks of nodes with similar infection probabilities. Because SBM is based on links, we predefined uninfected hosts as a separate group and conducted the SBM analysis exclusively on the infected hosts.

We identified two host groups based on the SBM analysis. Together with the uninfected group, this resulted in three *host infection profiles*. We also identified seven *protozoan infection profiles* (**Figure 2A**, **Figure S1**). The host infection profiles varied in size, with 271 individuals classified into the first profile, 205 into the second, and 365 into the third. Protozoan infection profile size ranged from a single OTU to 21 OTUs. We also found variation in host node degree. Specifically, rats from the first profile were infected by 0 protozoa OTUs (the uninfected profile), rats from the second profile were infected by 1–7 (mean = 1.61) OTUs, and rats from the third profile were infected by 1–10 (mean = 3.36) OTUs (**Figure 2B**). The protozoan infection profiles also varied in degree, as OTUs from profiles 1, 5, and 6 (all are of the genera *Tritrichomonas* or *Hypotrichomonas*) were more prevalent, infecting 15–36% of hosts, whereas OTUs from profiles 2–4 and 7 were rarer, infecting only 0.3–8% of hosts (**Figure 2C**, **Figure S2**).

The host infection profiles differed not only in node degree but also in connectivity patterns, as captured by the block connectivity matrix Θ (**Figure 2A**). Hosts in profile 1 were uninfected and showed no associations with any protozoan OTUs. In contrast, host infection profile 2

was characterized by infection with the prevalent *Hypotrichomonas* OTU. This was indicated by strong connectivity to protozoan profile 5, along with sporadic infections with diverse, low-prevalence OTUs (protozoan profiles 4 and 7). Host infection profile 3 exhibited higher overall infection levels, with particularly strong associations to the highly prevalent *Tritrichomonas* OTUs (protozoan profiles 1, 2, and 6).

Heterogeneous connectivity patterns were also found for the protozoa profiles, as some maintained a consistent low infection rate (e.g., profile 7), while others exhibited strong preferences to particular host infection profiles (e.g., parasite profiles 1, 5, and 6). Several protozoan profiles consisted of specific taxa (profiles 1-3, 5, 6, and 8 were mostly *Tritrichomonas* and *Hypotrichomonas*), while profiles 4 and 7 contained diverse taxa.

Overall, these distinct connectivity patterns demonstrate that the SBM effectively captured infection profiles, revealing latent groups in host-parasite interactions.

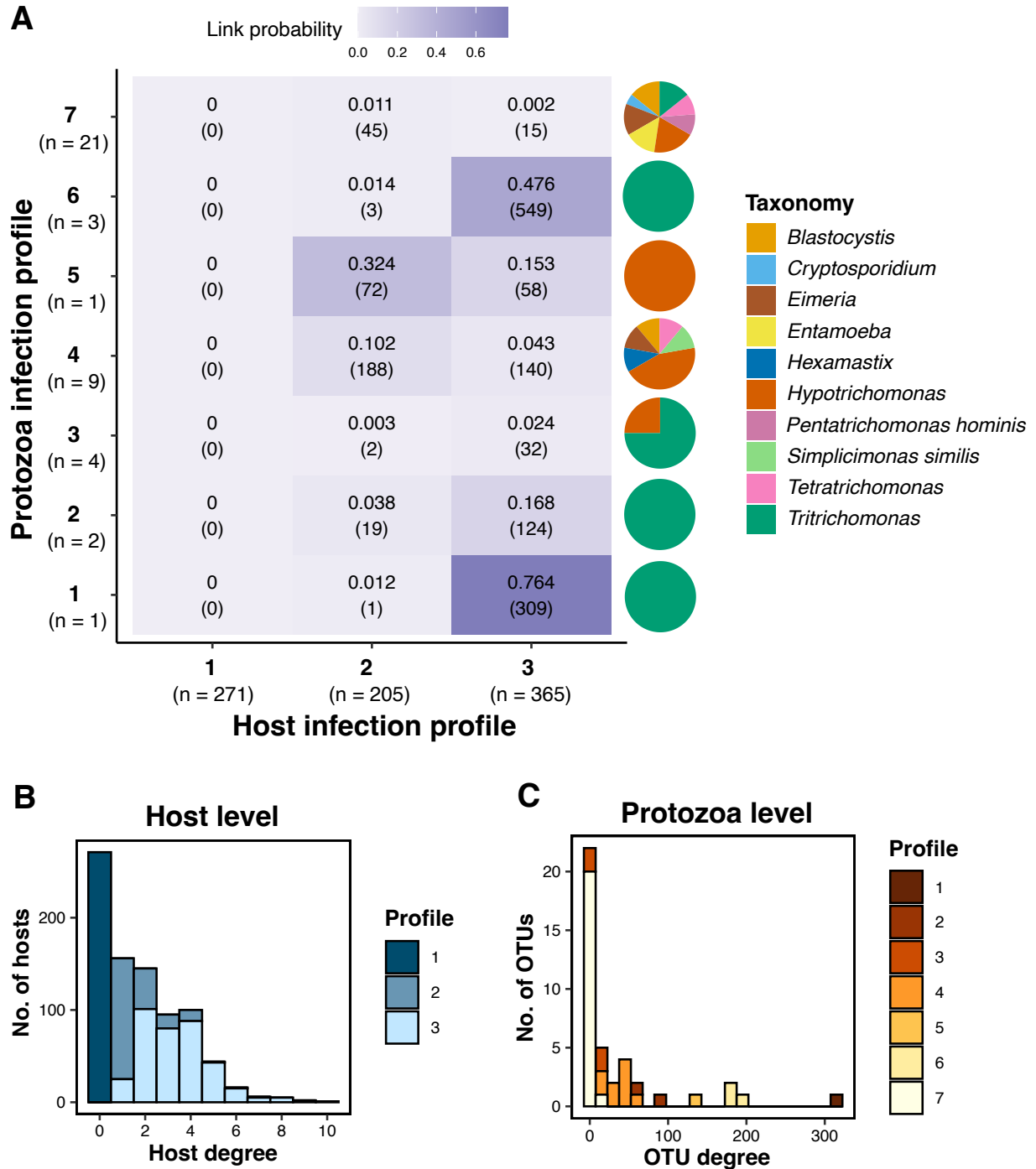


Figure 2: Infection profiles of *Rattus rattus* individual hosts and protozoan OTUs. (A) The block connectivity matrix Θ resulting from the stochastic block model (SBM) analysis of the rat-protozoa network. Rows and columns represent protozoa and host groups, respectively (n indicates the number of OTUs or hosts in each group). Cell color reflects the SBM-predicted link probability between a host and an OTU that belong to certain groups. The link probability is denoted in each cell and the total number of observed links between groups is shown in parentheses. Pie charts depict the proportion of OTUs from each protozoan taxa (at the species/genus level) within each protozoan infection profile. **(B)** Host degree distribution (protozoan OTU richness with which a host is infected). **(C)** Protozoa OTU degree distribution (number of host individuals an OTU infects). Colors indicate protozoa and host infection profiles (SBM groups).

Host traits outweigh environmental features in predicting host infection profiles

We trained an XGBoost multi-classification model to predict host infection profile membership. Our analysis considered class imbalance (uneven sizes of the three host infection profiles). We predicted host infection using both host and environmental variables as features. These variables were selected based on their known influence on host exposure to parasites and susceptibility to infection (see Feature collection and processing in *Materials and methods* for details; **Table 1**). Host variables included body mass, body condition, sex, and age class—traits that affect behavior and immune function. We also included nematode co-infection status and gut microbiome composition (relative abundance of four key microbial families), both of which can modulate host immunity and infection outcomes ([35,37](#)). Environmental variables included habitat type (derived via vegetation PCA to capture habitat structure), distance to the nearest village center (a proxy for human disturbance), and small mammal community composition (densities of *R. rattus*, native small mammals species, and other non-native species, including *Suncus* shrews and house mice *Mus musculus*). These factors influence environmental exposure to parasites by shaping habitat conditions, contact rates, and parasite persistence in the environment ([40,41](#)). Together, these variables capture host-level intrinsic and extrinsic sources of variation likely to shape infection profiles.

Our model consistently outperformed a no-skill classifier in predicting host profiles (weighted precision = 0.53, weighted recall = 0.54, weighted F1-score = 0.53, weighted balanced accuracy = 0.64, and Matthews correlation coefficient (MCC) = 0.28; **SI notes 3 and 4**). To evaluate the model’s ability to distinguish among profiles, we further assessed one-vs-all performance across decision thresholds. The Area Under the Receiver Operating Characteristic Curve (AUC-ROC) exceeded the random expectation of 0.5 for all profiles (AUC = 0.726, 0.606, 0.738, for profiles 1, 2, and 3, respectively). Similarly, the Precision-Recall Curve was above the no-skill baselines for the three host profiles (AUC = 0.548 [no-skill of 0.322] for profile 1; AUC = 0.331 [0.244] for profile 2; and AUC = 0.656 [0.434] for profile 3). Therefore, we can predict infection profiles based on the features we selected.

However, not all features contribute equally to prediction. To identify the features driving model predictions of host infection profiles, we used SHapley Additive exPlanations (SHAP) analysis, which quantifies each feature’s contribution to classification outcomes ([44](#)). To assess the relative importance of host traits versus environmental factors, we summed absolute SHAP values within each category. While both categories of features influenced model predictions, host traits’ features contributed 40% more than environmental features (1.24 compared to 0.88 mean absolute SHAP) (**Figure 3A**). This trend was consistent across all host infection profiles, with host features consistently showing higher absolute SHAP values. However, for host profile 3, the environmental features contributed almost as much as the host traits. The pattern held even when we considered only the top six host features to match the number of environmental features, with host features still exhibiting higher mean absolute SHAP values (1.16 vs. 0.88).

The most influential host features included host body mass and the relative abundance of the gut microbial families *Prevotellaceae* and *Muribaculaceae* (**Figure 3B**). Body mass is known to affect infection risk, and it is often linked to body condition, age, and sex ([50–54](#)). However, these later features ranked low in their importance. Salient environmental features included rat and other non-native species densities at a site, together with the site’s vegetation structure (PC1) (**Figure 3B**). Notably, the relative importance of all features varied among host infection profiles, with some playing a crucial role in predicting certain profiles while being less relevant for others. For example, rat density was an important feature in predicting host infection profiles

1 and 3, whereas its contribution to the prediction of profile 2 was lower (**Figure 3B**).

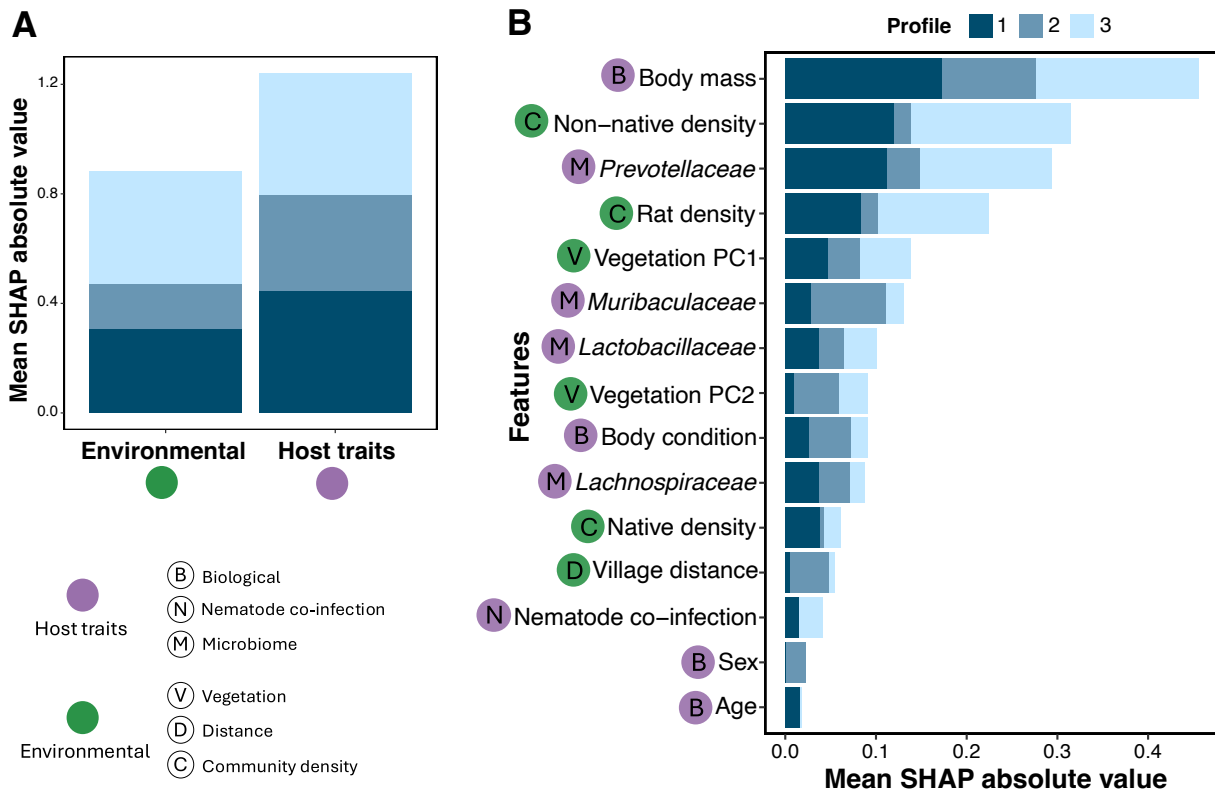


Figure 3: Importance of host traits and environmental variables in predicting host infection profiles. (A) Mean absolute SHAP values for host and environmental features across all host infection profiles, highlighting their relative contributions. (B) Feature importance, measured as the mean absolute SHAP values across hosts for all infection profiles. Features are ranked from most important (top) to least important (bottom). In both panels, bar colors depict host infection profiles. Circles indicate host (purple) and environmental (green) features, with the letter inside each circle depicting the corresponding sub-category of the feature per the legend at the bottom-left (**Table 1**).

Host traits and ecological gradients structure host infection profiles

To further investigate how the most important features influence the prediction of specific host infection profiles, we used SHAP dependency plots, which visualize the relationship between feature values and SHAP values. A positive SHAP value for a given infection profile indicates that the model is more likely to classify hosts with those specific feature values into that profile, whereas a negative SHAP value suggests a lower probability of classification into that profile (44).

