Features that matter: evolutionary signatures that predict viral transmission routes

Maya Wardeh*[1,2], Jack Pilgrim[2], Melody Hui[2], Aurelia Kotsiri[2], Matthew Baylis[2], Marcus SC Blagrove*[2]

1) Department of Computer Science, University of Liverpool, Liverpool, UK

2) Institute of Infection, Veterinary and Ecological Sciences, University of Liverpool, Liverpool, UK

* = corresponding authors maya.wardeh@liverpool.ac.uk & marcus.blagrove@liverpool.ac.uk

ABSTRACT

Routes of virus transmission between hosts are key to understanding viral epidemiology. Different routes have large effects on viral ecology, and likelihood and rate of transmission. For example, respiratory and vector-borne viruses together encompass the majority of high-consequence animal and plant outbreaks. However, the specific transmission route(s) can take months to years to determine, undermining the efficiency of mitigation efforts. Here, we identify the vial features and evolutionary signatures which are predictive of viral transmission routes, and use them to identify potential routes for fully-sequenced viruses – we perform this for both viruses with no observed routes or viruses with potentially missing routes. This was achieved by compiling a dataset of 24,953 virus-host associations with 81 defined transmission routes, constructing a hierarchy of virus transmission encompassing those routes and 42 higher-order modes, and engineering 446 predictive features from three (virus, host, and network) perspectives. We integrated those data and features, to train 98 different ensembles of LightGBM classifiers, each incorporating five different class-balancing approaches. Using our trained ensembles, we demonstrated that all features contributed to the prediction for at least one of routes and/or modes of transmission, demonstrating the utility of our multi-perspective approach. Our approach achieved ROC-AUC=0.991, and F1-score=0.855 on average across all modelled transmission mechanisms; and was able to achieve high levels of predictive performance for high-consequence respiratory (ROC-AUC=0.990, and F1-score=0.864) and vector-borne transmission (ROC-AUC=0.997, and F1-score=0.921). Our work ranks the viral features in order of their

contribution to prediction, per transmission route, and hence identifies the genomic evolutionary signatures associated with each route. Together with the more matured field of viral host-range prediction, our predictive framework could provide early insights into the potential for, and pattern of viral spread; facilitate rapid response with appropriate measures; and significantly triage the time-consuming investigations to confirm the likely routes of transmission. Moreover, the strength of our approach in high-consequence transmission routes showcases that our methodology has direct utility to pandemic preparedness.

INTRODUCTION

Mounting an effective response to an emerging infectious virus requires determining all critical information, as quickly as possible. In recent years, significant focus has been placed on understanding what determines host range and potential for spillover[e.g.1–5]. However, the transmission route, i.e. the pathway a virus uses to physically get from an infected host to an uninfected host, still requires months or even years to thoroughly investigate; this was most apparent during the first six to nine months of the SARS-CoV-2 pandemic, where the relative importance of lingering aerosols versus fomite transmission was still being determined[6,7]. Furthermore, secondary transmission routes, such as sexual transmission of both Zika[8] and Ebola[9] viruses, are often only ascertained during or after significant outbreaks. The ability to identify all epidemiologically significant transmission routes of a virus, with high accuracy, computationally, with minimal information, and as quickly as possible, is therefore of paramount importance to future emerging infectious viruses.

The transmission routes of a virus are also fundamentally intertwined with its ecology, epidemiology[10], and its potential for host shifting and spillover[11]. They determine how the virus spreads within and between different host species and within populations, therefore influencing the potential, severity, and geographical extent of outbreaks. Transmission routes such as respiratory via droplets or aerosol can result in rapid virus spread through a dense population. Influenza A, many coronaviruses, and the more benign rhinoviruses all benefit from this transmission mechanism to

cause swift outbreaks worldwide[12]. Conversely, vector-borne viruses tend toward a more varied outbreak speed, for instance Usutu gradual spread across Europe[13], compared with the El Niño-driven rapid spread of Zika through South America[14] – but nonetheless can produce wide ranging and long-term establishment, for example dengue[15] and bluetongue[15]. Different epidemiological patterns are also seen for other sets of transmission routes, such as vertical, sexual, and water-borne[10,16].

Here, we describe a novel approach to investigate which features and evolutionary signatures of a virus may determine the transmission route between hosts. These predictive features are then used to rapidly identify the transmission routes of fully-sequenced viruses with no observed routes. We also consider the potential for additional, so far un-observed, transmission routes in known pathogens (i.e., already have at least one observed transmission route).

To facilitate computational prediction of transmission routes, we first compiled a dataset of known transmission routes of the animal and plant viromes, to their hosts, from published literature, and use these data to construct a hierarchy of transmission mechanisms. We then established a field-bridging and uniform methodology, across the plant and animal viromes, to define routes of transmission based on each virus-host association rather than a 'per virus' definition, because, in some cases, the same species or strain of virus may utilise a different range of transmission routes to infect different hosts. For example, Influenza A is faecal-orally transmitted in waterfowl[17] but undergoes respiratory transmission in humans[18]. In other cases, very closely related but different viruses may utilise a diverse set of transmission routes in different hosts. For instance, whilst many viruses in the family Flaviviridae are exclusively vector-borne, some are also vertically or sexually transmitted in both vertebrate and vector populations[19], further some, e.g. Hepatitis C virus, are blood-borne and do not replicate in arthropods[20]. Given the above two scenarios, we expanded the concept of virus phylogenetic neighbourhood[21] to incorporate characteristics of the hosts, **and the network linking viruses with their known hosts**, into a unified framework. We applied this framework, termed: 'Virus-host integrated neighbourhoods', to synthesise predictive features, capturing both virus and host similarities as well as the topological characteristics of the underlying network.

In order to enable parameterisation of transmission routes that are closely interlinked with host taxonomy (e.g. seed and pollen-borne routes are strictly limited to plants, milk-borne transmission in mammals, and egg related routes in non-mammalian animals), we incorporated similarity between hosts to differentiate between those categorically different routes.

Finally, as virus structure has been shown to constrain virus transmission[22], and biases in the virus genome composition can also inform the transmission mechanism deployed by the virus, as well as correlating with virus reservoirs and vectors[21], we synthesised a complementary array of variables from the full genome sequences of viruses.

We combined the above into lightGBM ensembles – a gradient boosting framework that constructs robust models by combining numerous weak learners (decision trees) through iterative training to minimise prediction errors. We used our trained ensembles to firstly identify which combinations are most predictive of 98 transmission routes deployed by animal- and plant-infecting viruses to their known hosts; and secondly to predict which of those mechanisms are applicable to virus-host associations without observed transmission routes. We were able to predict potential transmission routes for 3,115 such associations.

This is the most taxonomically broad study of its kind, to demonstrate the ability to use the sequence and morphological information increasingly available within the first few days of an outbreak to highly accurately identify the transmission pathways of a novel virus, across the animal and plant viromes. This will provide early insights into the potential for, and pattern of spread of a virus; facilitate rapid response with appropriate measures; and significantly triage the time-consuming investigations to confirm the likely routes.

RESULTS

Hierarchy of virus transmission: We identified 82 *routes* of virus transmission in animals and plants by searching the relevant literature. We constructed a hierarchy of transmission mechanisms incorporating those routes as the lowest classification (leaves), and higher classifications of

transmission (non-leaf nodes) which we termed *modes* (Figure 1.A). Supplementary Table 3 fully describes our routes/modes.

We captured data on 81 transmission *routes* of 4,446 viruses to 5,317 host species (total of 24,953 virus-host associations, Supplementary Table 4, Supplementary Datasets 1-2), by performing a series of literature searches (see methods). Where at least one *route* of transmission was identified, we used those data to populate higher levels (*modes*) in our hierarchy. Figure 1 illustrates the distribution of observed transmission routes and modes of virus-host associations across our suggested hierarchy (Figure 1.A), as well as between our viruses and hosts (Figures 1.B and 1.C).

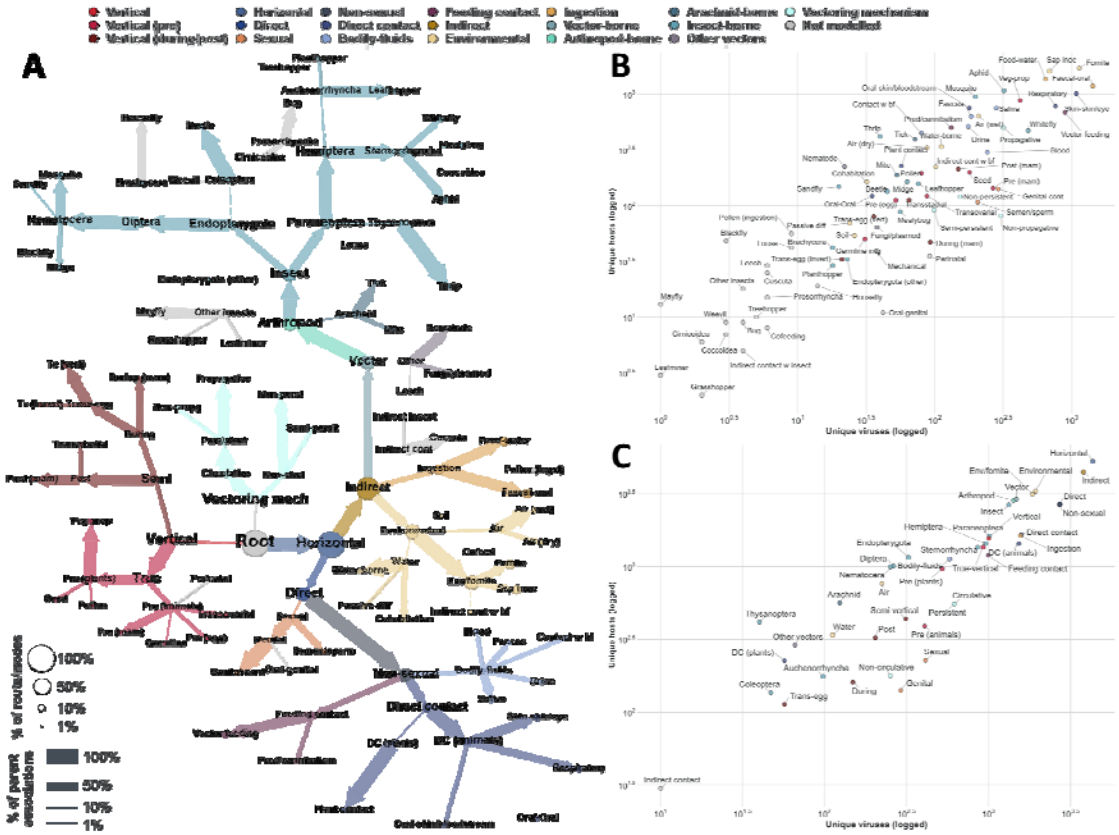


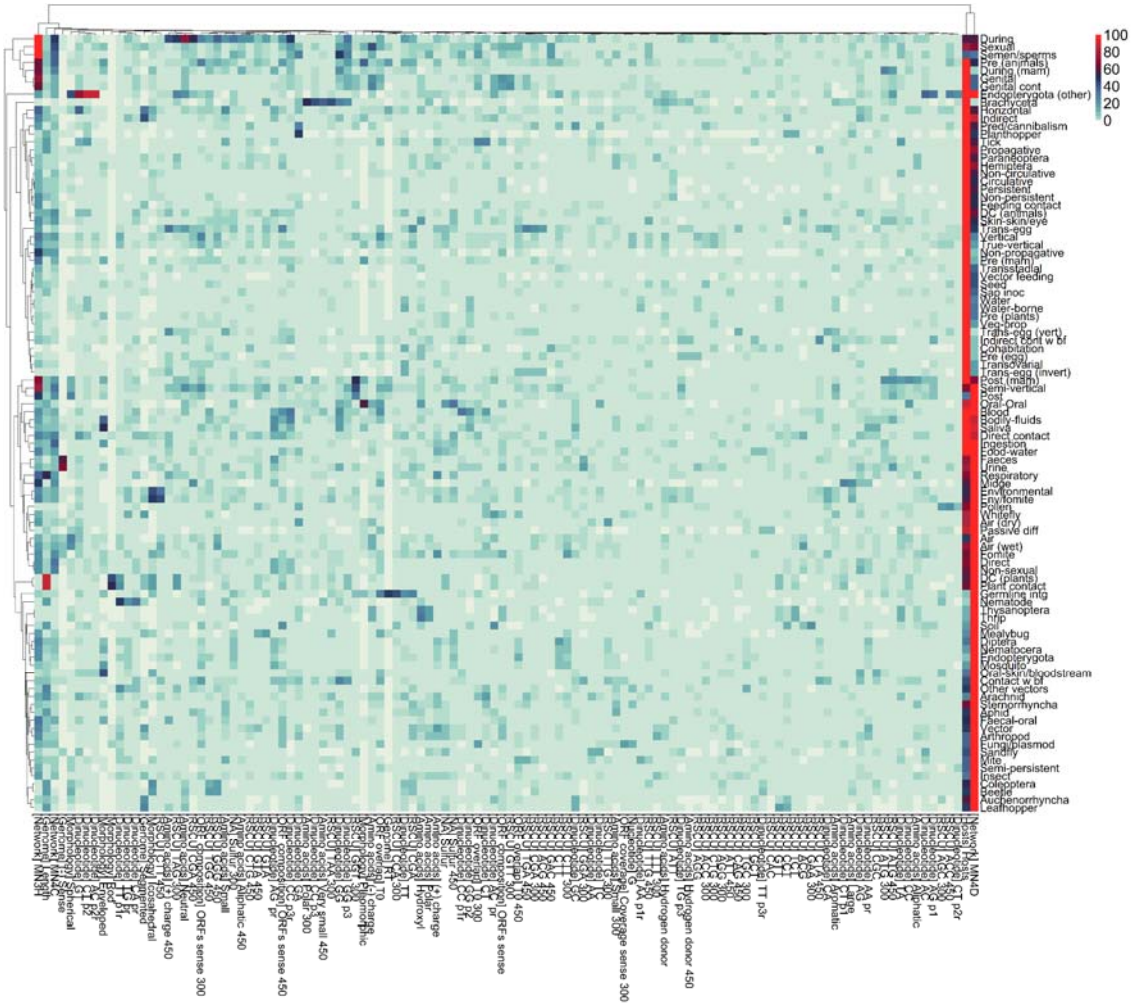
Figure 1 – Overview of observed transmission routes/modes. **Panel A – Our suggested hierarchy.** Nodes represent transmission routes/modes identified in this study. Edges link parent modes (nodes with at least one child) with their offspring (e.g. indirect and direct transmission are two modes of horizontal transmission). Nodes and edges are coloured by the mode of transmission; routes not modelled in this study (due to insufficient data, n=18, modelled routes = 59), and conceptual nodes (root and vectoring mechanism) are coloured in light grey. Node size is proportional to percent of unique virus-host associations (of 24,953 associations) where the virus is transmitted to the host species via the corresponding route/mode. Thickness of edges is proportional to the percent of the

parent associations identified to where the virus transmitted to the host species via the child route/mode (e.g. of 10,021 associations transmitted by arthropods [40.16% of included associations], 92.63% are insect-borne, and 9.5% are arachnid-borne). Supplementary Figure 5 visualises for each unique route/mode pair (no route/mode in the pair is an ancestor/offspring of the other), the percent of virus-host associations (of total included), whereby the virus is known to be transmitted to the host via both pathways. **Panel B – Transmission routes identified in this study.** Points represent transmission *routes* (Supplementary Table 3) and are coloured by transmission modes. X axis represents the number of observed unique viruses per route. Y axis represents the number of observed unique host species per route. **Panel C – Transmission modes identified in this study.** Given a virus-host association, we considered the virus to be transmitted to the host via a parent *mode* (e.g. Zika virus is insect-borne to humans), if it were transmitted by at least one *route* that is also an offspring/descendant of the parent mode (e.g. zika virus is mosquito-borne to humans), in our hierarchy (Figure 1.A). Points colours in panels B and C follow the same scheme as nodes in panel A.