For most features, we observed clear trends between feature values and model predictions of infection profiles (**Figure 4**, **Figure S4**). For example, rats with a body mass below 100 g were consistently classified into infection profile 1, which is associated with no infection, whereas larger rats were more often assigned to host profile 3, characterized by higher infection richness. Particularly interesting features were the relative abundance of the bacterial families *Prevotellaceae* and *Muribaculaceae*. A high relative abundance of *Prevotellaceae* in a host was influential in distinguishing host profile 3 (high infection). In contrast, a high relative abundance of *Muribaculaceae* was associated with host profile 2 (low infection) (**Figure 4**). These bacterial families have been shown to play key roles in digestion, immune regulation, and gut health

(55–57).

Among the environmental features, higher non-native species and rat densities were associated with a marked shift in predicted host infection profiles from 1 (no infection) to 3 (higher infection richness). This pattern was also evident along the vegetation gradient: negative values of vegetation PC1—indicative of less disturbed sites—were more associated with host profiles 1 and 2, and less so with profile 3. However, some features showed weak or no clear trends with specific infection profiles (e.g., distance from the village center; **Figure S4**).

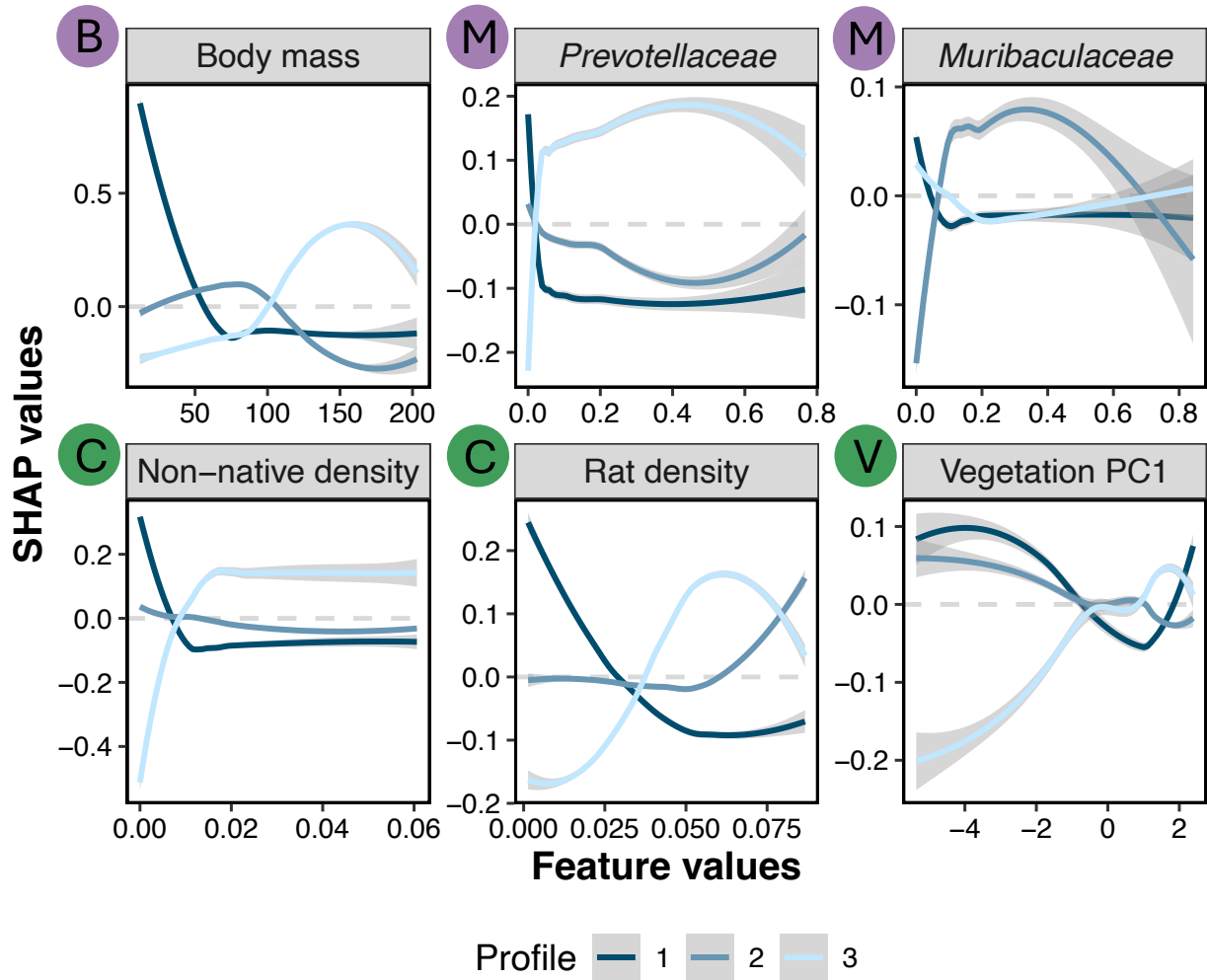


Figure 4: Key host traits and environmental factors predict infection profiles with distinct trends in SHAP values. SHAP dependency plots show individual hosts' SHAP values (y-axis) as a function of feature values (x-axis). For clarity, we visualized only the LOESS-derived trend line and its associated confidence interval for the 841 host samples, rather than displaying all individual host data points. Line colors indicate host infection profiles. The dashed gray line marks a SHAP value of zero (i.e., no predictive value). LOESS line above and below 0 indicates that the classification of a profile is more or less likely, respectively. The top and bottom rows present the three most important host and environmental features, respectively ([B] = Biological; [M] = Microbiome; [C] = Community density; [V] = Vegetation).

Discussion

Understanding what drives infection heterogeneity is key to predicting parasite transmission and spillover risk. However, identifying patterns in infection heterogeneity and its underlying factors remains challenging. By combining stochastic block modeling (SBM) and machine learning with detailed field data on the introduced *Rattus rattus* in Madagascar, we identified three distinct host infection profiles—ranging from low infection and weak protozoan associations to high infection. Host traits were 40% more important than environmental factors in shaping these profiles, with body mass, gut microbiome composition, and small mammal community density being the strongest predictors.

Our approach offers new insights into the processes underlying host–parasite interactions by identifying infection profiles that capture meso-scale network structure: groupings of hosts based on the composition and connectivity of their parasite assemblages. Unlike host-parasite pairwise analysis, infection profile detection accounts for co-infection patterns and variation in parasite generalism, which are ubiquitous in nature (58,59). For instance, hosts in profiles 2 and 3 differed not only in parasite richness but also in parasite composition, with a shift from highly prevalent *Tritrichomonas* OTUs in profile 3 to less prevalent *Hypotrichomonas* and other rare taxa in profile 2. These infection profiles also offer predictive insights into host heterogeneity, revealing epidemiologically relevant differences. For example, Hosts in profile 3 were associated with generalist parasites and may act as superspreaders, while those in profile 2 had lower infection rates but harbored a greater diversity of parasite taxa. These contrasting patterns suggest distinct roles in infection dynamics and spillover risk. Supporting this, profile 3 was more strongly linked to *Tritrichomonas*, whereas profile 2 was associated with *Blastocystis* and *Eimeria*. Those three protozoan taxa were detected in local human populations (unpublished data; (60)). Thus, the relevance of each profile to potential spillover events may differ.

Body mass was the dominant predictor of infection heterogeneity. Body mass is a well-established correlate of infection risk, often serving as a proxy for body condition, which can influence immune function and physiological resilience to parasites (50). It also correlates with age and sex, since younger and female rats tend to be smaller. Age influences infection patterns because younger individuals have had less time to accumulate parasites but may be more susceptible to initial infections (52), whereas male behavioral traits—greater movement, territoriality, and aggression—can increase exposure and infection rates (54). Because body mass correlates with age, sex, and body condition, it likely integrates multiple underlying physiological processes (e.g., immune maturity, energy reserves) that drive parasite susceptibility. None of those traits emerged as top predictors on their own—possibly because their signal was outweighed by mass in the model. Future work that disentangles how age, sex, and body condition independently influence parasitism will help clarify which specific processes underlie the emergence of infection profiles.

Gut microbiome composition was also an important feature. While the microbiome is closely tied to host health, the direction of causality remains unclear: dysbiosis (an imbalance in microbial communities) may increase susceptibility, or alternatively, parasite infection itself can disrupt the microbiome and trigger dysbiosis (61,62). Moreover, microbiome composition shifts along environmental gradients such as land-use change, potentially influencing host–parasite dynamics (36,63). Therefore, regardless of causality, our findings highlight the importance of jointly considering microbiome and environmental factors, and suggest that the microbiome may serve as a useful indicator of infection heterogeneity in wild populations. Notably, unlike the microbiome, nematode co-infection did not emerge as an important predictor, although it is

known to modulate host immunity and alter susceptibility to other parasites (64,65). More detailed research into how nematodes and protozoa jointly shape infection profiles is needed, particularly because their transmission pathways are similar.

Among the environmental variables, the most influential features were small mammal population densities. These community-level factors played a key role in shaping infection profiles, particularly in the transition from no infection (host profile 1) in more natural habitats to higher infection rates (host profile 3), which were more common in disturbed environments (e.g., villages, rice fields, and agroforests). Higher densities of rats and other non-native species were associated with these disturbed sites (**Figure S7**), suggesting that increased host abundance and altered community composition facilitate parasite transmission. One mechanism by which small mammal density can affect infection dynamics is through density dependence. Denser populations lead to more frequent contact events and greater environmental contamination (38), especially for gut protozoa that rely on fecal-oral transmission (42). While vegetation and distance from the village were also included as proxies for land-use change, they were less predictive than small mammal density (**Figure S4**). This suggests that the primary mechanism through which land-use change influences infection is not habitat structure per se, but rather its impact on the abundance and composition of the local host community (66).

Our study advances understanding of how host traits and environmental variation interact to shape infection heterogeneity. Nevertheless, several limitations highlight directions for future research. First, we were unable to disentangle the relative contribution of host susceptibility (likelihood of infection after exposure) from exposure (likelihood of encountering parasites) (32). Many predictors, such as gut microbiome composition, likely reflect both processes: the microbiome can influence immune function and infection risk (67,68), but could also reflect shifts with diet and environment, serving as a proxy for exposure (63,69). Disentangling these processes will require experimental or longitudinal designs. Second, our model showed limited predictive performance, particularly for the low-infection host profile 2. This is likely in part due to class imbalance, as profile 2 comprised the smallest proportion of hosts (24.3%). Additionally, our model does not explicitly capture stochasticity in infection dynamics (70). However, among the deterministic factors considered, we were able to distinguish those with greater predictive importance. Improving model performance and biological interpretability may require increased sampling effort and the inclusion of higher-resolution data—such as parasite load, behavioral metrics, and immunological markers—which can also clarify underlying mechanisms.

In summary, our study demonstrates that infection profiles provide a powerful framework for uncovering structured heterogeneity in host-parasite interactions at the individual level. By integrating network-based approaches with machine learning and rich ecological and biological data, we quantified the relative contributions of intrinsic host traits and extrinsic factors in shaping protozoan infection patterns of rat hosts along a land-use gradient. This approach moves beyond traditional analyses, capturing complex co-infection dynamics and highlighting the functional roles of different host groups in parasite transmission. As anthropogenic disturbance continues to reshape host communities and hence parasite dynamics and spillover risk, such integrative frameworks will be critical for advancing predictive models of infection and informing strategies for surveillance and intervention.

Materials and Methods

Study site and sampling

Rattus rattus were collected in the vicinity of three villages (Mandena, Sarahandrano, Andatsakala) in the SAVA Region of northeastern Madagascar, in the area surrounding Marojejy National Park (**SI note 1**). The park consists of natural moist evergreen forests spanning a wide elevation range, from lowland areas to mountain peaks exceeding 2000 m. At each of the three villages, small mammals were sampled across seven habitat types (sites) representing a degradation gradient: (1) semi-intact natural forest inside the national park, (2) secondary forest, (3) *savoka* (brushy regrowth), (4) agroforest (vanilla plantation), (5) mixed agriculture (sugarcane/coffee plantation), (6) flooded rice fields, and (7) the village itself. Sites in each village setting were located approximately 500 m apart. At each site, a grid of 121 live traps (arranged in an 11×11 configuration with 10 m spacing) was deployed, supplemented by two pitfall trap lines outside the grid, each containing 11 buckets. During the study period, each site was sampled for six consecutive nights during three different seasons. Sampling was conducted at Mandena between October 2019 and September 2020, at Sarahandrano between November 2020 and September 2021, and at Andatsakala between October 2021 and August 2022.

Protozoa DNA extraction and lab work

We used DNA metabarcoding to detect a range of protozoa in rats, identify nematode infections, and characterize the rat microbiome. Approximately 1g of feces was preserved in either nucleic acid preservation (NAP) buffer ([71](#)) or Zymo DNA/RNA Shield (Zymo Research, Irvine, California). Two different storage solutions were used due to complications with lab supplies during the COVID-19 pandemic. Mean sequencing read abundance did not differ between the two sample types. DNA was extracted from fecal samples using Zymo MiniPrep Fecal kits (Zymo Research, Irvine, California) according to manufacturer directions.

We performed PCR with the G4 primer set ([72,73](#)) to amplify 18S ribosomal DNA from a wide range of eukaryotes in the rat fecal DNA extracts. Forward and reverse primers contained 8-nucleotide barcodes with a Hamming distance of at least 4. PCR reactions were carried out in 15µL volumes consisting of: 3 µL of each forward and reverse primer (2 µM stock concentration); 7 µL from a Mastermix comprised of 0.7 µL of Amplitaq Gold polymerase, 150 µL MgCl₂, 150 µL Amplitaq Gold buffer, 12 µL BSA, 6 µL DMSO, and 344 µL water; and up to 2 µL template DNA (1–100 ng total). Cycling conditions were: 10-minute hot-start activation, 35x cycles of 15 s at 95°C, 30 s at 57°C, 40 s at 72°C, and a final 5-min extension at 72°C. DNA concentrations were then measured, pooled, normalized, and purified using MinElute columns prior to multiplexing with additional libraries. The final library for each village was sequenced three times on an Illumina MiSeq (v3 2 × 300 bp, 25 M reads) at the UC Davis Genome Center. Sequences were demultiplexed using cutadapt (v.3.4) with zero error tolerance ([74](#)). We used the *dada2* bioinformatics pipeline ([75](#)) to filter and trim amplicons (minimum length = 100, 15% PhiX removed), remove errors, dereplicate, infer amplicon sequence variants (ASVs) using the pseudo-pooling method, merge pairs, remove chimeras, and combine the three ASV read tables from the different villages into one table.

Bioinformatics and protozoa OTU processing

We calculated the relative read abundance of each ASV and excluded reads that accounted for less than 1% of a sample’s relative read abundance to avoid potential sequencing errors or tag jumps. We excluded any sample with fewer than 500 total reads due to potential amplification

or sequencing failure (n=38). Due to less certainty in protozoal identifications, we then used a consensus approach to assign taxonomy to ASVs: we used both the 'assignTaxonomy' function in *dada2* (minimum bootstraps = 50) and the 'IdTaxa' function in the *DECIPHER* package in R (76) to generate two identifications for each ASV using the SILVA non-redundant database clustered at 99% similarity (v.132). For any ASV with a mismatching identification, we queried the sequence in the NCBI GenBank database to assign a final 'consensus' ID.

Next, we clustered phylogenetically similar ASVs into OTUs at 97% similarity using the 'Clustered' function from the *DECIPHER* package. Taxonomy was assigned to each OTU based on its most common ASV. Because the G4 primer set amplifies both parasitic and non-parasitic eukaryotic DNA, we then manually filtered all OTUs to those that are known or suspected parasites of mammals. Consequently, our protozoan community represents OTUs with small genetic variations that may influence their pathogenic traits. In total, our dataset includes 841 individual rat hosts and 41 protozoan OTUs spanning 10 genera and 4 phyla (**Figure S2**).

Network construction and detection of infection profiles

We constructed a bipartite network representing individual rat hosts and protozoa OTUs, where an edge is present (1) if an OTU infects a host and absent (0) otherwise (**Figure 1C**). Then, we identified infection profiles using an SBM (77). In the bipartite version of the SBM, hosts are assigned to $Q^{(1)}$ groups and protozoa parasites to $Q^{(2)}$ groups. The interactions between these groups are governed by a block connectivity matrix Θ , which encodes the probability of infection between each pair of host and parasite blocks (77). Consequently, the probability that a link exists between a host i and a parasite j , which belong to groups c_x and c_y respectively, is $P_{ij} = \Theta_{c_x c_y}$. This structure implies that nodes within the same block are "stochastically equivalent" and thus have similar probabilities of being infected by parasites from a given parasite block.

We implemented SBM using a commonly used Variational Expectation-Maximization (VEM) algorithm, which iteratively estimates latent memberships and model parameters (77). The algorithm partitions the network into groups and calculates the likelihood of such clustering, considering the membership of nodes in groups. The algorithm consists of two main steps: (1) updating the posterior probabilities of host and parasite assignments to latent blocks and (2) optimizing model parameters—including the block connectivity probabilities Θ —to maximize the likelihood of the observed data. To determine the optimal number of host and parasite blocks, we used the Integrated Completed Likelihood (ICL) criterion, which balances model fit and complexity by penalizing model size. This approach ensures robust and interpretable clustering (77).

Feature collection and processing

To investigate the determinants of parasite infection patterns, we measured a set of host traits and environmental variables for each host (**Figure 1B**, **Table 1**, and see **SI note 2** for detailed explanations). These variables might influence infection patterns by affecting both host exposure and susceptibility to infection. We assessed six host variables: (1) body mass, (2) body condition, calculated by Body Condition Index (BCI), (3) sex, (4) age, categorized as sub-adult and adult, (5) nematode co-infection, measured as presence/absence of nematodes, and (6) gut microbiome composition, measured as the relative abundance of four microbial families that were significantly correlated with the first two principal coordinates in a PCoA (**Figure S9**). Host mass, body condition, sex, and age can influence behavior (e.g., home range and social interactions) as well as physiological traits (50, 54, 78). Co-infection with macroparasites can further influence infection with microparasites. Specifically, nematodes are known to modulate the host immune system,

potentially altering susceptibility to co-infection by other parasites (35,79). The gut microbiome plays a critical role in host metabolism and immune function, serving as an indicator of overall health (37,80). Variation in microbial composition and relative abundance has previously been linked to several diseases and may reflect either an increased vulnerability of the host or a response to the disease itself (61,67).

We measured three environmental variables: (1) habitat structure (obtained via a vegetation PCA that distinguishes between tree-dominated and herbaceous-dominated sites; **Figure S6**), (2) distance to the village center (a proxy for habitat disturbance), and (3) small mammal community composition (population densities of rats, native species, and other non-native species including shrews and house mice). Population densities were calculated as the average number of individuals captured per sampling trap for every site and season. Vegetation type and proximity to the village are covariates that can influence the composition and survival of parasites in the environment, thus altering exposure risk to rats (41,81). Additionally, small mammal population density can impact infection patterns, as higher contact rates in denser populations may facilitate parasite transmission within and between host species (38,40). Thus, differences in species densities across sites may lead to distinct parasite transmission patterns among rat populations. The correlation between pairs of features ranged between -0.71 and 0.67 with average absolute value of 0.13 (**Figure S3**).

Table 1: Summary of features used in the XGBoost model. See **SI note 2** for detailed explanations on sampling and measurement methods.

Category	Sub-category	Feature	Type	Scale	Explanation	References
Host traits	Biological	Body mass	Continuous	Individual	Rat body mass [gram]	(50,51)
		Body condition	Continuous	Individual	Body Condition Index value by age and sex	
		Sex	Binary	Individual	Male / female	(54,82)
		Age	Binary	Individual	Sub-adult / adult	(52,78,83)
	Nematode co-infection	Nematode co-infection	Binary	Individual	Infection by any nematode species [0/1]	(35,65,79)
	Microbiome	<i>Lachnospiraceae</i>	Continuous	Individual	Family relative abundance [0-1]	(36,61,67,84)
		<i>Lactobacillaceae</i>	Continuous	Individual	Family relative abundance [0-1]	
		<i>Muribaculaceae</i>	Continuous	Individual	Family relative abundance [0-1]	
		<i>Prevotellaceae</i>	Continuous	Individual	Family relative abundance [0-1]	
Environmental	Vegetation	Vegetation PC1	Continuous	Site	Habitat attributes PC1	(41,81,85)
		Vegetation PC2	Continuous	Site	Habitat attributes PC2	
	Distance	Village distance	Continuous	Individual	Distance in [m] from the nearest village center	
	Community density	Rat density	Continuous	Site	Density of rat population at the site	(38,40,86,87)
		Non-native density	Continuous	Site	Density of non-native (shrews and house mice) at the site	
		Native density	Continuous	Site	Density of native sp. populations at the site	

Training a classification model to predict infection profiles

Using both host traits and environmental variables as suites of features, we trained an XGBoost (eXtreme Gradient Boosting) multi-classification model to predict host infection profile membership (SBM-identified groups)(**Figure 1D**). XGBoost is a distributed decision tree machine learning algorithm based on gradient boosting that efficiently handles structured data and captures complex patterns through ensemble learning (88). Machine learning methods, like XGBoost, effectively capture complex non-linear relationships between variables, which traditional linear models cannot represent (89,90).

To ensure that all data points were tested at least once while mitigating overfitting and improving the robustness of the analysis, we implemented a stratified nested cross-validation (CV) approach, maintaining class distributions across splits. The outer loop used a 3-fold cross-validation, where the dataset was split into three equal parts, and each subset was used once as a test set while the remaining two served as the training set. Within the inner loop, we conducted a 5-fold CV for hyperparameter tuning. Hence, within each training set, the data were further divided into five subsets, with four used for training and one for validation. This process was repeated for each fold, optimizing hyperparameters across different partitions of the data. To address data imbalance (uneven sizes of the three host infection profiles) and prevent the model from being biased toward the majority profile, we applied host profile weights inversely proportional to profile frequencies, assigning higher weights to minority profile samples during model training.

For hyperparameter tuning, we performed a grid search over key parameters, including maximum tree depth, learning rate (eta), column and row subsampling rates, L1/L2 regularization parameters (alpha, lambda), and minimum child weight (**Table S1**). The selected parameter search space aimed to minimize model complexity, thereby reducing the probability of overfitting. For each configuration, we trained the model using 300 boosting rounds, with early stopping based on validation loss to prevent overfitting. In each outer loop iteration, the best-performing hyperparameters (selected based on multi-class log-loss) were used to train a model, which was then evaluated on the outer test dataset. This process produced three models, corresponding to the three outer loops, each predicting a different fold of the dataset. The output of each model was a probability distribution over possible infection profiles, constrained to sum to one. The predicted profile was determined as the one with the highest probability.

To evaluate performance, we used a range of metrics designed to capture different aspects of predictive ability (91,92) (**SI note 3**). Since we had three infection profiles (see *Results*), we used metrics for multi-class classification. The evaluation was based on a 3×3 confusion matrix, which recorded the number of true positives and false positives for each infection profile. We used common evaluation metrics such as accuracy, weighted precision, weighted recall, weighted F1-score, weighted balanced accuracy, and Matthews correlation coefficient (MCC). To benchmark model performance against chance, we analytically derived expected metric values for a random classifier using proportional guessing based on profile prevalence. In addition, for each infection profile, we assessed one-vs-all performance across decision thresholds using Area Under the Receiver Operating Characteristic Curve (AUC-ROC) and Area Under the Precision-Recall Curve (PR-AUC). This allowed us to evaluate the model’s ability to distinguish among profiles and capture precision–recall tradeoffs, especially in the context of class imbalance. See **SI note 3** for detailed explanations.

Feature importance analysis

To interpret the contribution of each host or environmental variable to the classification task, we employed both XGBoost’s gain metric and SHapley Additive exPlanations (SHAP) values (**Figure 1E**). SHAP explains machine learning model predictions by quantifying the contribution of each input feature. Specifically, SHAP values measure how much each feature increases or decreases the model’s prediction compared to the average prediction ([44](#)). To determine these values, SHAP evaluates the model’s output across different subsets of features and calculates the difference in predictions when a feature is included versus when it is omitted, then averaging these differences across all possible feature combinations. This ensures that the contribution of each feature is measured while accounting for interactions with other features. The magnitude of a SHAP value represents the strength of a feature’s influence on the prediction, with larger absolute values indicating a greater impact.

We assessed global feature importance by averaging absolute SHAP values across all hosts in the test subset of the three-fold models (totaling 841 samples), and then summing these mean values across all infection profiles. Additionally, we aggregated feature importance by category (host traits versus environmental) by summing SHAP values within each category, enabling a comparative assessment of their relative influence on classification. We then used the SHAP values themselves to create dependence plots capturing how each feature’s effect varied across its observed range (e.g., the range of host masses), illustrating potential relationships with the profiles. Positive SHAP values indicate that a feature increases the prediction compared to the baseline model output, “pushing” the model towards a specific infection profile. In contrast, negative SHAP values decrease the prediction compared to the baseline, moving the model away from predicting that infection profile.

Code and data

All analyses were conducted in R (v4.2.1) ([93](#)). We used the R package *blockmodels* (v1.1.5) ([77](#)) for the SBM analysis, the package *XGBoost* ([88](#)) for model training, with the package *caret* ([94](#)) for cross-validation and performance evaluation. ROC and PR curves were created using the *pROC* ([95](#)) and *PRROC* ([96](#)) packages. SHAP values were computed with the package *shapviz* ([97](#)).

Acknowledgments

We sincerely thank all members of our Madagascar Health team for their invaluable support and thoughtful feedback. We are especially grateful to the Mention Zoologie et Biodiversité Animale, Université d’Antananarivo, Madagascar National Parks, and the Direction des Aires Protégées, des Ressources Naturelles Renouvelables et des Ecosystèmes for administrative aid and issuing research permits. Jacques Tahinarivony for the botanical work at the different study sites. Our appreciation also extends to the Duke Lemur Center–SAVA Conservation office for their logistical support. We are deeply thankful to the local communities and authorities who welcomed us, as well as to the numerous field assistants from the villages where our research took place.

Funding

This work was supported by the Israel Science Foundation (grant 1281/20 to SP), the U.S. National Science Foundation (grant DEB 2308460 to GT and CLN), the U.S.-Israel Binational Science Foundation (grant 2022721 to SP), the NIH-NSF-NIFA Ecology and Evolution of Infectious

Diseases program (award R01-TW011493 to CLN), a Duke University Provost’s Collaboratory Award to CLN, and the Human Frontier Science Program (grant RGY0064/2022 to SP).

Author contributions

Conceptualization: MM, IT, SP; Data sampling: VS, SMG, TMR; Sample creation and molecular and bioinformatic analysis: GT; Formal analysis: MM, IT; Funding acquisition: CLN, SP, GT; Supervision: SP; Writing – original draft: MM, IT, SP, GT; Writing – review and editing: MM, IT, SP, GT, CLN, BB, SMG, VS, TMR.

Competing interests

The authors declare that they have no competing interests.

Data and materials availability

All data and code needed to evaluate the conclusions in the paper are present on the GitHub repository https://github.com/MadagascarEEID/rat_protozoa_infection_profiles.