Predictors of transmission mechanism: Of our total of 446 features, virus-host integrated neighbourhoods (Network, MN4D and MN3H, see methods and Supplementary Note 5) and hosts similarity (Hosts) features made the largest contributions to predictions across all routes/modes (Figure 2): 'MN4D' (top predictor of 51.02% of modelled pathways, see methods), 'Hosts' (45.19%), and 'MN3H' (3.06%). Those three features were in the top ten predictors of 96.94%, 98.98% and 72.45% of included route/modes, respectively (Figure 2). The remainder Network feature, 'MN4C' (see methods, Supplementary Note 5), was a top ten predictor in 45.92% of our routes/modes.

Of the genomic structure features (9 in total, Supplementary Table 5 fully describes all virus features utilised in this study), genome length, and whether the virus is 'segmented', 'retro-transcribing', or 'negative-sense', were top ten predictors of 39.8%, 13.27%, 4.08 and 3.06% of our routes/modes, respectively. Genome 'length' was an important predictor (locally normalised mean absolute SHAP value ≥20) for five routes/modes (Figure 2) including: direct contact (plants) (locally normalised absolute SHAP value = 85.92, vs globally normalised mean absolute SHAP value = 30.86), respiratory (54.34 vs 19.42), and Arachnid-borne (24.265 vs 9.91) transmission. Segmentation had locally normalised mean SHAPs ≥20 for four routes/modes: Auchenorrhyncha-borne (22.62 vs 14.04, locally normalised vs globally normalised), leafhopper-borne (29.40 vs 17.47), indirect (40.73 vs 17.07) and horizontal (22.63 vs 14.045) transmission. Retro-transcribing achieved a similar level of normalised mean SHAP values only for germline integration transmission (45.02 vs 14.71), and '(-) sense' for urine (69.18 vs 21.70) and faeces (57.94 vs 18.95) transmission.

Of our binary morphology features (16 in total, Supplementary Table 5), five were in the top ten predictors of our models: ‘icosahedral’ (17.35%), ‘enveloped’ (14.29%), ‘spherical’ (10.2%), ‘pleomorphic’ (6.12%), and ‘rod shaped’ (2.04%). These features had a locally normalised mean SHAP value ≥20 for 14 routes/modes, including: ‘enveloped’ for oral-skin/bloodstream (42 vs 20.44), and saliva (50.81 vs 15.23) transmission, ‘rod shaped’ for plant contact (53.56, 17.87), ‘icosahedral’ for environmental/fomite (47.18 vs 14.7), and ‘pleomorphic’ for oral-oral (57.68 vs 13.98) transmission.


100
80
60
40
20
0
Leafhopper
Coleoptera
Insect
Sandfly
Mite
Semi-persistent
Aphid
Fecal-oral
Vector
Arthropod
Eukaryotic
Mealybug
Nematode
Plant contact
Germine mtg
Air (wet)
Fomite
Passive diff
Air (dry)
Whitefly
Rodent
Fomite
Urine
Respiratory
Faeces
Edod-water
Ingestion
Direct contact
Bodily-fluids
Oral-Oral
Semi-vertical
Trans-egg (invert)
Transovarial
Cohabitation
Trans-egg (vert)
Vector feeding
Non-propagative
Vertical
Trans-egg
Skin-skin/eye
Feeding contact
Non-persistent
Circulative
Non-circulative
Propagative
Lick
Elanthopper
Fed/cannibalism
Indirect
Horizontal
Brachycera
Endopterygota (other)
Genital cont
Genital
During (mam)
Semen/sperms
During
Hemiptera
Diptera
Hymenoptera
Thysanoptera
Lepidoptera
Nematocera
Endopterygota
Mosquito
Oral-skin/bloodstream
Contact w bl
Other vectors
Arachnid
Sternorrhyncha
Hydrogen donor
Hydrogen donor 450
Coverage sense 300
Coverage 300
Coverage 450
Hydroxyl
Sulfur
Polar
Polar 300
Polar 450
Aromatic
Aromatic 450
Aromatic 300
Aromatic 450
Aromatic 300
Aromatic 450
Aromatic 300

predictions of the 98 route/mode modelled in this study (X-axis). We performed hierarchical clustering on both rows and columns, using the R package pheatmap. the resulting dendrogram is displayed (top and left).

Of ORF composition features (9 total, Supplementary Table 5), 'ORFs sense 450', 'ORFs sense 300', and 'ORFs sense' were top ten predictors for 14.29%, 12.24%, and 12.24% of our routes/modes, respectively. These features were important predictors for nine routes/modes, with 'ORFs sense 450' and 'ORFs sense 300' ranked 4th predictors for soil (21.55 vs 8.57), and vertical (25.77 vs 6.16) transmission, respectively.

Of ORF overlap features (21 total, Supplementary Table 5), 'T0', 'T0 450', and 'T0 300' were top ten predictors for 15.31%, 8.16%, and 5.1% of our models, respectively. These features were important predictors for four routes/modes, including 'T0 300' for tick-borne (27.29 vs 11.09). Of the remainder ORF category (ORF coverage, 6 features, Supplementary Table 5), only 'coverage sense 300' featured as a top ten predictor (for fungi/plasmodiophorids transmission (15.97, 7.24)).

Nineteen amino acid biases (57 total, Supplementary Table 5) were top ten predictors, six of which were top ten predictors for >10% of our models: 'sulphur' (13.27%), 'sulphur 450' (13.27%), '(+) charge' (12.24%), '(+) charge 450' (12.24%), 'hydroxyl' (11.22%), and '(-) charge' (10.2%). Our amino acid biases were important predictors for 21 routes/modes (Figure 2), and ranked 2nd predictors for five routes/modes: 'hydroxyl' bias for thrip- (30.18 vs 17.29) and Thysanoptera-borne (27.56 vs 16.09); 'polar 300' bias for Brachycera-borne (52.4 vs 15.97); '(+) charge 450' bias for trans-egg (vert) (21.82 vs 13.03); and 'neutral' bias for during (mammalian birth) transmission (72.86 vs 9.12).

Of nucleotide biases (4 total, Supplementary Table 5), only 'G' bias was a top ten predictor (three routes/modes), whereas 32 dinucleotide biases (128 total, Supplementary Table 5) were top ten predictors for at least one route/mode, with six biases ranked in the top ten predictors for >10% of our models: 'AG pr' (AG bias in the reverse complement of the sequence, 14.29%), 'GT' (13.27%), 'GA' (11.22%), 'GG' (11.22%), 'TT' (10.2%), and 'GG p3' (GG bias in position 3-1 within codon reading frames, 10.2%). Dinucleotide biases were important predictors for 37 routes/modes, and ranked 2nd

predictors for two routes: ‘TT p1r’ bias for nematode (46.08 vs 25.23), and ‘GG’ bias for semen/sperm transmission (45.14 vs 11.21).

Finally, of 190 'RSCU features (Supplementary Table 5), 44 ranked in the top ten predictors of at least one route/mode. Eight were top ten predictors for >5% of our models: ‘CGA’ (CGA bias in predicted codons of length≥36, 10.2%), ‘GGG 450’ (8.16%), ‘ATG 450’ (8.16%), ‘TAG 300’ (TAG bias in predicted codons of length≥300, 7.14%), ‘TGG 450’ (TGG bias in bias in predicted codons of length≥450, 7.14%), ‘GTT’ (7.14%), ‘ATT’ (6.12%), and ‘CCT’ (6.12%). RSCU features were important predictors for 25 routes/modes, and were ranked second predictors for two routes: ‘CTA 450’ for indirect contact with bodily-fluids (20.51 vs 11.02), and ‘GAA 300’ for soil (23.4 vs 9.3) transmission. ‘CCA 300’ had the largest globally normalised mean absolute SHAP value out of all RSCU features for transmission via germline integration (44.06 vs 14.4).

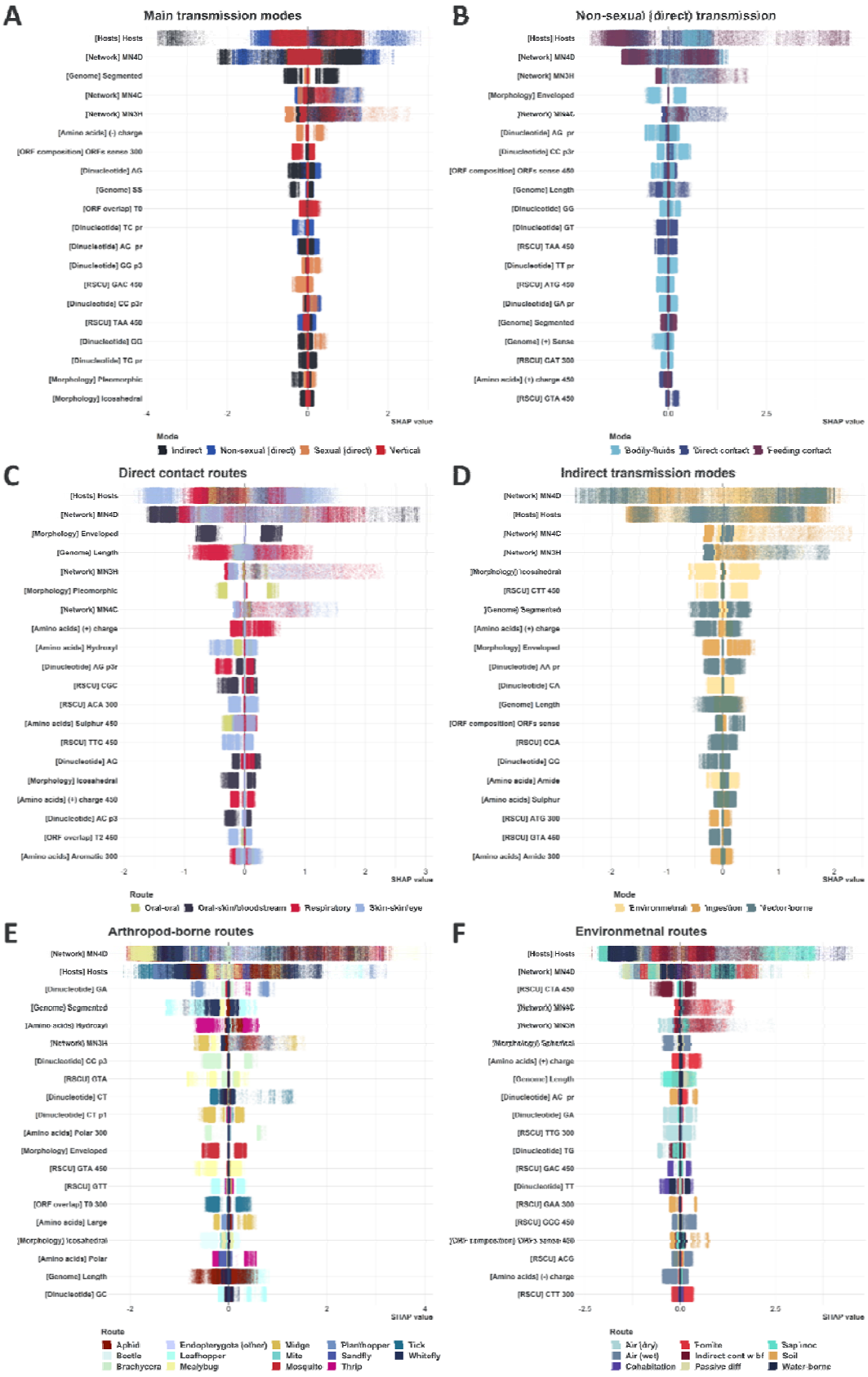
Instance-level predictors of transmission mechanism: Figure 3 visualises the instance-level contributions of top twenty features, by spread of variance in each sub-plot/category, for six categories: main transmission modes (3.A), direct (non-sexual) transmission modes (3.B), direct contact routes (3.C), indirect transmission modes (3.D), arthropod-borne routes (3.E), and environmental transmission routes (3.F). Overall, our virus-host integrated neighbourhoods (Network, MN4D) and hosts similarity (Hosts) features had the most spread of variance across the six categories. For our main modes of transmission (Figure 3.A) ‘segmented’ genome was the viral feature with most spread in variance: ranked 3rd predictor of indirect transmission (Locally normalised mean absolute SHAP value = 40.74 vs globally normalised = 17.07), but only 18th (6.95 vs1.44), 42nd (0.56 vs 0.25), and 87th (0.35 vs 0.08) predictor of sexual, non-sexual, and vertical transmission modes, respectively.

For our direct transmission modes (Figure 3.2), Enveloped status was the viral feature with most variance: ranked 3rd predictor of bodily-fluids transmission (37.16 vs 12.82), 63rd of feeding contact (0.11 vs 0.07), and 95th of direct contact transmission (0.10 vs 0.03). ‘enveloped’ was also the viral feature with most spread in variance in our direct contact routes (Figure 3.C): ranked 2nd predictor of skin/bloodstream (42.00 vs 20.44), but had virtually no influence on the remaining routes: 82nd for

skin-skin/eye (0.11, 0.05), 107th for respiratory (0.05, 0.02), and had no influence at all on oral-oral transmission (mean SHAP value =0).

Icosahedral structure was the viral feature with most spread in variance in our main indirect transmission modes (Figure 3.D): ranked 3rd predictor of environmental (40.31 vs 13.72), 27th of ingestion (1.37 vs 0.61), and only 63rd of vector-borne transmission (0.28 vs 0.17). Conversely, 'segmented' was an important predictor of vector-borne – ranked 4th (16.06 vs 9.84), and environmental (4.92 vs1.67) transmission modes, but not ingestion – ranked 102nd (0.04 vs 0.02). Whether the virus is segmented or not had also large spread in variance of contribution to predicting certain insect-borne routes/modes (Figure 3.D, Supplementary Figures 10.A and 10.D); ranked as 3rd predictor of Auchenorrhyncha- (22.63 vs 14.05) and leafhopper-borne (29.4vs 17.47) transmission; 4th of whitefly- (15.54vs 7.76), Sternorrhyncha- (7.94 vs 5.16), and insect-borne (7.82, vs 5.62) transmission. But it had no influence (mean SHAP value = 0) on Brachycera, Coleoptera, Diptera, Endopterygota, midge, mite, mosquito, Nematocera, planthopper, sandfly, thrip, Thysanoptera, and tick-borne transmission.

'CTA 450' bias and 'spherical' morphology were the viral feature with most spread in variance when predicting environmental transmission routes (Figure 3.E). 'CTA 450' ranked 2nd predictor of transmission via indirect contact with bodily-fluids (20.51, 11.02), but only 100th of air (dry) (0.02, 0.01) transmission. Spherical morphology ranked 3rd predictor of air (wet) transmission (24.21, 8.62), but had no influence on soil and fomite transmission (mean SHAP value = 0). Supplementary Figures 8-12 visualise the remainder routes/modes not included in Figure 3.


A
Main transmission modes
[Hosts] Hosts
[Network] MN4D
[Genome] Segmented
[Network] MN4C
[Network] MN3H
[Amino acids] (-) charge
[ORF composition] ORFs sense 300
[Dinucleotide] AG
[Genome] SS
[ORF overlap] T0
[Dinucleotide] TC pr
[Dinucleotide] AC pr
[Dinucleotide] GC p3
[RSCU] GAC 450
[Dinucleotide] CC p3r
[RSCU] TAA 450
[Dinucleotide] GG
[Dinucleotide] TG pr
[Morphology] Pleomorphic
[Morphology] Icosahedral
-4
-2
0
2
SHAP value
Mode
Indirect
Non-sexual (direct)
Sexual (direct)
Vertical
B
Non-sexual (direct) transmission
[Hosts] Hosts
[Network] MN4D
[Network] MN3H
[Morphology] Enveloped
[Network] MN4C
[Dinucleotide] AG pr
[Dinucleotide] CC p3r
[ORF composition] ORFs sense 450
[Genome] Length
[Dinucleotide] GG
[Dinucleotide] GT
[RSCU] TAA 450
[Dinucleotide] TT pr
[RSCU] ATG 450
[Dinucleotide] GA pr
[Genome] Segmented
[Genome] (+) Sense
[RSCU] CAT 300
[Amino acids] (+) charge 450
[RSCU] GTA 450
0.0
2.5
SHAP value
Mode
Bodily-fluids
Direct contact
Feeding contact
C
Direct contact routes
[Hosts] Hosts
[Network] MN4D
[Morphology] Enveloped
[Genome] Length
[Network] MN3H
[Morphology] Pleomorphic
[Network] MN4C
[Amino acids] (+) charge
[Amino acids] Hydroxyl
[Dinucleotide] AG p3r
[RSCU] CGC
[RSCU] ACA 300
[Amino acids] Sulphur 450
[RSCU] TTC 450
[Dinucleotide] AC
[Morphology] Icosahedral
[Amino acids] (+) charge 450
[Dinucleotide] AC p3
[ORF overlap] T2 450
[Amino acids] Aromatic 300
-2
-1
0
1
2
3
SHAP value
Route
Oral-oral
Oral-skin/bloodstream
Respiratory
Skin-skin/eye
D
Indirect transmission modes
[Network] MN4D
[Hosts] Hosts
[Network] MN4C
[Network] MN3H
[Morphology] Icosahedral
[RSCU] CTT 450
[Genome] Segmented
[Amino acids] (+) charge
[Morphology] Enveloped
[Dinucleotide] AA pr
[Dinucleotide] CA
[Genome] Length
[ORF composition] ORFs sense
[RSCU] CCA
[Dinucleotide] CC
[Amino acids] Amide
[Amino acids] Sulphur
[RSCU] ATG 300
[RSCU] GTA 450
[Amino acids] Amide 300
-2
-1
0
1
2
SHAP value
Mode
Environmental
Ingestion
Vector-borne
E
Arthropod-borne routes
[Network] MN4D
[Hosts] Hosts
[Dinucleotide] GA
[Genome] Segmented
[Amino acids] Hydroxyl
[Network] MN3H
[Dinucleotide] CC p3
[RSCU] GTA
[Dinucleotide] CT
[Dinucleotide] CT p1
[Amino acids] Polar 300
[Morphology] Enveloped
[RSCU] GTA 450
[RSCU] GTT
[ORF overlap] T0 300
[Amino acids] Large
[Morphology] Icosahedral
[Amino acids] Polar
[Genome] Length
[Dinucleotide] GC
-2
0
2
4
SHAP value
Route
Aphid
Beetle
Brachycera
Endopterygota (other)
Leafhopper
Mealybug
Midge
Mite
Mosquito
Planthopper
Sandfly
Thrip
Tick
Whitefly
F
Environmetnal routes
[Hosts] Hosts
[Network] MN4D
[RSCU] CTA 450
[Network] MN4C
[Network] MN3H
[Morphology] Spherical
[Amino acids] (+) charge
[Genome] Length
[Dinucleotide] AC pr
[Dinucleotide] GA
[RSCU] TTG 300
[Dinucleotide] TG
[RSCU] GAC 450
[Dinucleotide] TT
[RSCU] GAA 300
[RSCU] GGG 450
[ORF composition] ORFs sense 450
[RSCU] ACG
[Amino acids] (-) charge
[RSCU] CTT 300
-2.5
0.0
2.5
SHAP value
Route
Air (dry)
Air (wet)
Cohabitation
Formite
Indirect cont w bf
Passive diff
Sap inoc
Soil
Water-borne

.1

241

Figure 3 – Instance-level feature-contribution to various transmission routes/modes. Instance-level SHAP values provide means to quantify the contribution of each selected feature made to that particular (virus-host association) prediction. Here, we averaged instance-level SHAP values generated by all constituent models of each of our top-10 ensembles (50 models per each included route/mode). In each sub-plot, features were ordered by the spread of their variance (max(variance)-min(variance) across all routes/modes included in each sub-plot), and the top 20 features (from most to least spread) were selected. Points represent virus-host associations (instances) and are coloured by the underlying route/mode. The Y axes represent the selected features (category of each features between brackets). The X axes represent SHAP values. Positive SHAP values indicate that the feature has contributed towards a positive prediction for the instance (the virus is transmitted to the host species via route/mode). Negative SHAP values indicate that the feature has contributed towards a negative prediction (the virus is not transmitted to the host via route/mode). Larger magnitudes indicate that the feature has a stronger influence on the prediction for the given instance.

Prediction of routes/modes: We were able to predict (mean top-10 ensemble probability > 0.5) at least one route/mode for 3,108 out of 3,708 (83.82%) virus-host instances (2,004 viruses (8.1.63%) to 1,300 host species (87.84%)) (Figure 4); Of those instances, 2,969 were predicted to be transmitted horizontally (80.07% vs 98.73% of observed associations), and 249 were predicted to be transmitted vertically (6.715% vs 15.15%). Indirect (57.50% vs 81.25%), direct (34.98% vs 41.77%), non-sexual (32.42% vs 41.65%), and ingestion (26.54% vs 26.76%) *modes* were the most predicted after horizontal transmission. The faecal-oral *route* was the most predicted route of transmission for unknown associations (23.33% vs 20.07%), followed by sap inoculation (9.47% vs 19.36% - a plant-only route), vector feeding (8.60% vs 15.81%), and fomite (7.15% vs 23.93%).

Additionally, our models made in-sample predictions (routes/modes predicted with mean top-10 ensemble probability>0.5 for virus-host associations with at least one observed transmission route in our dataset) for 6,701 virus-host associations (26.85% of total associations). The top additional routes/modes predicted in-sample were as follows: true vertical transmission in plants - pre-plants (649 additional associations, 9.685% of total in-sample predictions vs 9.86% of observed and 6.96% of predicted for unknowns); non-sexual (direct) transmission (639, 9.535% vs 41.65% and 32.42%); transovarial transmission (611, 9.12% vs 0.82% and 4.64%); and vertical transmission (597, 8.91% vs 15.15% and 6.715%). Figure 4 visualises both in-sample as well as out-of-sample (unknowns) predicted routes/modes.

Given a virus-host association, we constructed a representative set of *unique transmission pathways* by traversing our hierarchy (Figure 1.A) from routes to root, and including all transmission *routes* of the given virus to the focal host, as well as any transmission *modes* that are not ancestors of any already included routes/modes. We predicted that 1,068 instances have a single unique transmission pathway (29.33% of unknown instances vs 30.89% of associations with at least one observed route and 64.525% for our in-sample predictions); 953 (26.17% vs 33.25% and 22.53%) to have two unique pathways; and 1,020 (28.01% vs 35.855% and 12.94%) to have three or more unique pathways.

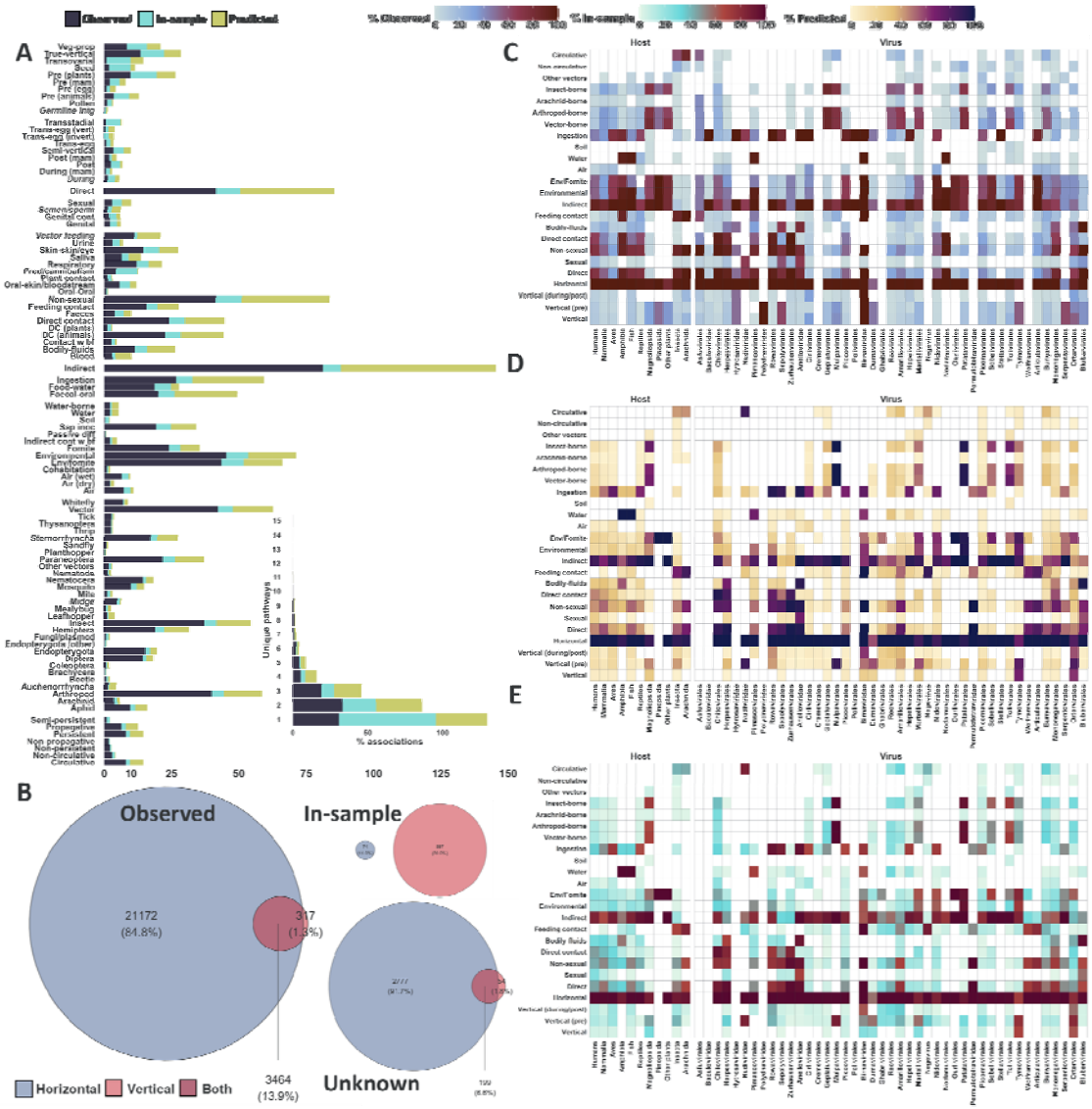


Figure 4 – Predicted transmission routes and modes. Panel A – Proportion of predicted unknowns (yellow, n=3,108, no known transmission route, mean probability cut-off>0.5), in-