References

1. R Poulin, *Parasitology* **134**, 763–776 (2007).
2. D. M. Tompkins, A. M. Dunn, M. J. Smith, S. Telfer, *J. Anim. Ecol.* **80**, 19–38 (2011).
3. K. L. VanderWaal, V. O. Ezenwa, *Funct. Ecol.* **30**, 1606–1622 (2016).
4. S. H. Paull *et al.*, *Front. Ecol. Environ.* **10**, 75–82 (2012).
5. J. M. Trauer *et al.*, *Clin. Infect. Dis.* **69**, 159–166 (2019).
6. M. G. M. Gomes *et al.*, *J. Theor. Biol.* **540**, 111063 (2022).
7. T. E. Stewart Merrill, S. R. Hall, C. E. Cáceres, *Ecology* **102**, e03245 (2021).
8. J. Sherry, E. H. Rego, *Annu. Rev. Genet.* **58**, 183–209 (2024).
9. A. B. Pedersen, A. Fenton, *Trends Ecol. Evol.* **22**, 133–139 (2007).
10. F. Venter, K. R. Matthews, E. Silvester, *Proc. Biol. Sci.* **289**, 20212155 (2022).
11. P. T. J. Johnson, J. C. de Roode, A. Fenton, *Science* **349**, 1259504 (2015).
12. R. Runghen, R. Poulin, C. Monlleó-Borrull, C. Llopis-Belenguer, *Trends Parasitol.* **37**, 445–455 (2021).
13. F. Bordes *et al.*, *J. Anim. Ecol.* **84**, 1253–1263 (2015).
14. T. A. Dallas, P. Jordano, *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **376**, 20200361 (2021).
15. D. P. Lima Jr, H. C. Giacomini, R. M. Takemoto, A. A. Agostinho, L. M. Bini, *J. Anim. Ecol.* **81**, 905–913 (2012).
16. T. A. Dallas *et al.*, *Oikos* **128**, 23–32 (2019).
17. S. Pilosof, S. Morand, B. R. Krasnov, C. L. Nunn, *PLoS One* **10**, e0117909 (2015).
18. S. Allesina, M. Pascual, *Ecol. Lett.* **12**, 652–662 (2009).
19. M. Rosvall, J.-C. Delvenne, M. T. Schaub, R. Lambiotte, in *Advances in network clustering and blockmodeling*, ed. by P. Doreian, V. Batagelj, A. Ferligoj (Wiley, 2018), pp. 71–87.
20. R. Poulin, B. R. Krasnov, D. Mouillot, *Trends Parasitol.* **27**, 355–361 (2011).
21. J.-B. Leger, J.-J. Daudin, C. Vacher, *Methods in Ecology and Evolution* **6**, 474–481 (2015).
22. E. B. Baskerville *et al.*, *PLoS Comput. Biol.* **7**, e1002321 (2011).

23. S. Cobo-López, V. K. Gupta, J. Sung, R. Guimerà, M. Sales-Pardo, *PNAS Nexus* **1**, gac055 (2022).
24. G. Galai *et al.*, *Ecography (Cop.)*, e07430 (2024).
25. V. Miele, R. Ramos-Jiliberto, D. P. Vázquez, *J. Anim. Ecol.* **89**, 1670–1677 (2020).
26. S. Hunter-Barnett, M. Viney, *Parasitology* **151**, 594–605 (2024).
27. S. Rahelinirina *et al.*, *PLoS One* **5**, e14111 (2010).
28. L. A. Spencer, M. T. Irwin, *Heliyon* **6**, e05604 (2020).
29. M Rasoanoro, B Ramasindrazana, S. M. Goodman, M Rajerison, M Randrianarivelosia, *Malagasy Natiora* **13**, 65–75 (2019).
30. R. Poulin, *Evolutionary ecology of parasites: (second edition)* (Princeton University Press, Princeton, NJ, ed. 2, 2007).
31. P. T. J. Johnson, J. T. Hoverman, *J. Anim. Ecol.* **83**, 1103–1112 (2014).
32. A. R. Sweeny, G. F. Albery, *Funct. Ecol.* **36**, 1713–1726 (2022).
33. V. O. Ezenwa *et al.*, *Proc. Biol. Sci.* **283**, 20153078 (2016).
34. I. M. Cattadori, B Boag, P. J. Hudson, *Int. J. Parasitol.* **38**, 371–380 (2008).
35. N. A. Mabbott, *Front. Immunol.* **9**, 2579 (2018).
36. A. P. Bernardo-Cravo, D. S. Schmeller, A. Chatzinotas, V. T. Vredenburg, A. Loyau, *Trends Parasitol.* **36**, 616–633 (2020).
37. A. L. Gould *et al.*, *Proc. Natl. Acad. Sci. U. S. A.* **115**, E11951–E11960 (2018).
38. Z. Gajewski *et al.*, *Ecol. Lett.* **27**, e14385 (2024).
39. J. Cable *et al.*, *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **372** (2017).
40. M Begon *et al.*, *Epidemiol. Infect.* **129**, 147–153 (2002).
41. N. L. Gottdenker, D. G. Streicker, C. L. Faust, C. R. Carroll, *Ecohealth* **11**, 619–632 (2014).
42. H. Sulaiman, A. Kukreja, O. H. Cheng, in *International Encyclopedia of Public Health* (Elsevier, 2025), pp. 777–803.
43. A. Dumètre *et al.*, *Appl. Environ. Microbiol.* **78**, 905–912 (2012).
44. S. M. Lundberg, S.-I. Lee, *Neural Inf Process Syst* **30**, 4765–4774 (2017).
45. C Brouat *et al.*, *Mol. Ecol.* **23**, 4153–4167 (2014).
46. K. Scobie *et al.*, *Integr. Zool.* **19**, 66–86 (2024).
47. T. M. Randriamoria, B Ramasindrazana, V Soarimalala, S. M. Goodman, in *The new natural history of Madagascar*, ed. by S. M. Goodman (Princeton University Press, Princeton, NJ, 2014), pp. 2006–2008.
48. M. Dammhahn, T. M. Randriamoria, S. M. Goodman, *BMC Ecol.* **17**, 16 (2017).
49. V. Marzano *et al.*, *PLoS Negl. Trop. Dis.* **11**, e0005916 (2017).
50. C. A. Sánchez *et al.*, *Ecol. Lett.* **21**, 1869–1884 (2018).
51. E. S. Durkin, L. T. Luong, J. Bird, *Parasitol. Res.* **114**, 4169–4174 (2015).
52. M. Santin, A. Molokin, J. G. Maloney, *Parasit. Vectors* **16**, 177 (2023).
53. R. Izhar, F. Ben-Ami, *J. Anim. Ecol.* **84**, 1018–1028 (2015).

54. D. A. Grear, S. E. Perkins, P. J. Hudson, *Ecol. Lett.* **12**, 528–537 (2009).
55. R. H. Gellman *et al.*, *Cell Rep.* **42**, 113233 (2023).
56. C. L. Betancur-Murillo, S. B. Aguilar-Marín, J. Jovel, *Microorganisms* **11**, 1 (2022).
57. Y. Zhu *et al.*, *Nutrients* **16**, 2660 (2024).
58. A. O. G. Hoarau, P. Mavingui, C. Lebarbenchon, *PLoS Pathog.* **16**, e1008790 (2020).
59. A. W. Park *et al.*, *Proc. Biol. Sci.* **285** (2018).
60. T. M. Barrett *et al.*, *Am. J. Biol. Anthropol.* **185**, e25030 (2024).
61. M. Levy, A. A. Kolodziejczyk, C. A. Thaiss, E. Elinav, *Nat. Rev. Immunol.* **17**, 219–232 (2017).
62. H. Brüßow, *Microb. Biotechnol.* **13**, 423–434 (2020).
63. M. Markfeld *et al.* (2025).
64. S. P. Keegan, A. B. Pedersen, A. Fenton, *Proc. Biol. Sci.* **291**, 20240103 (2024).
65. V. O. Ezenwa, *Parasite Immunol.* **38**, 527–534 (2016).
66. A. Budria, U. Candolin, *Parasitology* **141**, 462–474 (2014).
67. S. P. Rosshart *et al.*, *Cell* **171**, 1015–1028.e13 (2017).
68. J. M. Leung, A. L. Graham, S. C. L. Knowles, *Front. Microbiol.* **9**, 843 (2018).
69. G. Fackelmann *et al.*, *Commun Biol* **4**, 800 (2021).
70. D. A. May, F. Taha, M. A. Child, S. E. Ewald, *Trends Parasitol.* **39**, 1074–1086 (2023).
71. M. Camacho-Sanchez, P. Burraco, I. Gomez-Mestre, J. A. Leonard, *Mol. Ecol. Resour.* **13**, 663–673 (2013).
72. J. F. Gogarten *et al.*, *Mol. Ecol. Resour.* **20**, 204–215 (2020).
73. L. R. Krogsgaard *et al.*, *Clin. Transl. Gastroenterol.* **9**, 161 (2018).
74. M. Martin, *EMBnet.journal* **17**, 10–12 (2011).
75. B. J. Callahan *et al.*, *Nat. Methods* **13**, 581–583 (2016).
76. A. Murali, A. Bhargava, E. S. Wright, *Microbiome* **6**, 140 (2018).
77. J.-B. Leger, *arXiv [stat.CO]* (2016).
78. S. J. Cornell, O. N. Bjornstad, I. M. Cattadori, B. Boag, P. J. Hudson, *Proc. Biol. Sci.* **275**, 511–518 (2008).
79. M. L. Rodgers, D. I. Bolnick, *Oecologia* **204**, 317–325 (2024).
80. G. Berg *et al.*, *Microbiome* **8**, 1–22 (2020).
81. J. A. Patz, T. K. Graczyk, N. Geller, A. Y. Vittor, *Int. J. Parasitol.* **30**, 1395–1405 (2000).
82. S. L. Klein, *Parasite Immunol.* **26**, 247–264 (2004).
83. D. L. Preston, L. P. Falke, M. Novak, *Funct. Ecol.* **39**, 91–102 (2025).
84. M. Chabé, A. Lokmer, L. Ségurel, *Trends Parasitol.* **33**, 925–934 (2017).
85. F. Kiene *et al.*, *Ecol. Evol.* **11**, 6766–6788 (2021).
86. J. E. H. Patterson, K. E. Ruckstuhl, *Parasitology* **140**, 803–813 (2013).
87. P. Arneberg, A. Skorping, B. Grenfell, A. F. Read, *Proc. Biol. Sci.* **265**, 1283–1289 (1998).

88. T. Chen, C. Guestrin, presented at the Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.
89. S. M. Lundberg *et al.*, *Nat. Mach. Intell.* **2**, 56–67 (2020).
90. W. Manley, T. Tran, M. Prusinski, D. Brisson, *Peer Community J.* **3** (2023).
91. T. Poisot, *Methods Ecol. Evol.* **14**, 1333–1345 (2023).
92. M. Grandini, E. Bagli, G. Visani, *arXiv [stat.ML]* (2020).
93. R Core Team, *R: A Language and Environment for Statistical Computing*, Vienna, Austria, 2024.
94. Kuhn, Max, *Building Predictive Models in R Using the caret Package*, 2008.
95. X. Robin *et al.*, *pROC: an open-source package for R and S+ to analyze and compare ROC curves*, 2011.
96. J. Grau, I. Grosse, J. Keilwagen, *PRROC: computing and visualizing precision-recall and receiver operating characteristic curves in R*, 2015.
97. M. Mayer, *shapviz: SHAP Visualizations*, 2025.
98. J. G. Caporaso *et al.*, *ISME J.* **6**, 1621–1624 (2012).
99. R. B. Gasser, N. B. Chilton, H Hoste, I Beveridge, *Nucleic Acids Res.* **21**, 2525–2526 (1993).

Supplementary information

Table S1: Hyperparameter grid search settings for the XGBoost model.

Hyperparameter	Values	Description
max_depth	1, 2, 3	Maximum depth of a tree; controls model complexity.
eta	0.002, 0.01, 0.1	Learning rate; scales the contribution of each tree.
colsample_bytree	0.4, 0.6	Fraction of features sampled per tree.
min_child_weight	6, 8	Minimum sum of instance weight needed in a child.
subsample	0.5, 0.7	Fraction of training data sampled per tree.
lambda	5, 10	L2 regularization term on weights; reduces overfitting.
alpha	2, 5	L1 regularization term on weights; encourages sparsity.

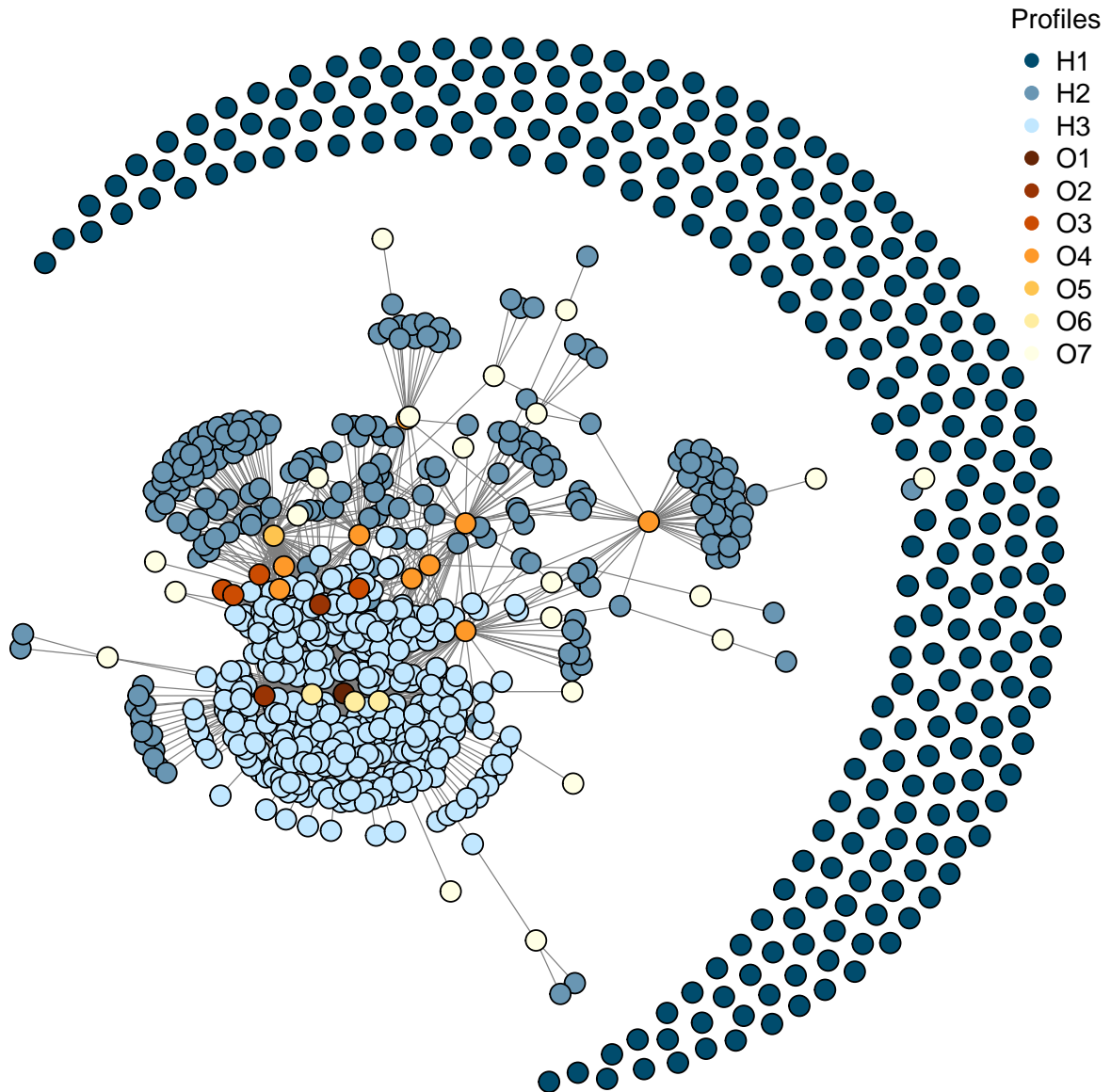


Figure S1: Visualization of the rat-protozoa bipartite network. Nodes represent individual rat *Rattus rattus* hosts and protozoa OTUs, with edges indicating infections. Blue-shaded nodes denote host infection profiles, while yellow-to-red shaded nodes represent protozoa infection profiles.