sample predictions (cyan, n=6,701, hitherto unobserved routes predicted with mean probability cut-off>0.5 for associations with at least one observed route) and observed virus-host instances (dark blue, n= 24,953, at least one observed transmission route). Vertical and horizontal modes were removed from the bar plot for better visualisation (represented in panel B as Venn diagrams). The inset (bottom) represent the percent of unique pathways for unknowns and in-sample predictions, as well as observed for each virus-host association. Supplementary Figures 5, 6, and 7 visualise the percent of virus-host associations, whereby the virus is observed, predicted within sample, and predicted for unknowns, to be transmitted to the host via each pair of unique pathways, respectively. **Panel B – Horizontal and vertical transmission.** Venn diagrams represent horizontal transmission (blue) and vector transmission modes (red) for observed, in-sample predictions (with at least one previously observed route, but that is not observed to be transmitted via the corresponding mode), and unknown (out of sample) predictions (instances without observed routes), respectively. **Panel C – Proportion of host-virus instances transmitted by each main route/mode per each host group or virus order.** Rows represent main transmission modes/routes. Columns represent main host groups, and virus orders. Proportions are calculated by the number of instances known to be transmitted via given route/mode per each category (e.g. humans), divided by the total number of instances in the category. Some routes/modes were grouped together for better visualisation (e.g. Vertical (pre), insect-borne). **Panel D – Proportion of host-virus instances, without previously observed transmission route, predicted to be transmitted by each main route/mode per each host group or virus order.** Rows represent main transmission modes/routes. Columns represent main host groups, and virus orders. Proportions are calculated by the number of instances predicted to be transmitted via given route/mode per each category, divided by the total number of instances in the category. **Panel E – Proportion of host-virus instances, with at least one previously observed transmission route, predicted to be transmitted by each main route/mode per each host group or virus order.** Rows represent main transmission modes/routes. Columns represent main host groups, and virus orders. Proportions are calculated by the number of instances predicted to be transmitted via given route/mode per each category, divided by the total number of instances in the category.

Prediction dependencies: We utilised Mutual Information (MI) to quantify the relationship between predictions (top-10 ensemble mean probability > 0.5) for a given route/mode and those of its sibling route(s)/modes(s) - children of the same parent node in our hierarchy (Figure 1). Figure 5.A visualises the resulting normalised MI estimates. Our normalised MI ranged between 0.00002 (predation/cannibalism) and 0.027 (Brachycera-borne), suggesting a very weak to weak correlation (very limited to limited relationship or dependency) between the predictions of the focal route and those of its siblings. The routes with highest normalised MI were: Brachycera-borne (0.027), air (wet) (0.025), mosquito-borne (0.022), and midge-borne (0.021). The modes with highest MI were: during (mammalian birth) (0.023), true vertical (0.022), post (birth, e.g. via breast feeding) (0.018), and Arachnid-borne (0.017).

Model performance: We evaluated the performance of our predictive framework against held-out test sets (see methods and Supplementary Note 7 for full details). Overall our framework achieved average ROC-AUC = 0.988±0.017 (values closer to 1 indicates high accuracy in distinguishing between positive and negative instances); and average F1-score (>0.5 probability cut-off, commonly used for imbalanced classes, not used in tuning or selection of models) = 0.806±0.169 (values closer to 1 indicates high accuracy in identifying positive instances while minimising potentially false positives). We ranked our class-balancing ensembles (n=50, see methods), trained for each pathway (n=98), based on the average of four metrics: AUC, PR-AUC, Precision, and 1-Brier score, we selected the top-10 performing ensembles to produce final predictions. Our top-10 ensembles achieved ROC-AUC=0.991±0.012, and F1-score=0.855±0.143. Figures 5.B and 5.C visualise the performance of constituent models over 50 iterations, as well as for constituent models of our top-10 performing ensembles against our selection metrics, respectively. Table 1 lists the average performance against ten metrics. Supplementary Figures 17 and 18 visualise the performance of constituent models against all ten metrics. Supplementary Figure 19 illustrates our post-hoc assessment of in-sample predictions of our top-10 selected ensembles of five class balancing techniques at >0.5 mean probability threshold.

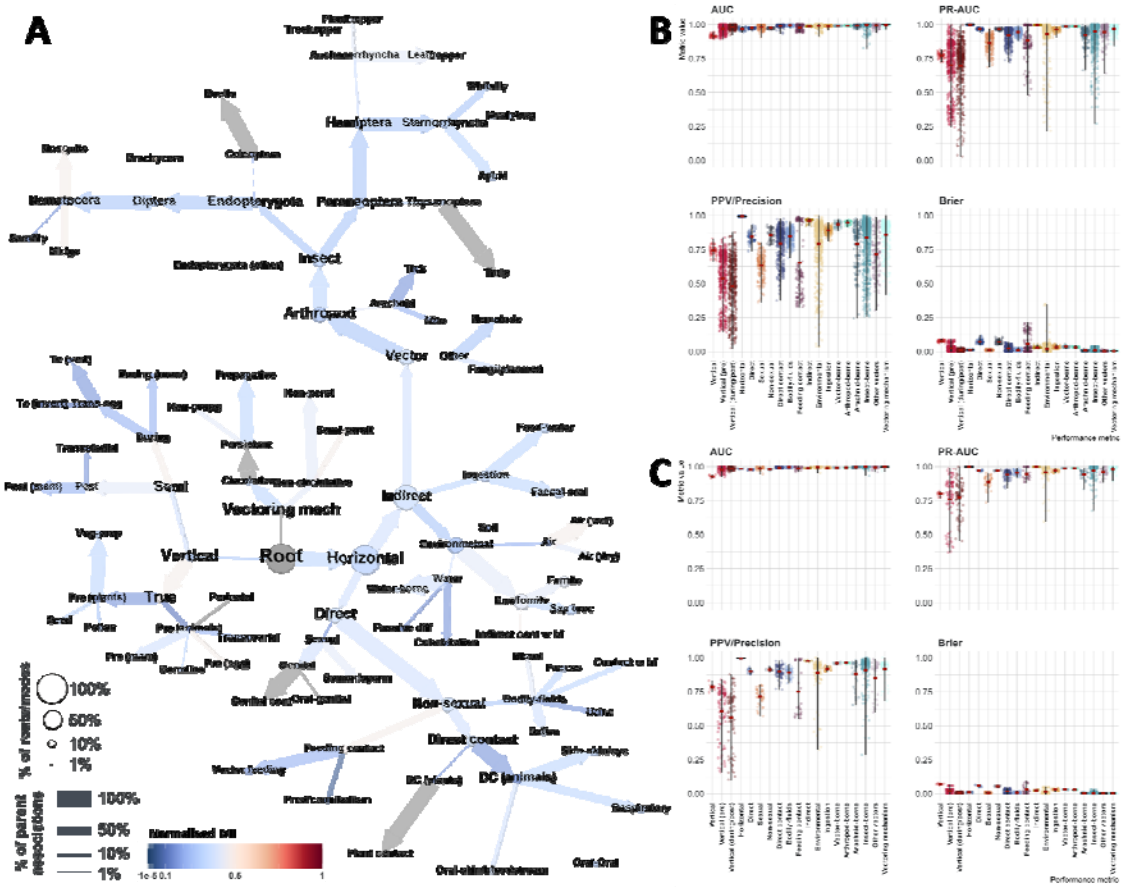


Figure 5 – Prediction dependencies and performance assessment. Panel A – Prediction dependencies. Nodes represent transmission routes/modes modelled in this study. Edges indicate our hierarchy (Figure 1). Nodes are coloured by Normalise Mutual Information (MI) estimates between the mean probabilities (derived from our top-10 ensembles) of instances predicted by the route/mode represented by the node (mean probability > 0.5), and corresponding instance-wise probabilities of their siblings (nodes with the same parent node). Grey nodes indicate structural nodes (not transmission related), and only children (routes without modelled siblings). Nodes sizes and thickness of edges are same as Figure 1. MI quantifies how much knowing the value of one variable can tell us about the value of the other variable. If the resulting estimate is high, it indicates a strong relationship or dependency between the variables, meaning that knowing one variable provides useful information about the other. To assess the statistical significance of the MI estimates, we compared each estimate to a null distribution using bootstrapping (n=2,000). Supplementary Figure 13 visualises the resulting p-values. Supplementary Figures 14-16 visualises decencies between probabilities for routes/modes and knowledge of their siblings, predictions of their siblings, as well as knowledge of routes/modes and resulting probabilities of their siblings, respectively. **Panel B – Performance assessment of constituent models of our ensembles of five class balancing techniques against 50 held-out test sets (per route/mode) at >0.5 probability threshold.** Points represent mean values of each performance metric, resulting from each of 50 individual transmission route/mode class-balancing ensembles (n=98). Violin plots show the kernel probability density of the data at different values (per mode). Red horizontal lines indicate the mean over all runs per mode. Points and violin plots are coloured by transmission mode. **Panel C – Performance assessment of constituent models our top-**

10 selected ensembles of five class balancing techniques against 10 held-out test sets (per route/mode) at >0.5 probability threshold. Points represent mean values of each performance metric, resulting from 10 individual class balancing ensembles per route/mode (n=98). Violin plots show the kernel probability density of the data at different values (per mode). Red horizontal lines indicate the mean over all runs per mode. Points and violin plots are coloured by transmission mode. Supplementary Table 9 provides full definitions of included performance metrics. For Brier score, in panels B and C, values closer to 0 indicate better performance, and those closer to 1 indicate worse performance.

	All Ensembles (n=50)	**Top-10 Ensembles**
ROC-AUC	0.988±0.017 [Vertical = 0.917 (±0.011), Cohabitation = 0.999 (± 0.0004)]	0.991±0.012 [Vertical = 0.931 (±0.008), Cohabitation = 0.999 (± 0.0001)]
PR-AUC	0.891±0.147 [Trans-egg (invert)= 0.332 (±0.167), Horizontal= 0.999 (±0.0002)]	0.922±0.109 [Transovarial= 0.446 (±0.052), Horizontal= 0.999 (±0.00004)]
F1-score	0.806±0.169 [Trans-egg (invert)= 0.237 (±0.086), Horizontal= 0.994 (±0.001)]	0.855±0.143 [Trans-egg (invert)= 0.293 (±0.073), Horizontal= 0.995 (±0.0007)]
NPV	0.987±0.048 [Horizontal= 0.546 (±0.091), Mealybug-borne= 0.999 (± 0.0002)]	0.987±0.042 [Horizontal= 0.601 (±0.059), Passive diffusion= 0.999 (± 0.0002)]
PPV/ Precision	0.755±0.206 [Trans-egg (invert)= 0.149 (±0.06), Horizontal= 0.995 (±0.001)]	0.829±0.179 [Trans-egg (invert)= 0.180 (±0.049), Horizontal= 0.996 (±0.0008)]
Recall/ Sensitivity	0.896±0.105 [Seed= 0.605 (±0.072), Vector feeding= 0.995 (±0.004)]	0.905±0.09 [Seed= 0.615 (±0.089), Vector feeding= 0.995 (±0.003)]
Specificity	0.976±0.052 [Horizontal= 0.656 (±0.066); Sandfly-borne= 0.999 (±0.001)]	0.984±0.037 [Horizontal= 0.672 (±0.061); Sandfly-borne= 0.999 (±0.001)]
MCC	0.797±0.156 [Trans-egg (invert)= 0.305 (±0.101), Vector feeding= 0.982 (±0.007)]	0.847±0.133 [Trans-egg (invert)= 0.384 (±0.079), Vector feeding= 0.990 (±0.0035)]
Brier	0.018±0.024 [Feeding contact= 0.144 (±0.047), Sandfly-borne= 0.002 (±0.001)]	0.014±0.016 [Vertical= 0.070 (±0.005), Passive diffusion= 0.001 (±0.0003)]
TSS	0.872±0.111 [Seed= 0.578 (±0.069), Vector feeding= 0.992 (±0.004)]	0.889±0.096 [Seed= 0.593 (±0.083), Vector feeding= 0.994 (±0.004)]

Table 1 – Average performance and standard deviation of all ensembles, and top-10 selected ensembles, across all routes/modes. Brier scores range from 0 (best performance) to 1 (worst performance); and MCC values range from +1 (best performance) to -1 (worst performance). Values in square brackets indicate worst and best performing ensembles, respectively. Top-10 ensembles were selected based in the average of four metrics (AUC, PR-AUC, PPV/Precision and Brier (1-actual score)). Supplementary Figures 17-18 visualise performance metrics of constituent models. Supplementary Figure 20 provides post-hoc assessment of in-sample predictions of our top-10 selected ensembles. Supplementary Dataset 3 lists performance assessment results for each route/mode.

DISCUSSION

This study uncovers the specific viral features and evolutionary signatures which predict the transmission routes each virus utilises between their respective hosts. We incorporated these features

and signatures to predict the transmission routes of 2,004 fully-sequenced viruses for which there are no known transmission routes/modes to 1,300 host species. Additionally, we identified an additional 19,396 transmission routes/modes potentially un-observed in 2,147 viruses and 2,355 host species with at least one route/mode observed. These predictions were made across a total of 4,076 animal and plant viruses. Consequently, this is the most taxonomically broad study of its kind, and is the first to utilise viral sequence, morphology, and host information, which are increasingly available within the first few days of an outbreak. This study, therefore, showcases the potential to provide early insights into the epidemiology of a newly emerging virus, and hence can be used to facilitate rapid response and significantly triage the time-consuming investigations to confirm the routes.