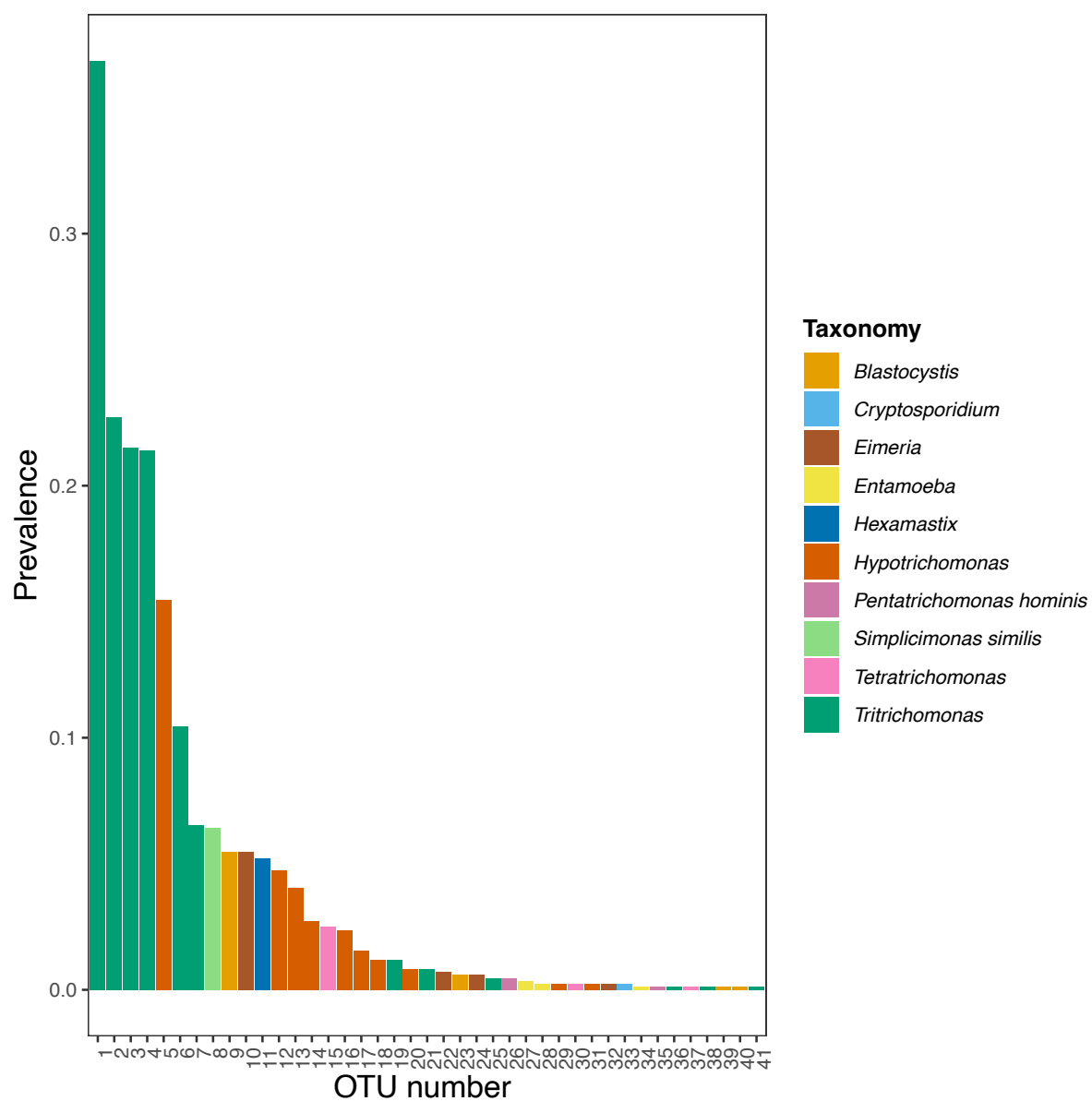


Figure S2: Protozoa OTUs prevalence distribution and taxonomy. The prevalence of 41 OTUs, measured as their occupancy out of the 841 hosts, ordered from highest to lowest. Colors represent the lowest taxonomic classification (genus or species) of each OTU.

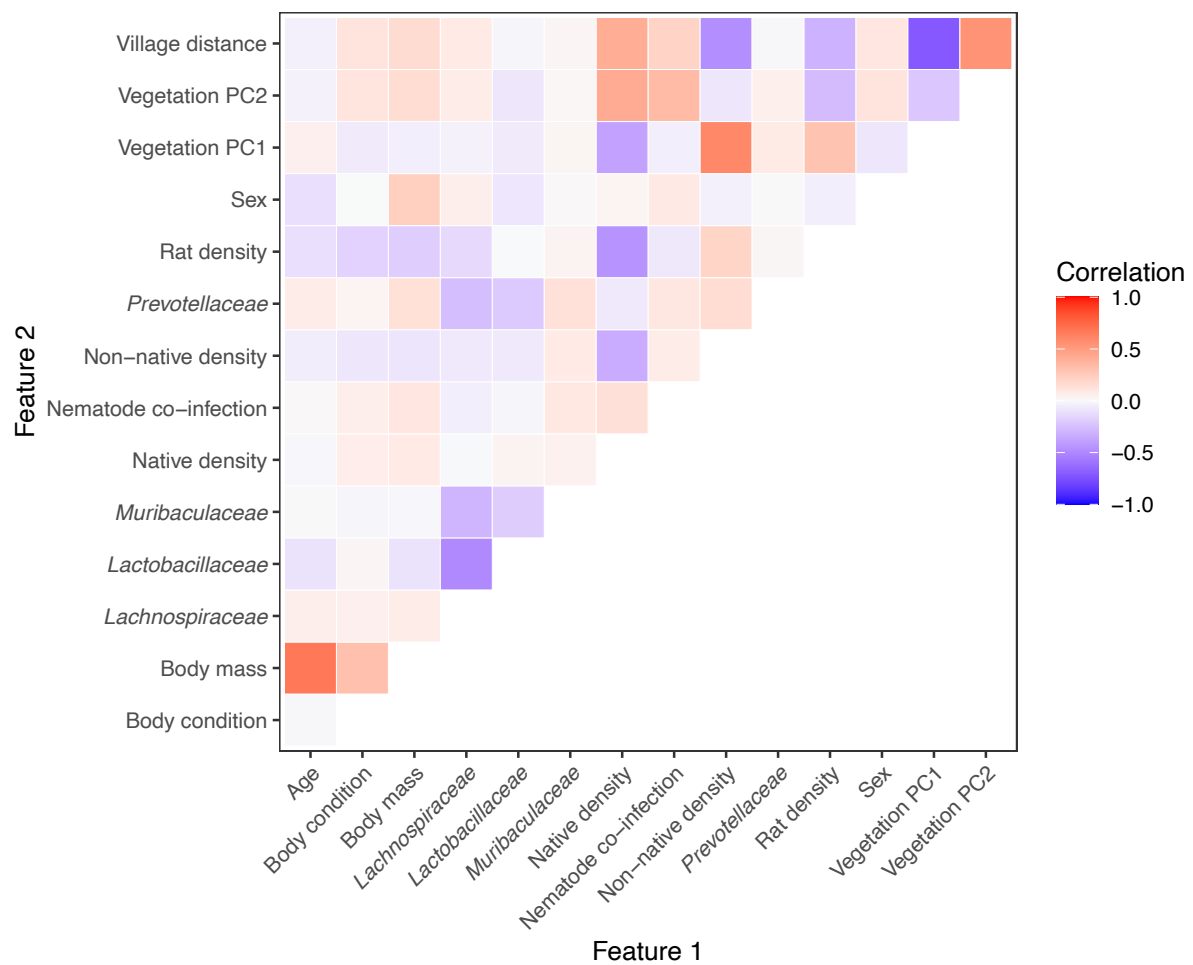


Figure S3: Correlations between features. Pairwise correlations were computed between all features. Pearson correlation was used for continuous–continuous pairs, point-biserial correlation for continuous–dichotomous pairs, and tetrachoric correlation when both features were dichotomous. Colors represent the strength and direction of the correlations: red indicates positive, and blue indicates negative relationships.

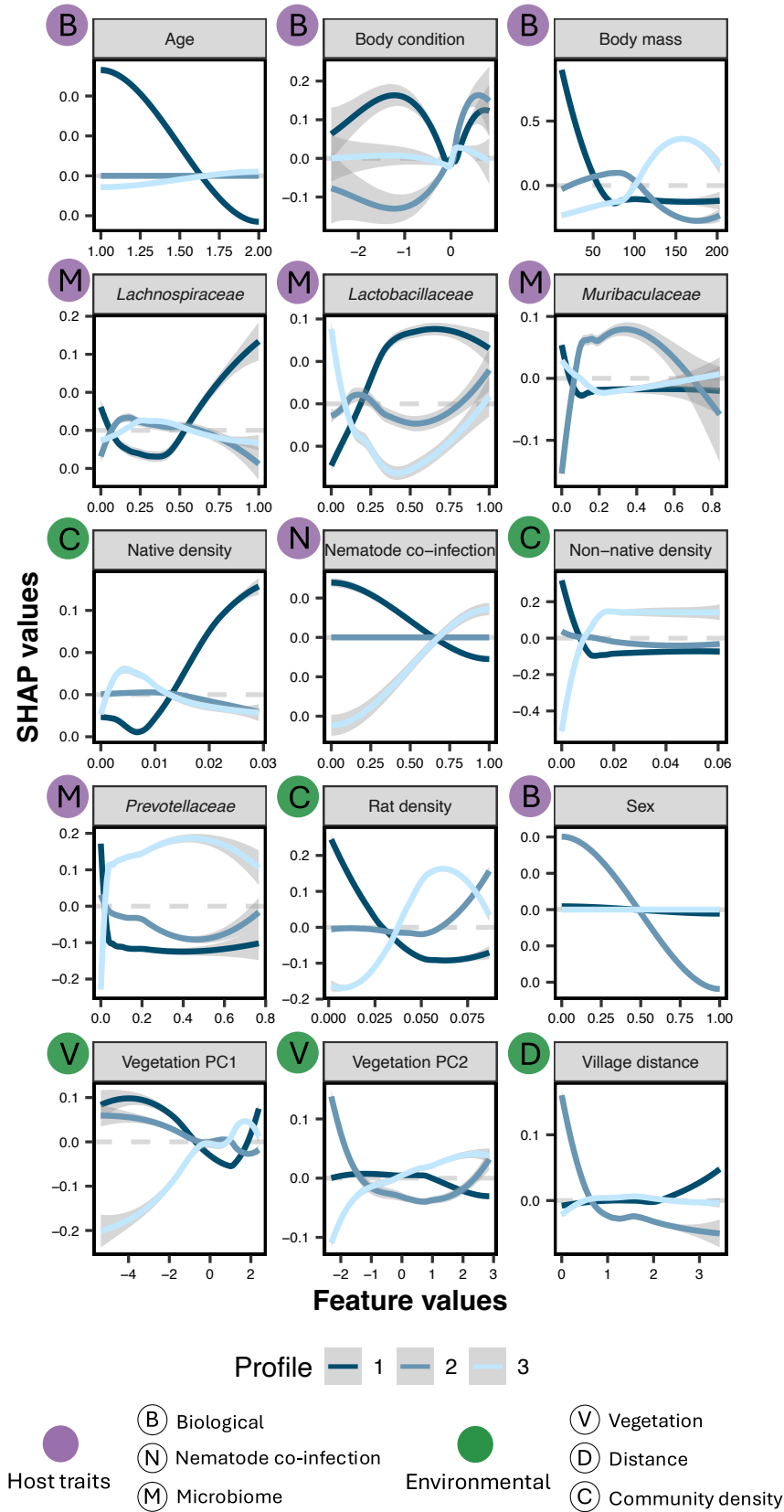


Figure S4: SHAP dependency plots for all features. SHAP dependency plots show individual hosts' SHAP values as a function of feature values. For clarity, we visualized only the LOESS-derived trendline and its associated confidence interval for the 841 host samples, rather than displaying all individual host data points. Line colors represent different host infection profiles, while the dashed gray line marks a SHAP value of zero. Y-axis values were rounded to two decimal places.

SI note 1: Study site and small mammal sampling

Small mammals were collected in the vicinity of three villages in the SAVA Region of northeast Madagascar, in the surroundings of Marojejy National Park. The village of Mandena (14.477049° S, 49.8147° E) was sampled between October 2019 and September 2020. A second village, Sarahandrano (14.607567° S, 49.647759° E), was sampled between November 2020 and September 2021, while a third village, Andatsakala (14.397276° S, 49.8820° E), was sampled between October 2021 and August 2022. In the vicinity of each village, seven sites were sampled along a degradation gradient: (1) semi-intact natural forest inside the national park, (2) secondary forest, (3) *savoka* (brushy regrowth), (4) agroforest (vanilla plantation), (5) mixed agriculture (sugarcane/coffee plantation), (6) flooded rice, and (7) the village itself. Sites near each village were located ~500 m apart.