To achieve this, we compiled a dataset of routes of transmission, and constructed a hierarchy of transmission modes (Figure 1A); we then generated predictive features from three complementary perspectives (viruses, hosts, and network), and applied those features to train ensembles of lightGBM models in different subset of our data for each of our 98 routes/modes. Our ensembles indicated that our network (virus-host integrated neighbourhoods) and host features were highly predictive of all transmission routes/modes (Figure 2); however, all our virus features were also predictive, or improved the predictions, of at least one route/mode, enabling us to distinguish between closely related routes/modes. This highlights the advantage of our multi-perspective approach to investigating mechanisms of virus transmission, and further emphasises the applicability of multi-perspective approaches, in line with the significant promise they have shown in predicting virus-host associations[1,2,23,24].

Application of multiple perspectives in predicting transmission routes

All of our features, from across the three (virus, host and network) perspectives, contributed to the predictions of at least one route/mode, demonstrating the utility of our multi-perspective approach. Other existing approaches which aim to predict viral phenotypes solely from one perspective – such as the viral sequence – will miss key features from the host and network perspectives which would

enhance their predictive performance. For example, our virus-host integrated neighbourhoods (network), and host similarity (hosts) features were the most informative predictors across all routes/modes, providing the large-scale structure of the viral transmission route landscape. Our host similarity was top-10 predictor of 97 routes (ranked 12th for nematode-borne transmission - average locally normalised absolute SHAP value = 72.86, vs average normalised mean absolute SHAP value = 35.59). Contribution of our network features varied between 'MN4D' (34.75 vs 73.11), 'MN3H' (19.31 vs 7.50), and 'MN4C' features (8.54 vs 3.10)

Our 442 viral features (often referred to as viral evolutionary signatures[21]) further enhanced accuracy and explainability at higher resolutions of the individual association and route/mode levels, and therefore improving distinction between similar sister routes/modes. For instance, envelope status was an important predictor of mosquito-borne transmission (ranked 3rd) but has no effect on predicting sister routes midge- and sandfly-borne transmission. Conversely, 'CT p1' bias (refer to Supplementary Note 3 and Supplementary Table 5 for full definition) was predictive of midge-borne transmission (ranked 4th) but has no effect on predicting mosquito-borne transmission, and negligible effect on predicting sandfly-borne transmission. This is mirrored throughout our hierarchy, including routes that are often interlinked: '(+) charge 450' bias ranked 4th predictor of faecal-oral transmission (locally normalised SHAP = 12.64), but only 9th for transmission via ingestion of food/water (4.09). Furthermore, and in line with previous studies[22], envelop status was a top predictor (ranked 5th) of faecal-oral transmission, but has no effect on predicting transmission via ingestion of food/water (Supplementary Figure 11.B). Indeed, our viral features can differentiate between mechanistically similar routes, for instance, 'spherical' morphology was highly predictive (ranked 3rd) of transmission via inhalation of wet environmental particles (air (wet)) but had little effect on predicting respiratory transmission (ranked 110th).

We next applied our trained ensembles to predict transmission mechanisms for viruses with no observed routes to their known hosts (3,108 virus-host associations). Proportionally, direct transmission routes were overall more likely to be predicted than indirect (Figure 4). However,

ingestion and faecal-oral were the most likely unobserved individual routes to be predicted. We also noted a significant underestimation in the number of plant viruses potentially transmitted by insect. Looking specifically at predictions for arthropod-borne routes/modes, there are 608 associations (16% of associations without observed routes in our dataset). Of these 24% are to animals, and 76% to plants. The greatest proportion is within the Hemiptera-borne routes (354 associations); the Hemiptera being a superorder of insects which contains the majority of the plant vectors[25]. Given the relative difficulty, time, and expense of demonstrating arthropod-borne transmission, which requires vector competence studies[26], it is not surprising that many such routes remain undetermined. Our approach could readily be used to triage the vast numbers of potential competence studies into high-likelihood and high-priority combinations.

Strength of predictions in potentially high-consequence transmission routes

Routes of transmission are fundamentally intertwined with virus ecology and epidemiology[10]; they determine how a virus spreads within and between host species and populations, therefore having a large effect on the rate of spread (R_0) and, for some routes, the geographical range of outbreaks[27,28]. Respiratory and vector-borne transmission frequently results in high-consequence outbreaks[12,15], and our methodology is particularly effective at identifying viruses with these two mechanisms. For example, for respiratory viruses, our top-10 selected ensembles approach achieved mean ROC-AUC=0.990, and mean F1-Score = 0.864; whereas our arthropod-borne classifiers (n=26) averaged ROC-AUC=0.997, and F1-score = 0.921 (Supplementary Dataset 3). This strength is likely driven by the large amount of data (e.g. 40% of our virus-host associations had at least one arthropod-borne route of transmission), and the large number of human-virus data assigned to respiratory transmission (33% of virus-human associations). Given the importance of these two classes from previous high-consequence outbreaks, the high density of data serves to improve our pipeline's predictive performance where it is needed most.

Limitations

We acknowledge certain methodological limitations and shortcomings in our study.

- Our approach is restricted to fully sequenced viruses with at least one known host species, therefore it cannot be applied to viruses without association to a specific host, or those that are only partially sequenced. Additionally, we could not integrate the full genome sequence of host species, as those data are lacking for the majority of species included. Similarly, we could not include life-history or other ecological traits of our hosts, as those data are not available for many species, and are not comparable between animals and plants. Thus, we had to rely on diversion times as the only proxy to differentiate between our hosts.
- Our method does not make assumptions, or use features, based on which specific parts of the virus genome, or which receptor or receptor binding proteins are commonly utilised in specific transmission routes. Instead, we synthesised a wide range of features (n=446) from three complementary perspectives. This 'no-preconceptions' approach enables us to analyse transmission routes/modes of viruses to their known hosts without being restricted by our current, and highly incomplete, knowledge of the specific biological and molecular mechanisms which govern mechanism of transmission. Whilst some of these details are known for a very limited number of well-studied viruses and hosts, they are unknown for the vast majority. Therefore, a machine learning study aiming for breadth of understanding across all transmission routes, viruses, and hosts cannot use these incomplete data. Despite this 'no-preconceptions' approach having this distinct advantage, it is also a limitation of the predictions, and may result in less accurate predictions for those well studied transmission routes/viruses/hosts for which important factors are well known.

Conclusion

This study is the first of its kind to demonstrate that the virus sequence morphology and host information, increasingly available in the first few days of an outbreak, can be used to accurately identify the transmission routes of a novel virus, across the animal and plant viromes. Importantly, we

have showcased that predictions can be achieved with high accuracy, including for respiratory and vector-borne routes/modes, which together encompass the majority of high-consequence outbreaks across animals and plants. Together with the more matured field of viral host-range prediction, much of the key information which is needed to assess the potential for a virus to cause a high-consequence outbreak can be predicted in the first few days, enabling rapid and targeted mitigation procedures and triage of the time-consuming confirmatory investigative.

METHODS

Viruses: Viruses with complete sequences were downloaded from Genbank[29]. Sequences labelled with the terms: 'vector', 'construct', 'vaccine', or 'clone' were excluded. Following this, the number of ambiguous bases for each sequence was identified, and sequences requiring more than 1,024 permutations to resolve were removed from further analyses. Segmented viruses were included only if all their segments met the previous criteria per strain (of virus). Finally, viruses without known animal or plant host species (below) were also excluded. The above protocol resulted in 6,803 virus species or strains that were included in this study. Supplementary Table 1 illustrates the distribution of these viruses.

Virus-host associations: Associations between the above viruses and their observed host species were compiled from the ENHanCEd Infectious Diseases Database[30], virus-host associations datasets[31–34], and literature[e.g.35], and were further supplemented with targeted literature searches . 28,661 associations between 5,750 host species (animals = 3,649, and plants = 2,101), and the 6,803 fully sequenced viruses. Supplementary Table 2 lists the distribution of these associations by virus Baltimore classification and host category. Supplementary Dataset 1 lists all included associations and their sources.

Hierarchy of transmission modes and routes. We identified 81 routes of virus transmission in animals and plants by searching relevant literature (Supplementary Table 3 provides details of all

included routes/modes). We adopted a three-fold strategy to capture whether our viruses are transmissible by one or more of these routes as follows:

1. **Title and abstract (TIABs) mining**: we extracted TIABs of PubMed papers linked to single virus species (i.e. excluding TIABs with multiple viruses). These TIABs were subsequently classified via keyword matching into the transmission routes described above. The resulting routes were then manually verified and erroneous associations removed.
2. **Manual extractions**: we manually extracted transmission routes of viruses for which no papers were identified by the previous step, as well as for routes not detected in the TIABs from various sources[35–38]. Supplementary Dataset 2 lists all sources.
3. **Within genus generalisations**: Finally, in certain cases, we assigned transmission routes to viruses not captured by the previous two steps, by taking a minimal agreement set of within-genus transmission routes. Supplementary Dataset 2 all sources.

Following a manual check for accuracy, and to remove erroneous routes, we were able to identify at least one transmission *route* (of total = 77 routes, Supplementary Table 3) for 4,446 viruses (65.35% of total – Supplementary Table 4) to 5,317 hosts (total of 24,953 virus-host associations).

Features: We engineered 446 features, in three perspectives, for each included virus-host association, and each modelled route/mode as follows (For full details of number of features, description and relevance to transmission mechanisms see Supplementary Notes 3-5):

1. Viral features (n=442): we categorised features engineered from the virus genome into the following groups: i) genomic features (length (1 feature), GC content (1 feature), and genomic structure (7 features); ii) Morphology (capsid) features (16 features); iii) replication (2 features); iv) Biases: nucleotide (4 features) and dinucleotide (128 features) biases calculated as genome-wide features; and Relative Synonymous Codon Usage[39] (RSCU, 204 features)) and Amino acids biases (57 features) computed using predicted Open reading frames (ORFs, at three ORF lengths ≥36, ≥300, ≥450); v) ORF features including ORF composition (12 features), genome coverage (6 features) and proportion of total ORF length with overlaps (21 features), also calculated at three

ORF lengths. Supplementary Note 3 lists full details of those features, Supplementary Table 5 summarises rationale for inclusion.

2. Host similarity (n=1): We obtained a time tree of 4,342 plant and animal species from the Time Tree of Life[40] (timetree.org). We used those data to compute diversion time distance between 99.98% of all included host species pairs (Supplementary Note 4). We utilised the resulting distances to calculate similarity between the host species of the focal virus-host association, and all hosts (excluding the focal hosts), for which at least one virus is known to be transmitted by the focal route/mode (Supplementary Note 4 lists full details).
3. Virus-host integrated neighbourhoods (n=3): We expanded the concept of virus phylogenetic neighbourhood[21] to incorporate characteristics of the hosts, and the network linking viruses and hosts, into a unified framework (Supplementary Note 5). Utilising the neighbourhood network of a focal virus-host association, given a focal transmission route/mode, we computed three predictive features:
    a. MN4D measures the average similarity between the focal association, and associations between viruses (excluding the focal virus) known to be transmitted via the given route/mode to hosts other than the focal host.
    b. MN3H indicates whether the focal host is susceptible, via the given route/mode to viruses which exhibit high sequence similarity to the focal virus.
    c. MN4C measures the average similarity between the focal association and associations between viruses (excluding the focal virus) known to be transmitted via the given route/mode to the focal host and at least one other host species.

Class imbalance: The number of observed virus-host instances varied per route/modes (Figure 1, Supplementary Figure 2) - ranging from 0.16% (vertical trans-egg transmission in invertebrates) to 98.71% (horizontal transmission). This presented varied, considerable, and bi-directional imbalance between observed (positive class), and unknown (negative class) transmission routes/modes for our virus-host associations. Due to the small proportion of observed instances per majority of

routes/modes (median = 763 associations, 3.06%), we elected to correct for class imbalance using a range of over-sampling and hybrid methods, rather than strict under-sampling (Supplementary Table 6).

We compared the performance of our selected class-balancing techniques across all modelled transmission routes/modes (n=98), over a single iteration of our pipeline (below, steps 1-5). Following an in-depth assessment against the corresponding held-out test-set (stratified, 10%), using a comprehensive set of performance metrics (Supplementary Table 8), the following techniques were selected (Supplementary Figure 3): SL-SMOTE (25%), MWMOTE (25%), SMOTE (NRAS), SMOTE-ENN (25%), and SMOTE-TL (25%).

Model training, optimisation, and validation: We trained a suite of 250 LightGBM (Lightweight Gradient Boosting Machines)[41] models per every route/mode in our study (n=98). LightGBM utilises boosting and gradient descent to build decision trees (weak learners) sequentially, based on the error obtained from previous iterations. Predictions are generated based on the sum of all trained trees. Unlike other implementations (e.g., XGBoost) which grow trees in level-wise manner, LightGBM implements a leaf-wise approach to tree growth, which enhances training and prediction speed.

For each of our routes/modes we applied the following steps (Supplementary Figure 4) in training our lightGBM models:

<u>Step 1 – Initialisation:</u> as some viruses are transmitted via different routes/modes between individuals of the same species, or between species, we initialised a set of virus-host associations for each of our routes/modes to contain all associations whereby the virus is known to be transmitted to the host species via the given route/mode (positive class), and the remainder of our 24,953 virus-host associations with at least one observed transmission route/mode after excluding all positive-class associations (negative class). We used the resulting set to perform 50 iterations of optimisation, training, and validation of LightGBM boosted decision trees models (steps 2-6).

Step 2 – Splitting: per each iteration (n=50), we split the above set into *training* (80%), *optimisation* (10%) and *validation* (held-out test, 10%) sets. Sets were drawn using stratified random sampling to maintain class ratio in the resulting splits.

Step 3 – Generating association derived features: we recomputed association-derived feature (host similarity and virus-host integrated neighbourhoods) for each set as follows: 1) Training: only associations within the training set (~80% of total) were used. 2) Optimisation: associations within both training and optimisation sets were used (~90%). 3) Validation: all available associations were used.

Step 4 – Class balancing: Each training set was then passed through our balancing pipeline, which generated five training sets: SL-SMOTE (25%), MWMOTE (25%), SMOTE (NRAS), SMOTE-ENN (25%), and SMOTE- TL (25%) (Supplementary Note 6). Integer and binary variables were adjusted, post balancing, by rounding.

Step 5 – Optimisation: A LightGBM model was then optimised for each of the resulting training sets (n=5) against the corresponding optimisation set (10%, not balanced, same set was used for all class-balancing models trained per each iteration). We performed Bayesian (model-based) optimisation, using two metrics: AUC and PRAUC, against the validation set to tune the hyperparameters (Supplementary Table 7). We set the tuning stopping to stagnation over 15 rounds, and optimisation was performed using the R packages: mlr3tuning and mlrintermbo.

Step 6 – Bagging of class-balancing techniques: We generated predictions by averaging probability outputs of the five constituent models (i.e. bagging). This approach allowed us to factor in uncertainties arising from our class-balancing processes.

Step 7 – Validation: Performance of resulting bagged ensembles and their constituent models was assessed against the stratified held-out test sets (10% of virus-host interactions per iteration, not balanced), using a comprehensive set of performance metrics (Supplementary Table 8).

Steps 2-7 above were performed using the R Packages: mlr3, mlr3extralearners, mlr3tuning, mlrintermbo, and lightgbm.

Construction of final ensembles. We excluded underperforming models by averaging predictions from our top-10 ensembles. Top-10 were selected by ranking the class-balancing ensembles (n=50), trained for each route/mode (n=98), using the average of four metrics: AUC, PR-AUC, Precision, and 1-Brier score. All four metrics were computed against the corresponding held-out test sets. A value closer to 1 (against held-out validation set) of the first three metrics indicates a better performance. The Area Under the ROC Curve (AUC) is widely used to evaluate the discrimination ability of binary classifiers. A higher AUC indicates better overall performance. For imbalanced datasets, Precision-Recall AUC (PR-AUC) is preferred, as it focuses on the positive class and is more sensitive to class imbalances. Precision (Positive Predictive Value) assesses the accuracy of positive predictions. Conversely, Brier scores closer to zero indicate more reliable predictions. Brier score measures the accuracy of probabilities generated by the models, thus it enables assessment of the confidence of a model. A more confident model would generate probabilities closer to 1 for instances of the positive class, and closer to 0 for instances of the negative class. By employing a combination of these metrics, we are able to comprehensively assess the performance of our ensembles, taking into account discrimination, class imbalances, positive prediction accuracy, and calibration aspects for more informed decision-making.

Feature importance: We quantified the contribution each of our features (n=446) made on predictions using SHAP (SHapley Additive exPlanations) values[42]. SHAP values were computed for each constituent model (n=50), and then averaged for each instance/feature combination for each of our top-10 ensembles. Instance-level SHAP values, for a given route/mode, represent the contribution each of our features made to the specific prediction made by the corresponding top-10 ensemble for that specific instance. Instance-level SHAP values can be either positive or negative: For a specific feature, a positive SHAP value indicates that its presence increases the prediction (probability) compared to the average prediction across all possible instances. In other words, positive SHAP values represent how much the specific feature (at the specific value indicated by the instance) contributes towards a positive prediction (virus is transmitted to host species via a specific

route/mode). Conversely, a negative SHAP value for a feature indicates that this specific feature (at the specific value indicated by the instance) decreases the resulting probability compared to the average prediction across all possible instances. The absolute magnitude of the SHAP value for a feature indicates the importance or influence of that feature on the model's prediction for a specific instance. Larger absolute SHAP values mean that the feature has a more significant impact on the prediction.

Since SHAP values are relative and represent the impact of a feature on a prediction, they are scale-free and do not have units. However, to facilitate the comparison of the contribution each of our features made to the predictions of each modelled route/mode (n=98), as well as to measure overall importance of our features, we normalised SHAP values at two levels as follows:

Local (locally normalised SHAPs): Provides relative importance within each route/mode and allows us to compare the relative importance of features within that specific route/mode. This is turn enable us to understand the contribution each of our feature made to the prediction of each modelled route/mode. Additionally, local normalisation facilitates model-specific interpretation, such as feature importance within the context of each individual route/mode, thus enhancing interpretability of our predictions.

Global (Globally normalised SHAPs): Provides relative importance across all models, which allows us to compare the relative importance of features across different routes/modes, which in turn enables us to identify consistent patterns of feature importance across multiple routes/modes. Additionally, normalising globally simplifies the comparison process by aligning the scales of SHAP values across all routes/modes. Thus, allowing for straightforward comparisons without the need to consider individual model characteristics.

Prediction dependencies: To maximise the utilisation of our models in mitigating against future emerging viruses, it was essential to measure their ability to differentiate between closely related routes/modes (e.g. mosquito-borne vs midge- or sandfly-borne). To this end, we evaluated the extent by which our predictions for a given route/mode depended on predictions of its sibling

route(s)/modes(s) - children of the same parent node in our hierarchy (Figure 1). This was achieved by utilising Mutual Information (MI) to determine if there exists a dependency between the mean probabilities (top-10 ensembles) of instances predicted to be transmitted by the focal route/mode (mean probability >0.5), and the corresponding mean probabilities obtained from the siblings' pipelines. Where the focal route has more than one sibling (e.g. mosquito-borne has two siblings: midge-borne and sandfly-borne), we used the maximum mean probability for each instance in our calculations.

MI quantifies how much knowing the value of one variable can tell us about the value of the other variable. If the resulting estimate is high, it indicates a strong relationship or dependency between the variables, meaning that knowing one variable provides useful information about the other. Notably, MI can capture both linear and nonlinear relationships, and is suitable for analysing data with highly skewed distributions, as is the case with our input data.

To assess the statistical significance of the MI estimates, we compared each estimate, obtained using mi.empirical function in the R package entropy, to a null distribution using bootstrapping (n=2,000). The null hypothesis assumes there is no dependency between the predictions of focal route and those of its siblings. The resulting p-value, calculated using empPvals function in the R package qvalue, indicates the probability of obtaining an MI value as extreme or more extreme than the observed value, assuming the null hypothesis is true.

We normalised MI estimates by dividing each by the maximum possible MI given the underlying sample size. Normalised MI ranges between [0,1]. Normalised MI = 0 indicate there is no information shared between the focal route/mode and its siblings, and therefore there is no relationship between the two. Normalised MI $\geq$ 0.7 suggests a strong correlation between the focal route/mode and its siblings, and changes in one are likely to be reflected in the other.

DATA AND CODE

All data and codes accompanying the manuscript can be found at: 10.6084/m9.figshare.24581421.

REFERENCES

690 1. Wardeh, M., Blagrove, M. S. C., Sharkey, K. J. & Baylis, M. Divide-and-conquer: machine-
691 learning integrates mammalian and viral traits with network features to predict virus-
692 mammal associations. *Nat. Commun. 2021 121* **12**, 1–15 (2021).

693 2. Wardeh, M., Baylis, M. & Blagrove, M. S. C. Predicting mammalian hosts in which novel
694 coronaviruses can be generated. *Nat. Commun. 2021 121* **12**, 1–12 (2021).

695 3. Becker, D. J. *et al.* Optimising predictive models to prioritise viral discovery in zoonotic
696 reservoirs. *The Lancet Microbe* **0**, (2022).

697 4. Gibb, R. *et al.* Mammal virus diversity estimates are unstable due to accelerating discovery
698 effort. *Biol. Lett.* **18**, 20210427 (2022).

699 5. Brierley, L. & Fowler, A. Predicting the animal hosts of coronaviruses from compositional
700 biases of spike protein and whole genome sequences through machine learning. *PLoS Pathog.*
701 **17**, (2021).

702 6. Greenhalgh, T. *et al.* Ten scientific reasons in support of airborne transmission of SARS-CoV-2.
703 *Lancet (London, England)* **397**, 1603–1605 (2021).

704 7. Pastorino, B., Touret, F., Gilles, M., de Lamballerie, X. & Charrel, R. N. Prolonged Infectivity of
705 SARS-CoV-2 in Fomites. *Emerg. Infect. Dis.* **26**, (2020).

706 8. Moreira, J., Peixoto, T. M., Siqueira, A. M. & Lamas, C. C. Sexually acquired Zika virus: a
707 systematic review. *Clin. Microbiol. Infect.* **23**, 296–305 (2017).

708 9. Thorson, A., Formenty, P., Lofthouse, C. & Broutet, N. Systematic review of the literature on
709 viral persistence and sexual transmission from recovered Ebola survivors: Evidence and
710 recommendations. *BMJ Open* **6**, e008859 (2016).

711 10. Cortez, M. H. & Weitz, J. S. Distinguishing between indirect and direct modes of transmission
712 using epidemiological time series. *Am. Nat.* **181**, (2013).

713 11. Plowright, R. K. *et al.* Pathways to zoonotic spillover. *Nat. Rev. Microbiol. 2017 158* **15**, 502–
714 510 (2017).

715 12. Johnston, S. & Holgate, S. Epidemiology of Viral Respiratory Tract Infections. *Viral Other*
716 *Infect. Hum. Respir. Tract* 1–38 (1996) doi:10.1007/978-94-011-7930-0_1.

717 13. Folly, A. J. *et al.* Evidence for overwintering and autochthonous transmission of Usutu virus to
718 wild birds following its redetection in the United Kingdom. *Transbound. Emerg. Dis.* (2022)
719 doi:10.1111/TBED.14738.

720 14. Caminade, C. *et al.* Global risk model for vector-borne transmission of Zika virus reveals the
721 role of El Niño 2015. *Proc. Natl. Acad. Sci. U. S. A.* **114**, 119–124 (2017).

722 15. Alkhamis, M. A. *et al.* Global emergence and evolutionary dynamics of bluetongue virus. *Sci.*
723 *Rep.* **10**, (2020).

724 16. Burrell, C. J., Howard, C. R. & Murphy, F. A. Epidemiology of Viral Infections. *Fenner White's*
725 *Med. Virol.* 185 (2017) doi:10.1016/B978-0-12-375156-0.00013-8.

726 17. Wille, M., Bröjer, C., Lundkvist, Å. & Järhult, J. D. Alternate routes of influenza A virus
727 infection in Mallard (Anas platyrhynchos). *Vet. Res.* **49**, 1–9 (2018).

728 18. Krammer, F. *et al.* Influenza. *Nat. Rev. Dis. Prim. 2018 41* **4**, 1–21 (2018).

729 19. Pierson, T. C. & Diamond, M. S. The continued threat of emerging flaviviruses. *Nat. Microbiol.*
730 **5**, 796–812 (2020).

731 20. Blitvich, B. J. & Firth, A. E. A Review of Flaviviruses that Have No Known Arthropod Vector.
732 *Viruses* **9**, (2017).

733 21. Babayan, S. A., Orton, R. J. & Streicker, D. G. Predicting reservoir hosts and arthropod vectors
734 from evolutionary signatures in RNA virus genomes. *Science (80-.).* **362**, 577–580 (2018).

735 22. Bushman, F. D., McCormick, K. & Sherrill-Mix, S. Virus structures constrain transmission
736 modes. *Nat. Microbiol.* **4**, 1778–1780 (2019).

737 23. Tseng, K. K. *et al.* Viral genomic features predict orthopoxvirus reservoir hosts. *bioRxiv*
738 2023.10.26.564211 (2023) doi:10.1101/2023.10.26.564211.

739 24. Blagrove, M. S. *et al.* Monkeypox virus shows potential to infect a diverse range of native
740 animal species across Europe, indicating high risk of becoming endemic in the region. *bioRxiv*
741 2022.08.13.503846 (2022) doi:10.1101/2022.08.13.503846.

742 25. Heck, M. Insect Transmission of Plant Pathogens: a Systems Biology Perspective. *mSystems* **3**,
743 (2018).

744 26. Wu, V. Y. *et al.* A minimum data standard for vector competence experiments. *Sci. Data 2022*
745 *91* **9**, 1–6 (2022).

746 27. Kraemer, M. U. G. *et al.* The global distribution of the arbovirus vectors Aedes aegypti and Ae.
747 albopictus. *Elife* **4**, (2015).

748 28. Leung, N. H. L. Transmissibility and transmission of respiratory viruses. *Nat. Rev. Microbiol.*
749 *2021 198* **19**, 528–545 (2021).

750 29. Sayers, E. W. *et al.* Database resources of the National Center for Biotechnology Information.
751 *Nucleic Acids Res.* (2018) doi:10.1093/nar/gky1069.

752 30. Wardeh, M., Risley, C., Mcintyre, M. K., Setzkorn, C. & Baylis, M. Database of host-pathogen
753 and related species interactions, and their global distribution. *Sci. Data* **2**, (2015).

754 31. Gibb, R. *et al.* Data proliferation, reconciliation, and synthesis in viral ecology. *bioRxiv*
755 2021.01.14.426572 (2021) doi:10.1101/2021.01.14.426572.

756 32. Olival, K. J. *et al.* Host and viral traits predict zoonotic spillover from mammals. *Nature* **546**,
757 646–650 (2017).

758 33. Shaw, L. P. *et al.* The phylogenetic range of bacterial and viral pathogens of vertebrates. *Mol.*
759 *Ecol.* **29**, 3361–3379 (2020).

760 34. Stephens, P. R. *et al.* Global Mammal Parasite Database version 2.0. *Ecology* **98**, 1476 (2017).

35. Sastry, K. S., Mandal, B., Hammond, J., Scott, S. W. & Briddon, R. W. Encyclopedia of Plant Viruses and Viroids. *Encycl. Plant Viruses Viroids* (2019) doi:10.1007/978-81-322-3912-3.