For sampling small mammals, a 100 m X 100 m grid of 121 live traps (11x11) was established, including 97 Sherman (H. B. Sherman Traps, Inc., Tallahassee, Florida, model LFA and XLK), and 24 Tomahawk (Tomahawk Live Trap, Hazelhurst, Wisconsin, model 201), placed 10 m apart and baited with peanut butter. Additionally, two pitfall lines were installed between 20-50 m outside of the grid, running in parallel to the grid edge. Each pitfall line was 100 m in length, with 11 buckets dug into the ground and placed every 10 m, and an 80 cm high vertically oriented plastic fencing bisecting each bucket, stapled to vertical stakes, and a flange touching the ground covered with soil and leaf litter to block the passage of small mammals and guide them to a bucket. Each plot was sampled for six consecutive nights and during three different sampling periods (before the wet season, after the wet season, and during the dry season).

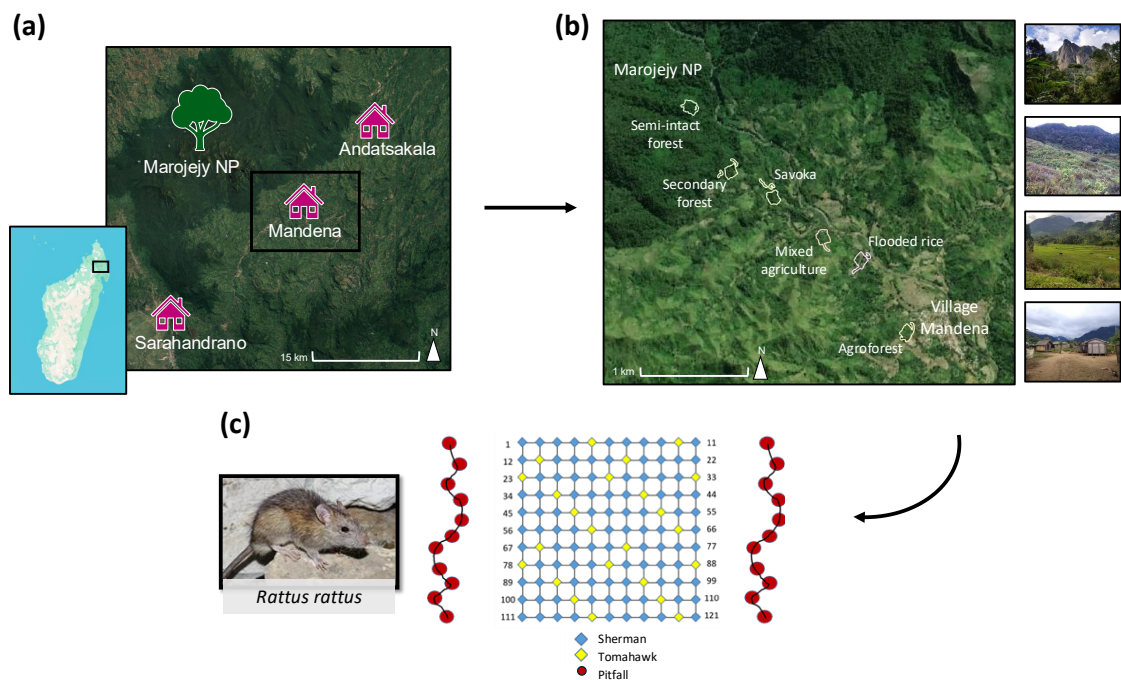


Figure S5: Study site and sampling scheme. (a) Sampling was conducted in northeastern Madagascar, in three different zones associated with three villages near Marojejy National Park. (b) In each village, seven distinct land-use types were sampled. The map illustrates the village of Mandena as an example. The images depict typical landscapes from top to bottom: semi-intact forest, *savoka*, flooded rice fields, and village plots. (c) In each plot, an 11×11 trapping grid consisting of Sherman and Tomahawk traps was installed, along with two pitfall lines. A total of 841 individual *Rattus rattus* were captured.

SI note 2: Measuring host traits and environmental features

To explore determinants of parasite infection patterns, we measured three environmental variables (vegetation, small mammal community, and distance to the village center) and six host variables (mass, body condition index, sex, age, nematode co-infection, and gut microbiome composition) (**Table 1**). The host variables and distance to the village are specific to each individual rat, while the environmental variables are specific to a site and common to all the rats captured at a specific site in a specific season.

Habitat attributes

We measured two environmental gradients across sites: vegetation and the distance from the village center. These variables collectively capture much of the natural and anthropogenic variation across the landscape and are related to the environmental reservoir of parasites. The distance to the nearest village was measured using a GPS logger as the shortest distance from the village center to the trap location where the rat was captured.

In addition, with the help of a specialist botanist, we measured habitat attributes in 16 plots (5m \times 5m) within the sampling grid at each site, conducting measurements three times (seasons) during the sampling period. At each plot, we assessed eight habitat characteristics: (1) number of trees, (2) number of dead logs, (3) tree diameter at breast height, (4) tree height, (5) percent canopy cover, (6) number of lianas, (7) herbaceous vegetation height, and (8) percent herbaceous vegetation cover. We averaged the measurements across all plots to calculate mean values for each site per season. To explore habitat variation between sites, we conducted a principal component analysis (PCA). Prior to analysis, all variables were centered at 0 and rescaled to have unit variance. The first two principal components explained 80.71% of the variation across sites (PC1: 59.51%, PC2: 21.2%) (**Figure S6**). Vegetation PC1 divides the more natural sites (semi-intact forest and secondary forest) from the more disturbed sites. PC2 is positively correlated with herbaceous vegetation cover and height.

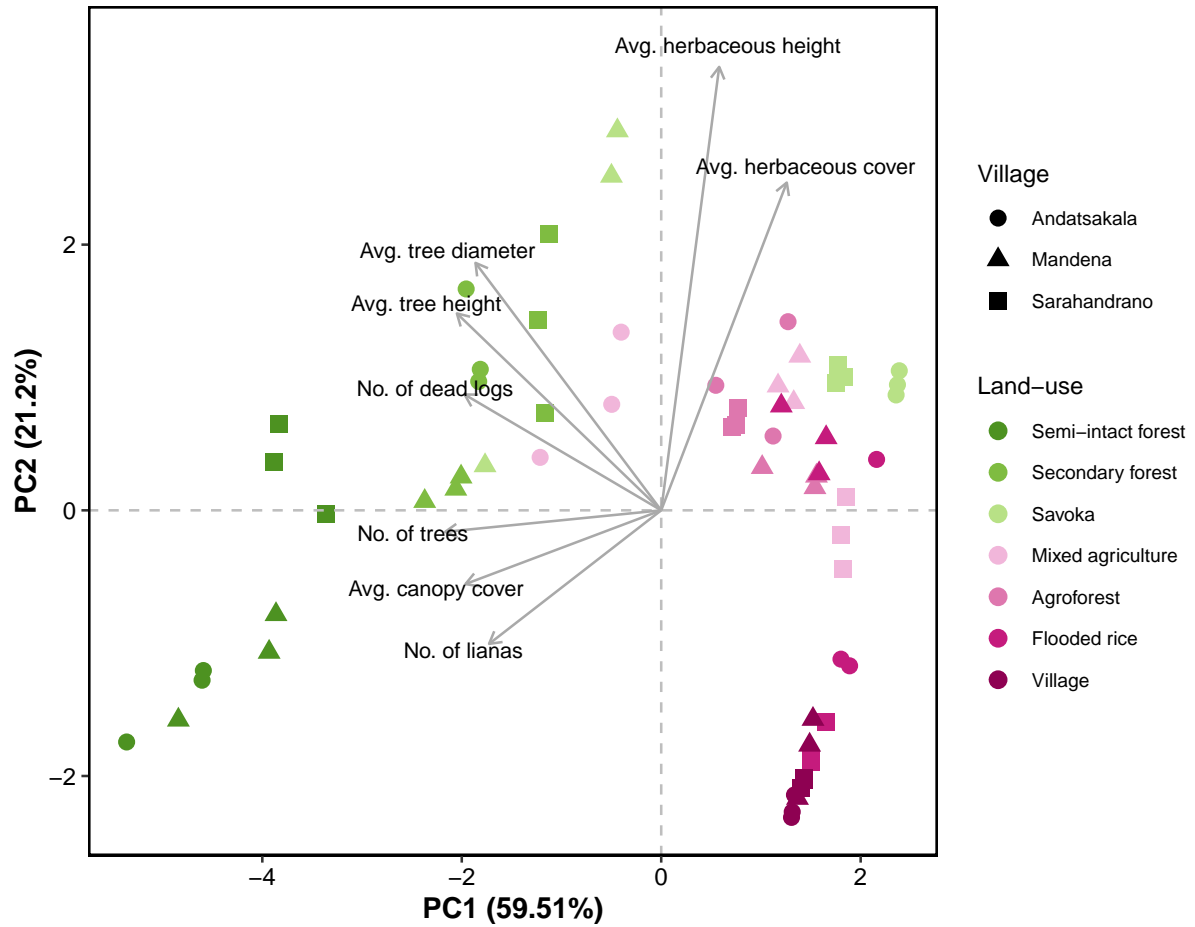


Figure S6: Vegetation PCA across land-use types. The first two principal components (PC1 and PC2) from the PCA of vegetation attributes. Each point represents a site in one season, with shape denoting the village and color representing the land-use type. Arrow length and direction indicate the contribution of each vegetation variable to the first two PCs.

Small mammal community

Population density can influence infection patterns, as higher contact rates in denser populations may increase parasite transmission within and between host species. Therefore, for each site and season, we measured small mammal density for (1) the rat (*Rattus rattus*) population, (2) other non-native species (including *Mus musculus* and *Suncus* spp.), and (3) native species (including members of the family Tenrecidae and subfamily Nesomyinae). Density was calculated as the total abundance of individuals in categories 1–3 at a given site-season, divided by the number of traps at that site. For the rat population, pitfall traps were excluded from calculations, as only two individuals (0.2%) were captured in them. The small mammal community significantly differed between more natural sites (semi-intact forest and secondary forest) and more disturbed sites (Figure S7).

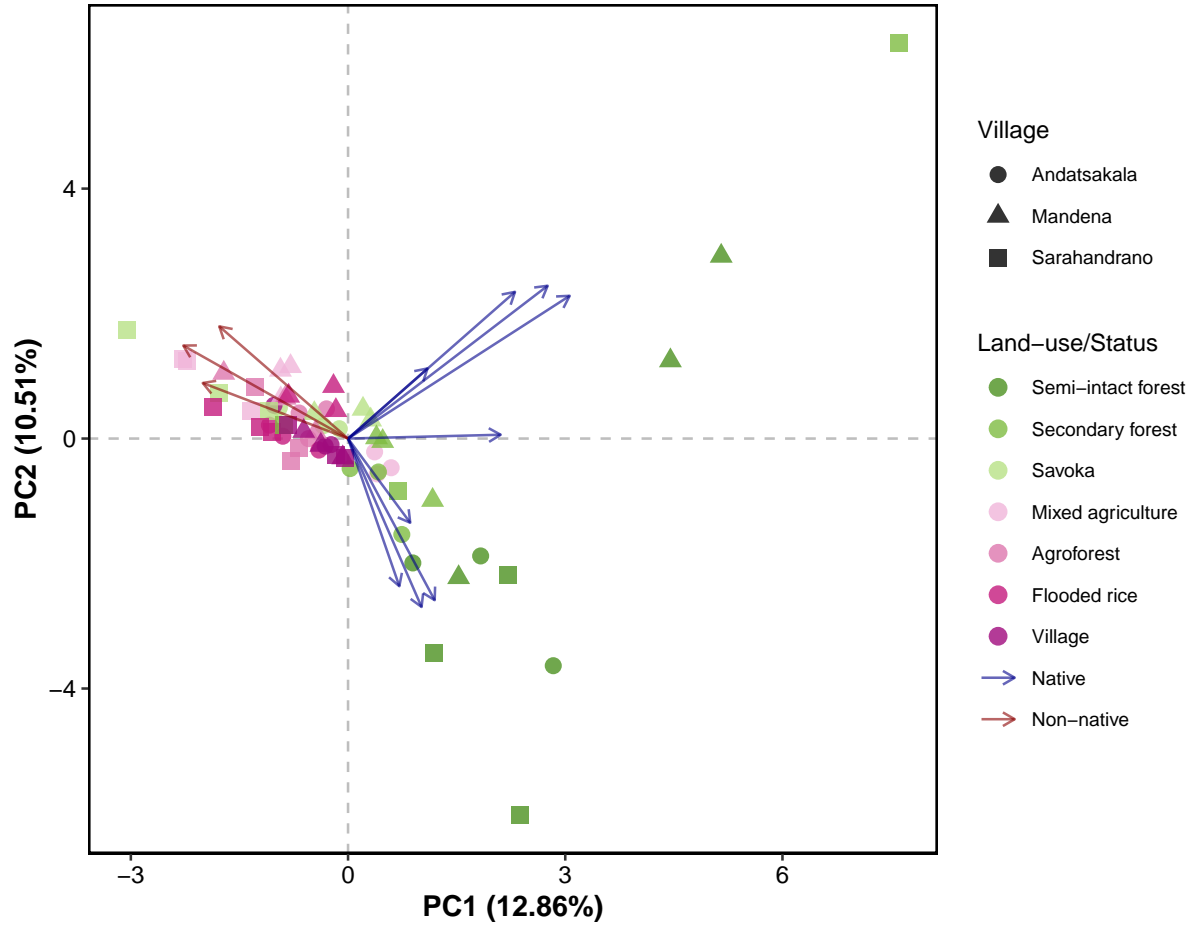


Figure S7: Small mammal community PCA across land-use types. The first two principal components (PC1 and PC2) from the PCA of small mammals species. Each point represents a site in one season, with shape denoting the village and color representing the land-use type. Arrow length and direction indicate the contribution of each species to the first two PCs. The arrow colors indicate native (blue) and non-native (red) species.

Body condition

To assess the physiological condition of individual rats, we calculated a Body Condition Index (BCI) based on the residuals from a linear regression of body mass on structural body length. Because body size and growth patterns can differ significantly between age classes and sexes, we calculated BCI separately for each combination of age group (sub-adult/adult) and sex. For each subgroup, we log-transformed both body mass (M) and head-body length (L) to linearize the allometric relationship. We then fit a linear model of the form:

$$\log(M_i) = \beta_0 + \beta_1 \log(L_i) + \varepsilon_i \quad (\text{S1})$$

where M_i is the mass of individual i , L_i is its head-body length, β_0 and β_1 are the intercept and slope of the regression, and ε_i is the residual. The residuals ε_i from this regression represent the BCI, with positive values indicating individuals heavier than expected for their body length (i.e., better condition), and negative values indicating poorer condition. These residuals were used as a continuous predictor of host condition in the subsequent statistical model.

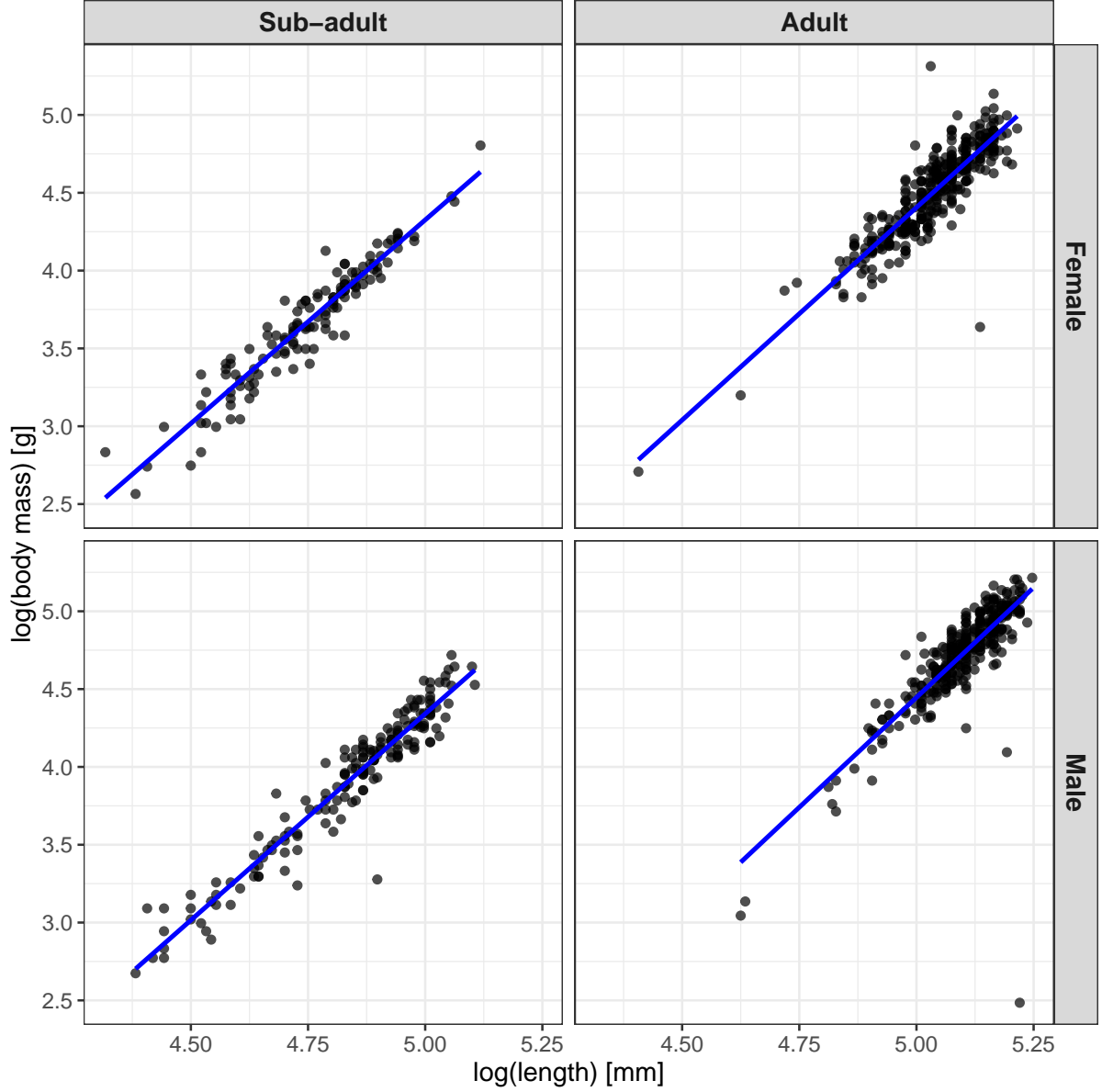


Figure S8: Body Condition Index (BCI) of individual rat hosts. BCI was calculated by the residuals from a linear regression of body mass [g] on structural body length [mm] for each combination of age group (sub-adult/adult) and sex (female/male).

Gut microbiome

DNA was extracted from ~1g feces collected from trapped small mammals using Zymo Quick-DNA Fecal/Soil Microbe Miniprep kits (cat #D6010) using manufacturer protocols. 16S metabarcoding was conducted using 515F–806R primers to target the V4 region of the 16S SSU rRNA (98). Each primer included an Illumina adapter, barcode, primer pad, and linker. Reactions were carried out in 25 μ L volumes consisting of 10 μ L of 1.25 μ M forward and reverse primer,

2 μL of DNA, and 13 μL of Platinum Hot Start PCR mastermix (ThermoFisher Scientific, cat #13000014). Reaction conditions were as follows: 95°C for 3min, 35x 98°C for 30secs, 58°C for 30secs, and 72°C for 30 s, followed by a final extension at 72°C for 5min. Concentrations were measured using Promega One Quantifluor kits on a Tecan plate reader. Samples were then normalized to 7 ng/ μL prior to pooling. The product was cleaned using magnetic beads (bead:DNA ratio was 0.8:1) and sequenced at UC Santa Barbara Biological Nanostructures Laboratory on an Illumina MiSeq (v3 chemistry, 2x300 bp, 24M reads).

Sequences were demultiplexed using cutadapt (v.3.4) with zero error tolerance (74). We then performed quality filtering steps using the *dada2* package in R (75). Specifically, we filtered and trimmed amplicons (minimum length = 100, 15% PhiX removed), inferred and removed errors, dereplicated sequences, inferred amplicon sequencing variants (ASVs) using the pseudo-pooling method, merged pairs, and removed chimeras. We assigned taxonomic identifications to ASVs using the *assignTaxonomy* function in *dada2*, using the SILVA nr99 SSU reference database (v.138.1).

We filtered out very rare ASVs with a relative abundance lower than 0.1% in a sample or those that occur in less than 1% of all individuals. Additionally, we removed all non-bacterial ASVs or those that were identified as 'Chloroplast' or 'Mitochondria'. Finally, we excluded 21 samples with fewer than 5000 total reads from our analysis. Filtering procedures resulted in 1,951 ASVs from an original total of 10,358.

We aggregated ASVs at the family level for each individual host. ASVs with unidentified families were excluded, resulting in the analysis of 1,770 ASVs (90.72% of all ASVs) classified into 55 families. To examine microbiome variation among individuals, we performed a principal coordinate analysis (PCoA). The first two principal coordinates explained 56.8% of the variation (PCo1: 36.5%, PCo2: 20.3%) (**Figure S9**). For better interpretability, we selected as features only the microbial families that exhibited a significant correlation with the first two principal coordinates (PCos). To achieve this, we utilized the equilibrium circle (or correlation circle), a graphical tool that helps interpret the contribution of variables to the principal coordinates. The radius of the equilibrium circle is given by $\text{Radius} \propto \left(\frac{d}{p}\right)^{0.5}$, where d represents the number of retained principal coordinates (two in our case), and p denotes the total number of original variables. A variable (microbial family) with a vector extending beyond the equilibrium circle indicates a strong correlation with at least one principal coordinate, signifying a major role in the ordination. Conversely, vectors within the circle reflect weaker correlations and lower contributions to the PCoA structure. Based on this criterion, we identified four microbial families as significant and included them as features in the final XGBoost model: (1) *Lachnospiraceae*, (2) *Lactobacillaceae*, (3) *Muribaculaceae*, and (4) *Prevotellaceae*.

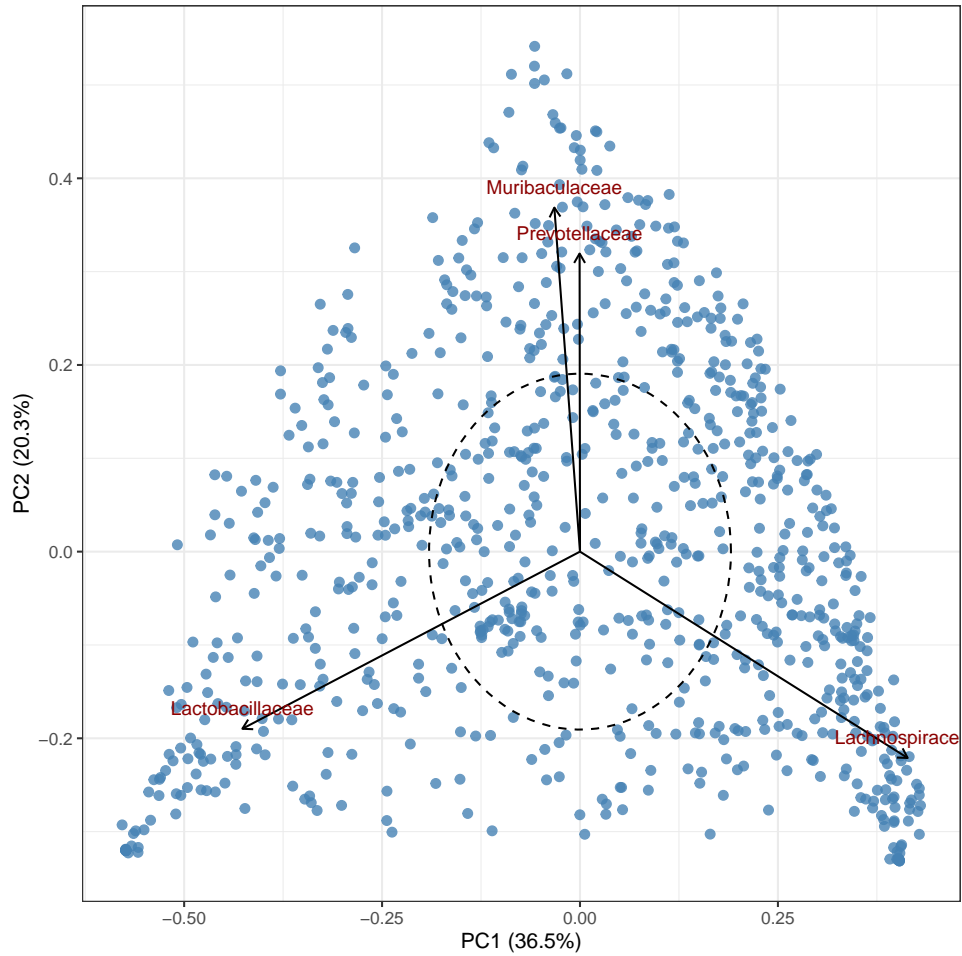


Figure S9: PCoA of the gut microbiome across individual rat hosts. The first two principal coordinates (PCo1 and PCo2) are shown. Each point represents an individual rat, with the circle marking the equilibrium threshold. Arrow length and direction indicate the contribution of each microbial family to the first two PCs, with only families whose arrows extend beyond the equilibrium circle displayed.

Nematode co-infection

We performed metabarcoding using the NC1/NC2 primer set (99) to amplify ITS2 ribosomal DNA from strongylid nematodes. Forward and reverse primers contained 8-nucleotide barcodes with a Hamming distance of at least 4. PCR reactions were carried out in 15 μ L volumes consisting of: 3 μ L of each forward and reverse primer (2 μ M stock concentration); 7 μ L from a Mastermix comprised of 0.7 μ L of Amplitaq Gold polymerase, 150 μ L MgCl₂, 150 μ L Amplitaq Gold buffer, 12 μ L BSA, 6 μ L DMSO, and 344 μ L water; and up to 2 μ L template DNA (1–100 ng total). Cycling conditions were: 10-minute hot-start activation, 35x cycles of 15 s at 95°C, 30 s at 55°C, 40 s at 72°C, and a final 5-min extension at 72°C. DNA concentrations were then measured, pooled, normalized, and purified using MinElute columns prior to multiplexing with additional libraries. The final library for each village was sequenced three times on an Illumina MiSeq (v3 2 \times 300 bp, 25 M reads) at the UC Davis Genome Center. Sequences were demultiplexed using cutadapt (v.3.4) with zero error tolerance (74). We used the *dada2* bioinformatics pipeline (75) to filter and trim amplicons (minimum length = 100, 15% PhiX removed), remove errors, dereplicate, infer amplicon sequence variants (ASVs) using the pseudo-pooling method, merge pairs, remove chimeras, and combine the three ASV read tables from

the different villages into one table. We then calculated the relative read abundance of each ASV and excluded reads that accounted for less than 1% of a sample's relative read abundance to avoid potential sequencing errors or tag jumps. We excluded a small subset of samples that failed to amplify across other primer sets, and used the assignTaxonomy function with minimum bootstraps = 50 to identify ASV sequences using the nemabiome ITS2 reference database (v 1.6.0). Next, we clustered phylogenetically similar ASVs into OTUs at 97% similarity using the 'Clusterized' function from the *DECIPHER* package. Taxonomy was assigned to each OTU based on its most common ASV.

To examine variation in nematode co-infection among individual hosts, we plotted the distribution of the number of nematode OTUs infecting each host (**Figure S10**). Because the distribution was highly skewed, with most rats infected by only a single nematode OTU, we included a binary variable in the final model indicating whether or not the host was infected by any nematode species (infected: $n = 626$; uninfected: $n = 215$).