36. Lefkowitz, E. J. *et al.* Virus taxonomy: The database of the International Committee on Taxonomy of Viruses (ICTV). *Nucleic Acids Res.* **46**, D708–D717 (2018).

37. Hulo, C. *et al.* ViralZone: A knowledge resource to understand virus diversity. *Nucleic Acids Res.* **39**, D576 (2011).

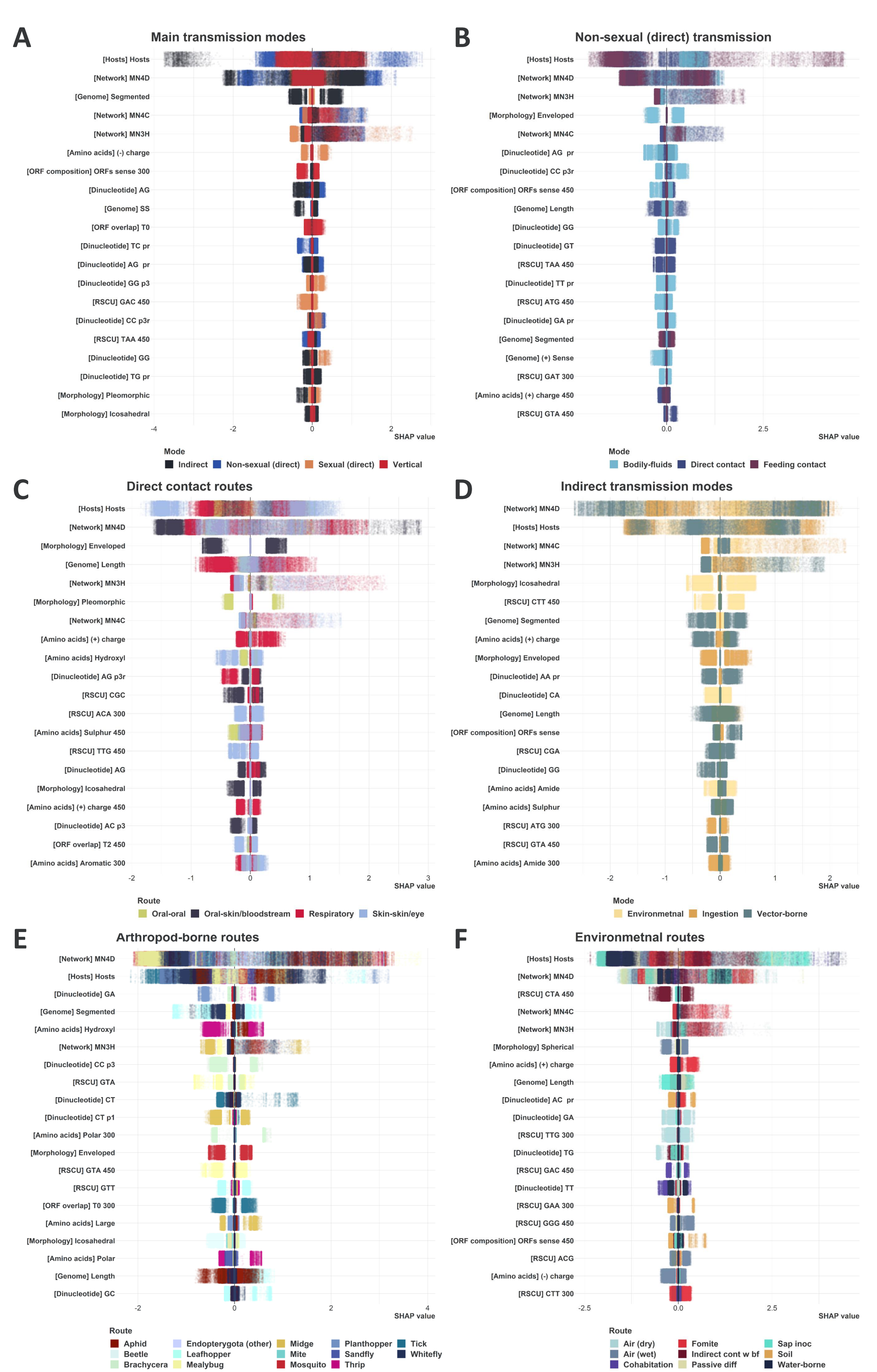
38. Woolhouse, M. E. J. & Brierley, L. Epidemiological characteristics of human-infective RNA viruses. *Sci. Data* **5**, (2018).

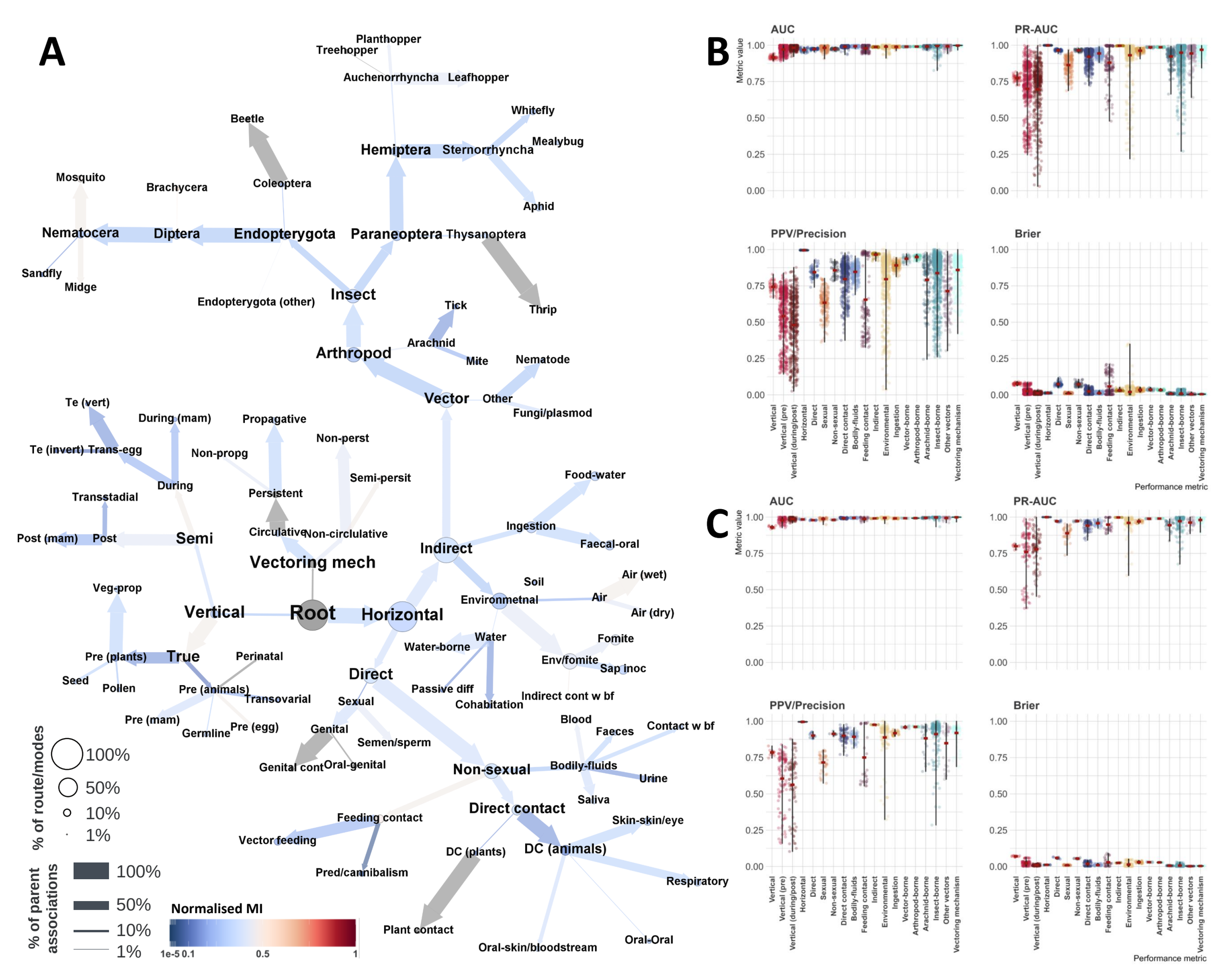
39. Sharp, P. M. & Li, W. H. The codon Adaptation Index--a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Res.* **15**, 1281 (1987).

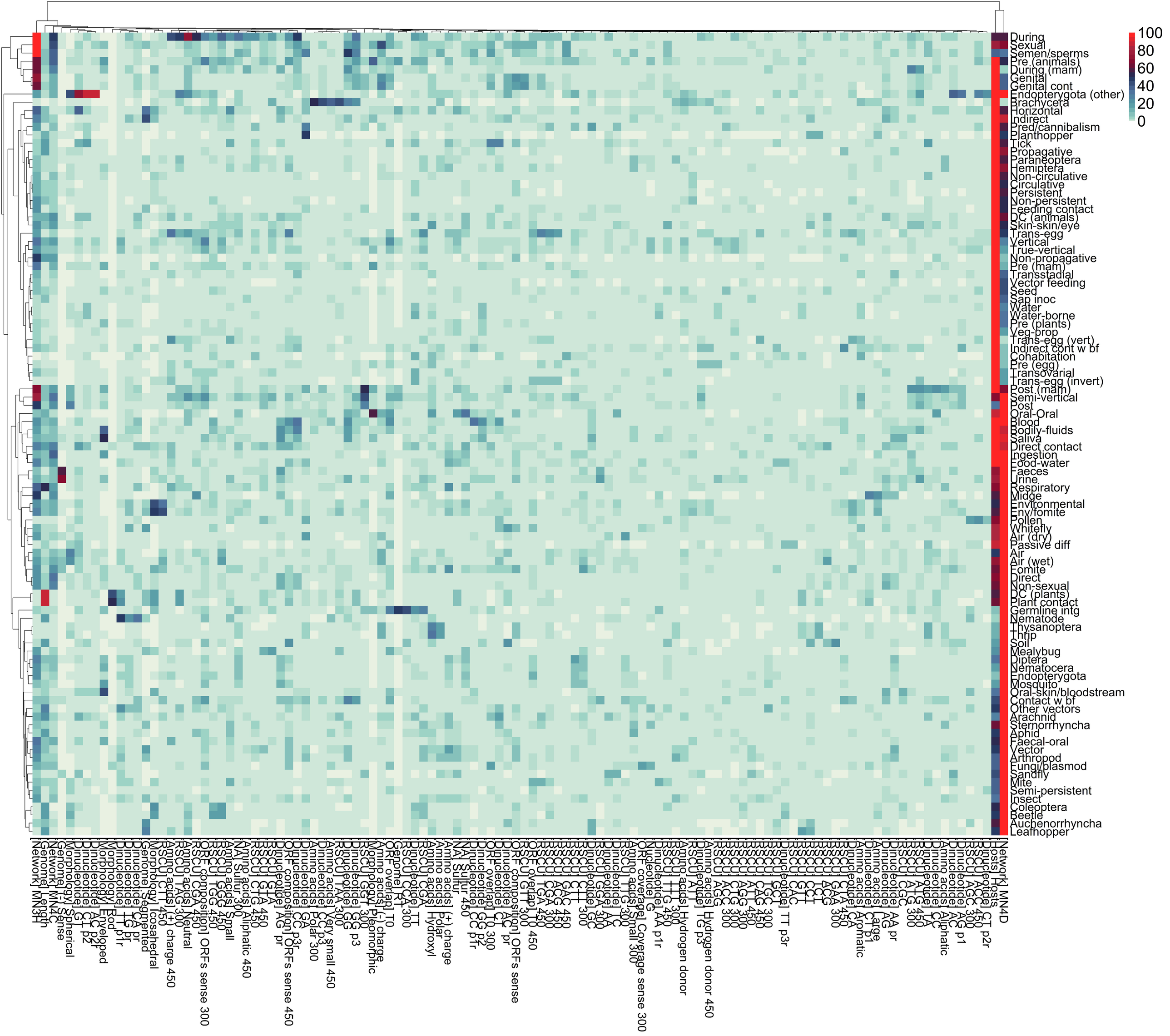
40. Kumar, S. *et al.* TimeTree 5: An Expanded Resource for Species Divergence Times. *Mol. Biol. Evol.* **39**, (2022).