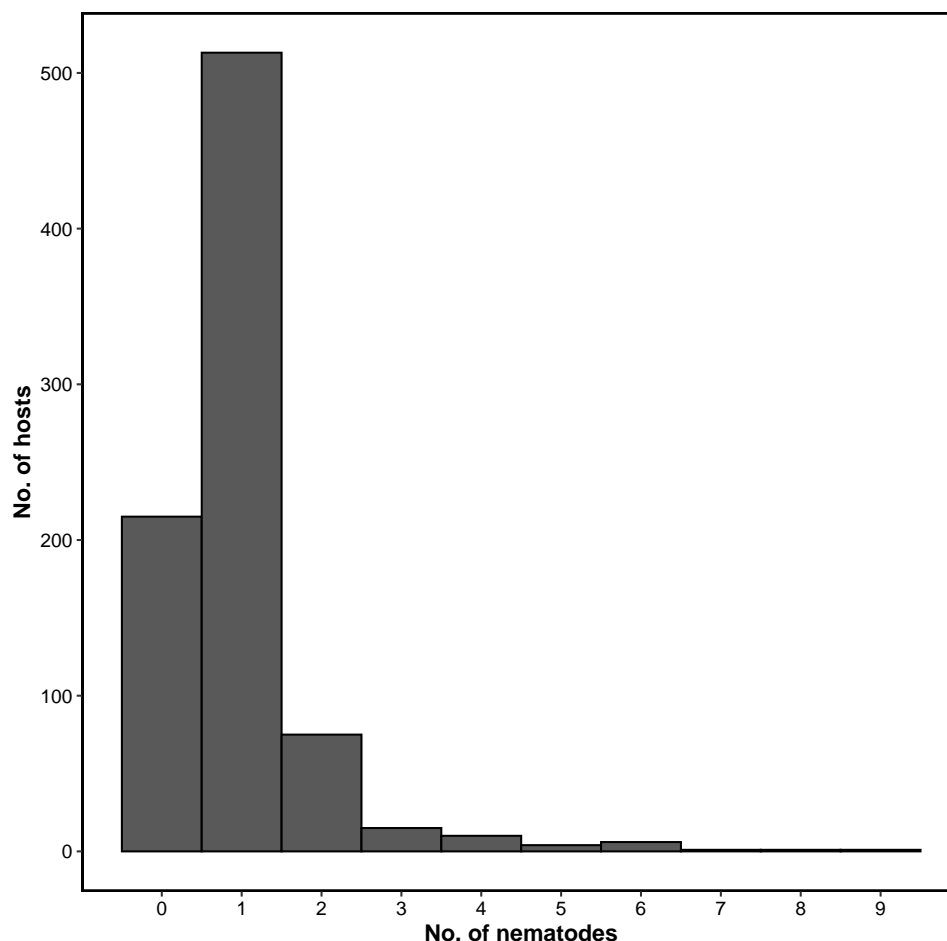


Figure S10: Distribution of the number of nematode OTUs infecting each host.

SI note 3: Details on model evaluation metrics

In multi-class classification models, the output is a probability distribution over the possible classes (i.e., the probabilities of a host being classified into one of the host infection profiles). This is achieved using a softmax function, which converts raw scores into probabilities that sum to 1. The predicted class is the one with the highest probability.

To evaluate the performance of our three-class classification model, we used a confusion matrix, which records the number of correctly and incorrectly classified instances for each class (92). The confusion matrix is structured as follows:

Table S2: Confusion Matrix for a 3-Class Model. A-C are host infection profiles.

Actual \ Predicted	Pred A	Pred B	Pred C
Actual A	TP_A	$FP_{B,A}$	$FP_{C,A}$
Actual B	$FP_{A,B}$	TP_B	$FP_{C,B}$
Actual C	$FP_{A,C}$	$FP_{B,C}$	TP_C

Each row represents the actual class, while each column represents the predicted class.

- TP_X (True Positives): Correctly classified instances of class X .

- $FP_{Y,X}$ (False Positives): Instances incorrectly classified as class Y when they actually belong to class X .

Since our dataset is imbalanced, we used evaluation metrics that give fair importance to each class based on the class size (i.e., fraction of hosts with the infection profile). These include accuracy, weighted precision, weighted recall, weighted F1-score, weighted balanced accuracy, and the Matthews Correlation Coefficient (MCC) (91).

Accuracy: Accuracy measures the overall correctness of the model:

$$\text{Accuracy} = \frac{TP_A + TP_B + TP_C}{N} \quad (\text{S2})$$

where N is the total number of samples.

Weighted Precision: Precision for class X is the proportion of correctly predicted X instances out of all instances predicted as X :

$$P_X = \frac{TP_X}{TP_X + \sum FP_{X,Y}} \quad (\text{S3})$$

The weighted precision is:

$$P_w = \sum_{X \in \{A,B,C\}} w_X P_X \quad (\text{S4})$$

where w_X is the proportion of actual instances of class X .

Weighted Recall: Recall (Sensitivity) for class X measures how many actual X instances were correctly classified:

$$R_X = \frac{TP_X}{TP_X + \sum FP_{Y,X}} \quad (\text{S5})$$

The weighted recall is:

$$R_w = \sum_{X \in \{A,B,C\}} w_X R_X \quad (\text{S6})$$

Weighted F1-score: F1-score is the harmonic mean of precision and recall for each class:

$$F1_X = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \quad (\text{S7})$$

The weighted F1-score is:

$$F1_w = \sum_{X \in \{A, B, C\}} w_X F1_X \quad (\text{S8})$$

Weighted Balanced Accuracy: Balanced accuracy accounts for class imbalance and is calculated as the mean recall across classes:

$$BA_X = \frac{TP_X}{TP_X + \sum FP_{Y,X}} \quad (\text{S9})$$

The weighted balanced accuracy is:

$$WBA = \sum_{X \in \{A, B, C\}} w_X BA_X \quad (\text{S10})$$

Matthews Correlation Coefficient (MCC): MCC is a more comprehensive metric that considers all values in the confusion matrix, and is, therefore, a balanced measure that can be used even if the classes are of very different sizes:

$$MCC = \frac{c \times s - \sum_{X \in \{A, B, C\}} p_X \times t_X}{\sqrt{\left(s^2 - \sum_{X \in \{A, B, C\}} p_X^2\right) \left(s^2 - \sum_{X \in \{A, B, C\}} t_X^2\right)}} \quad (\text{S11})$$

where:

c = sum of true positives across all classes

s = total number of samples

p_X = predicted counts for each class $TP_X + FP_{X,Y}$

t_X = actual counts for each class $TP_X + FP_{Y,X}$

MCC values range from -1 to $+1$. A coefficient of $+1$ indicates a perfect prediction, 0 indicates no better than a random prediction, and -1 indicates total disagreement between prediction and observation.

To further assess model performance, we compared all metrics against a theoretical no-skill classifier, whose expected values were analytically computed based on profile distributions using proportional guessing. In this approach, the classifier predicts each profile according to its prevalence in the dataset, favoring frequent profiles over rare ones. This reflects the natural class distribution without relying on learned patterns. For this classifier, accuracy, precision, and recall scale with profile frequencies, while balanced accuracy remains equivalent to a uniform guessing strategy, providing a simple baseline despite class imbalance. The theoretical MCC for an ideal random classifier is zero. This method offers a straightforward way to benchmark our trained model's performance against random chance, especially in datasets with imbalanced classes.

In addition, we used evaluation metrics in a one-vs-all manner, where the predicted class is considered positive, and the remaining classes are treated as negative. The model's performance was then evaluated across multiple threshold values (i.e., classifying a sample into class X only

if its probability exceeds threshold Y) by computing the Area Under the Receiver Operating Characteristic curve (AUC-ROC) and the Area Under the Precision-Recall Curve (PR-AUC) for each class separately.

ROC-AUC: The area under the receiver operating characteristic curve is a graphical representation of the actual positive rate (y-axis) versus the false positive rate (x-axis) of a model across different decision thresholds. The ROC-AUC score ranges from 0 to 1, where a score of 1 represents a perfect classification model, while a score of 0.5 represents a one-vs-all model with random guessing.

Although the ROC-AUC is a common measure, the number of true negatives in imbalanced data sets is very large, so even with a substantial number of false positives, the false positive rate might remain relatively small. This means that the ROC curve might not fully capture the cost of misclassifying a substantial number of the minority class instances. A better way to evaluate predictions in imbalanced data sets is by combining precision and recall metrics. Precision and recall provide a more granular understanding of a model’s performance because their trade-off highlights how well the model balances false positives and false negatives, offering insight into its effectiveness in identifying true cases under different thresholds.

PR AUC: To evaluate the tradeoff between precision and recall, the area under the PR curve provides a single number that summarizes the overall performance of a model across all possible classification thresholds. Like the ROC-AUC curve, the PR curve is calculated across all thresholds. We calculated a PR curve for each class. The no-skill baseline PR-AUC was derived from class prevalence, meaning that a random classifier’s expected precision equaled the fraction of positive instances in the dataset (i.e., the fraction of hosts in the infection profile out of all hosts).

SI note 4: Results of model evaluation

Our XGBoost model demonstrated strong overall performance, significantly predicting infection across all host profiles (**Figure S11**). All evaluation metrics exceeded those of a no-skill classifier, including accuracy (0.54), weighted precision (0.53), weighted recall (0.54), weighted F1-score (0.53), weighted balanced accuracy (WBA) (0.64), and Matthews correlation coefficient (MCC) (0.28) (**Figure S11A**).

The Area Under the Receiver Operating Characteristic Curve (AUC-ROC) was above the random value of 0.5 for all profiles (AUC = 0.726, 0.606, 0.738, for profiles 1, 2, and 3, respectively), indicating effective classification accuracy (**Figure S11B**). However, the relatively lower ROC curve for profile 2 (reflecting a low true positive to false positive rate) suggests that the model only weakly distinguished it from other profiles. The Precision-Recall Curve (PR-AUC) is useful for imbalanced datasets, as it captures the tradeoff between precision and recall. The PR-AUC was above the no-skill values (i.e., the fraction of hosts with each infection profile) for the three host profiles: AUC = 0.548 (no-skill of 0.322) for profile 1; AUC = 0.331 (no-skill of 0.244) for profile 2; and AUC = 0.656 (no-skill of 0.434) for profile 3 (**Figure S11C**). Overall, while the model performed well, it had greater difficulty accurately predicting profile 2.

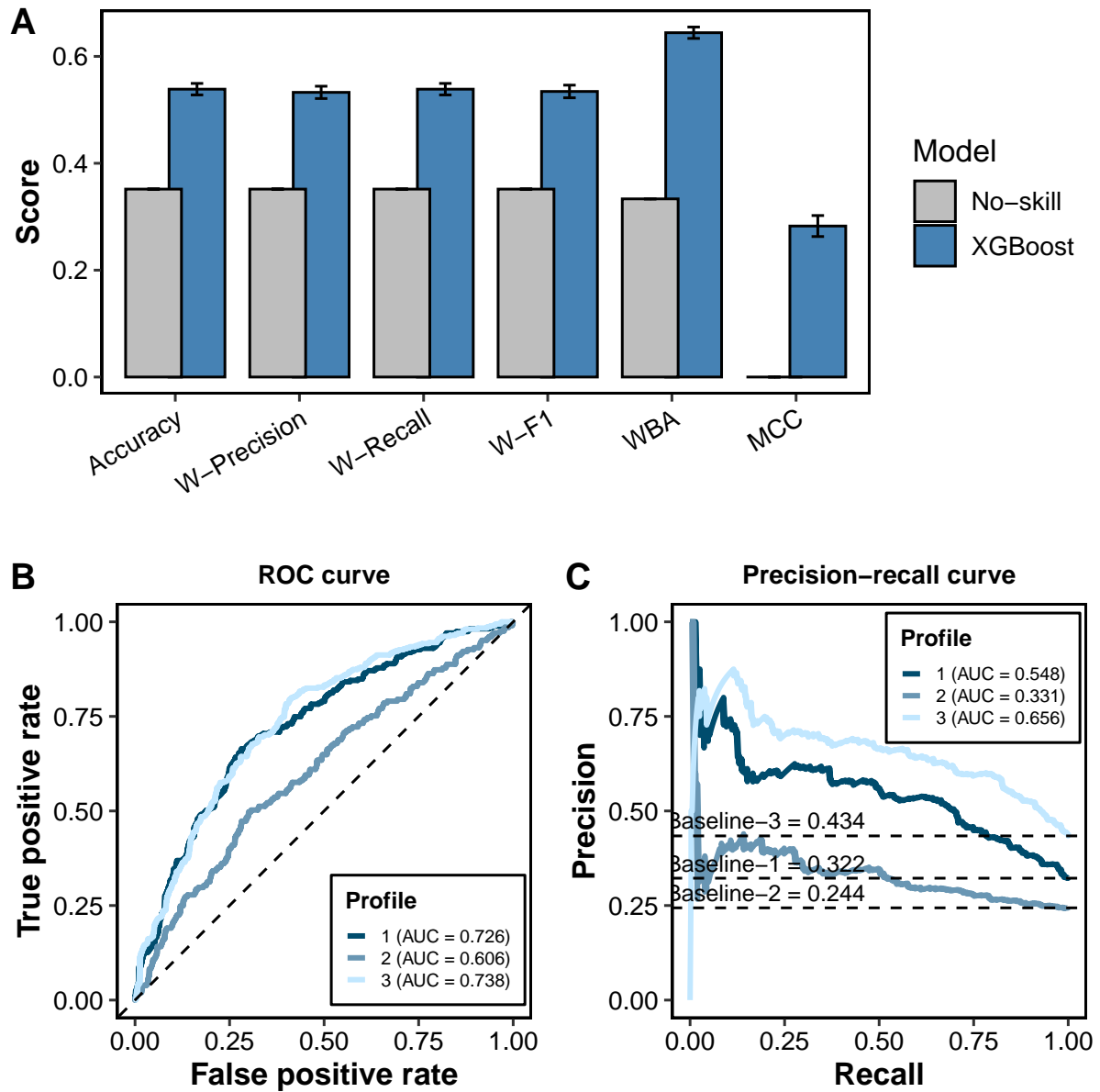


Figure S11: Evaluation of the XGBoost predictive model. (A) Comparison between the XGBoost model and a no-skill classifier in different evaluation metrics. The bars and error bars indicate the mean and standard deviation, respectively, of the three-fold models used for nested cross-validation. (B) The Area Under the Receiver Operating Characteristic Curve (AUC-ROC) for each profile separately. Colors indicate infection profiles, while the dashed black line indicates random model performance (AUC = 0.5). (C) Precision-Recall Curves (PR-AUC) for each profile separately. The dashed black lines indicate a no-skill model for each profile, derived from profile prevalence (i.e., a random expectation equal to the fraction of hosts with the infection profile). See **SI note 3** for detailed explanations on model evaluation.