41. Ke, G. *et al.* LightGBM: A Highly Efficient Gradient Boosting Decision Tree.

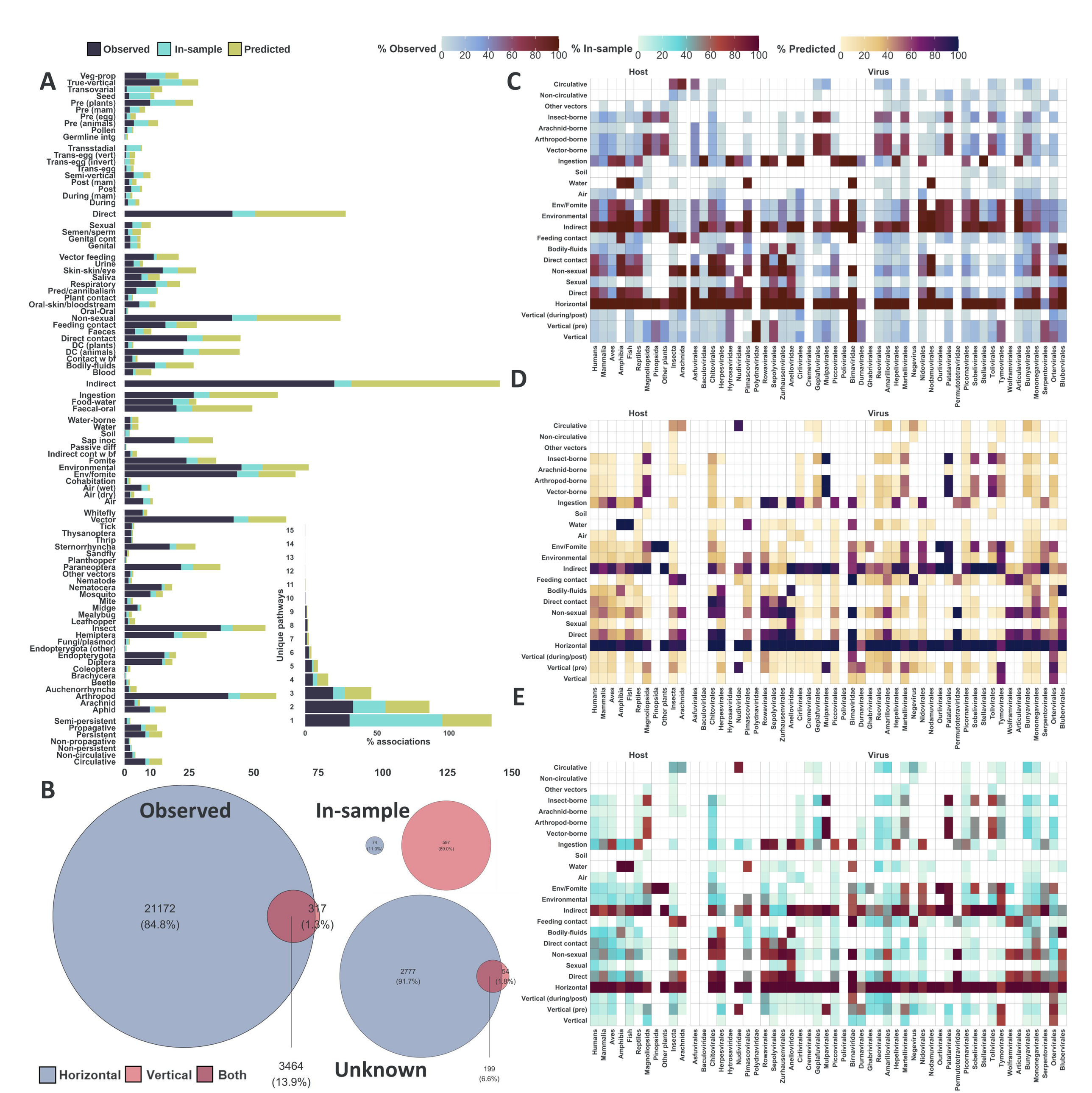
42. Lundberg, S. M. & Lee, S. I. A Unified Approach to Interpreting Model Predictions. *Adv. Neural Inf. Process. Syst.* **2017–December**, 4766–4775 (2017).

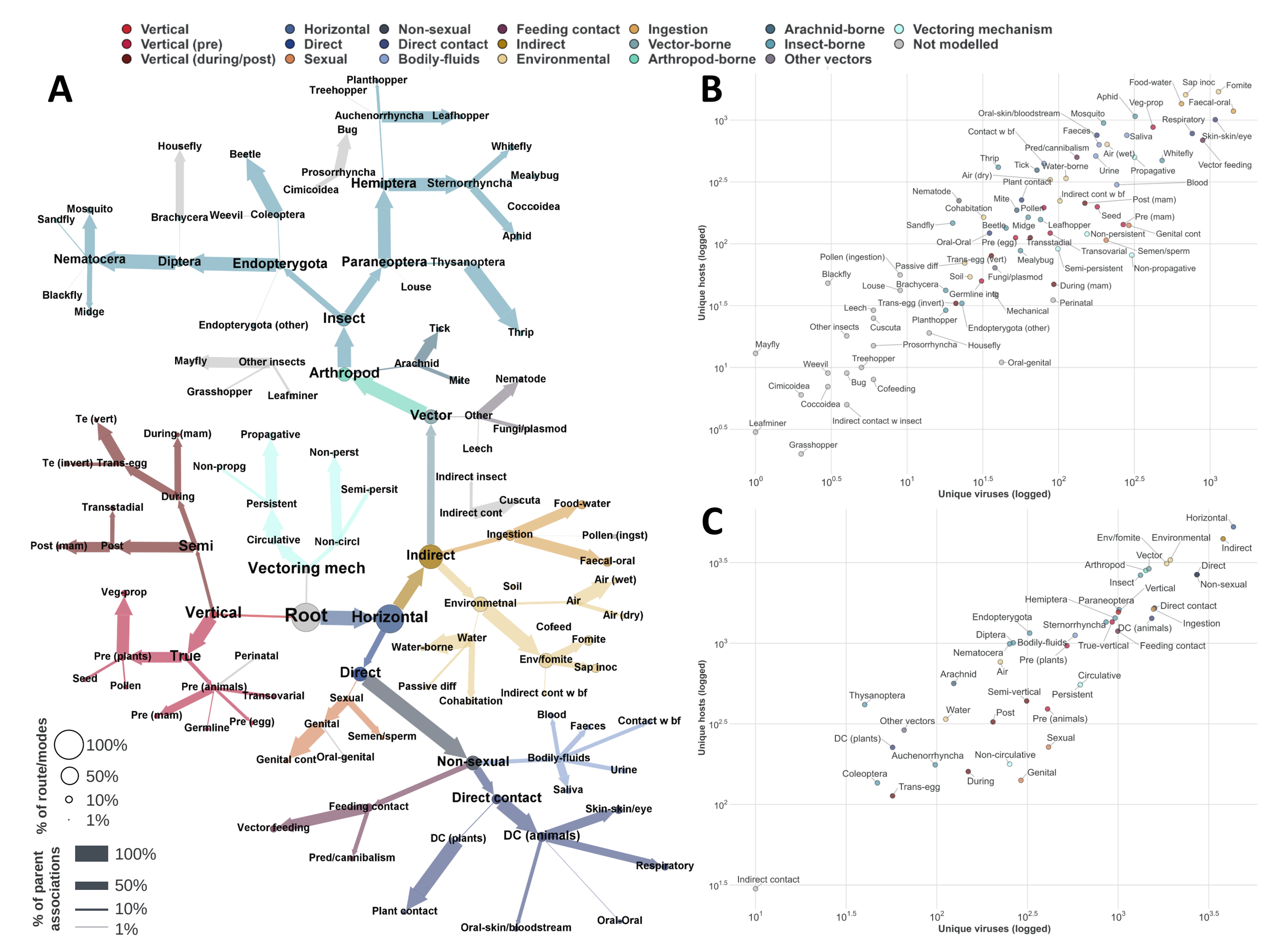

A
Main transmission modes
[Hosts] Hosts
[Network] MN4D
[Genome] Segmented
[Network] MN4C
[Network] MN3H
[Amino acids] (-) charge
[ORF composition] ORFs sense 300
[Dinucleotide] AG
[Genome] SS
[ORF overlap] T0
[Dinucleotide] TC pr
[Dinucleotide] AG pr
[Dinucleotide] GG p3
[RSCU] GAC 450
[Dinucleotide] CC p3r
[RSCU] TAA 450
[Dinucleotide] GG
[Dinucleotide] TG pr
[Morphology] Pleomorphic
[Morphology] Icosahedral
-4
-2
0
2
SHAP value
Mode
Indirect
Non-sexual (direct)
Sexual (direct)
Vertical
B
Non-sexual (direct) transmission
[Hosts] Hosts
[Network] MN4D
[Network] MN3H
[Morphology] Enveloped
[Network] MN4C
[Dinucleotide] AG pr
[Dinucleotide] CC p3r
[ORF composition] ORFs sense 450
[Genome] Length
[Dinucleotide] GG
[Dinucleotide] GT
[RSCU] TAA 450
[Dinucleotide] TT pr
[RSCU] ATG 450
[Dinucleotide] GA pr
[Genome] Segmented
[Genome] (+) Sense
[RSCU] GAT 300
[Amino acids] (+) charge 450
[RSCU] GTA 450
0.0
2.5
SHAP value
Mode
Bodily-fluids
Direct contact
Feeding contact
C
Direct contact routes
[Hosts] Hosts
[Network] MN4D
[Morphology] Enveloped
[Genome] Length
[Network] MN3H
[Morphology] Pleomorphic
[Network] MN4C
[Amino acids] (+) charge
[Amino acids] Hydroxyl
[Dinucleotide] AG p3r
[RSCU] CGC
[RSCU] ACA 300
[Amino acids] Sulphur 450
[RSCU] TTG 450
[Dinucleotide] AG
[Morphology] Icosahedral
[Amino acids] (+) charge 450
[Dinucleotide] AC p3
[ORF overlap] T2 450
[Amino acids] Aromatic 300
-2
-1
0
1
2
3
SHAP value
Route
Oral-oral
Oral-skin/bloodstream
Respiratory
Skin-skin/eye
D
Indirect transmission modes
[Network] MN4D
[Hosts] Hosts
[Network] MN4C
[Network] MN3H
[Morphology] Icosahedral
[RSCU] CTT 450
[Genome] Segmented
[Amino acids] (+) charge
[Morphology] Enveloped
[Dinucleotide] AA pr
[Dinucleotide] CA
[Genome] Length
[ORF composition] ORFs sense
[RSCU] CGA
[Dinucleotide] GG
[Amino acids] Amide
[Amino acids] Sulphur
[RSCU] ATG 300
[RSCU] GTA 450
[Amino acids] Amide 300
-2
-1
0
1
2
SHAP value
Mode
Environmetnal
Ingestion
Vector-borne
E
Arthropod-borne routes
[Network] MN4D
[Hosts] Hosts
[Dinucleotide] GA
[Genome] Segmented
[Amino acids] Hydroxyl
[Network] MN3H
[Dinucleotide] CC p3
[RSCU] GTA
[Dinucleotide] CT
[Dinucleotide] CT p1
[Amino acids] Polar 300
[Morphology] Enveloped
[RSCU] GTA 450
[RSCU] GTT
[ORF overlap] T0 300
[Amino acids] Large
[Morphology] Icosahedral
[Amino acids] Polar
[Genome] Length
[Dinucleotide] GC
-2
0
2
4
SHAP value
Route
Aphid
Beetle
Brachycera
Endopterygota (other)
Leafhopper
Mealybug
Midge
Mite
Mosquito
Planthopper
Sandfly
Thrip
Tick
Whitefly
F
Environmetnal routes
[Hosts] Hosts
[Network] MN4D
[RSCU] CTA 450
[Network] MN4C
[Network] MN3H
[Morphology] Spherical
[Amino acids] (+) charge
[Genome] Length
[Dinucleotide] AC pr
[Dinucleotide] GA
[RSCU] TTG 300
[Dinucleotide] TG
[RSCU] GAC 450
[Dinucleotide] TT
[RSCU] GAA 300
[RSCU] GGG 450
[ORF composition] ORFs sense 450
[RSCU] ACG
[Amino acids] (-) charge
[RSCU] CTT 300
-2.5
0.0
2.5
SHAP value
Route
Air (dry)
Air (wet)
Cohabitation
Fomite
Indirect cont w bf
Passive diff
Sap inoc
Soil
Water-borne


A
Planthopper
Treehopper
Auchenorrhyncha
Leafhopper
Whitefly
Beetle
Mealybug
Hemiptera
Sternorrhyncha
Mosquito
Brachycera
Coleoptera
Aphid
Nematocera
Diptera
Endopterygota
Paraneoptera
Thysanoptera
Sandfly
Midge
Insect
Endopterygota (other)
Tick
Thrip
Arachnid
Arthropod
Mite
Nematode
Vector
Other
Te (vert)
Fungi/plasmod
During (mam)
Propagative
Non-perst
Te (invert)
Trans-egg
Non-propg
Semi-persit
Food-water
During
Transstadial
Persistent
Post (mam)
Post
Semi
Circulative
Non-circlulative
Ingestion
Indirect
Faecal-oral
Vectoring mech
Air (wet)
Soil
Veg-prop
Environmetnal
Air
Vertical
Root
Horizontal
Air (dry)
Water
Fomite
Water-borne
Pre (plants)
True
Perinatal
Env/fomite
Sap inoc
Direct
Seed
Pollen
Pre (animals)
Passive diff
Transovarial
Sexual
Cohabitation
Indirect cont w bf
Pre (mam)
Blood
Germline
Pre (egg)
Genital
Faeces
Contact w bf
Semen/sperm
Genital cont
Oral-genital
Non-sexual
Bodily-fluids
Urine
Saliva
Direct contact
Feeding contact
Skin-skin/eye
Vector feeding
DC (plants)
DC (animals)
Pred/cannibalism
Respiratory
Plant contact
Oral-skin/bloodstream
Oral-Oral
% of route/modes
100%
50%
10%
1%
% of parent associations
100%
50%
10%
1%
Normalised MI
1e-5 0.1
0.5
1
B
AUC
PR-AUC
PPV/Precision
Brier
Metric value
1.00
0.75
0.50
0.25
0.00
Vertical
Vertical (pre)
Vertical (during/post)
Horizontal
Direct
Sexual
Non-sexual
Direct contact
Bodily-fluids
Feeding contact
Indirect
Environmental
Ingestion
Vector-borne
Arthropod-borne
Arachnid-borne
Insect-borne
Other vectors
Vectoring mechanism
Performance metric
C
AUC
PR-AUC
PPV/Precision
Brier
Metric value
1.00
0.75
0.50
0.25
0.00
Vertical
Vertical (pre)
Vertical (during/post)
Horizontal
Direct
Sexual
Non-sexual
Direct contact
Bodily-fluids
Feeding contact
Indirect
Environmental
Ingestion
Vector-borne
Arthropod-borne
Arachnid-borne
Insect-borne
Other vectors
Vectoring mechanism
Performance metric



100
80
60
40
20
0
During
Sexual
Semen/sperms
Pre (animals)
During (mam)
Genital
Genital cont
Endopterygota (other)
Brachycera
Horizontal
Indirect
Pred/cannibalism
Planthopper
Tick
Propagative
Paraneoptera
Hemiptera
Non-circulative
Circulative
Persistent
Non-persistent
Feeding contact
DC (animals)
Skin-skin/eye
Trans-egg
Vertical
True-vertical
Non-propagative
Pre (mam)
Transstadial
Vector feeding
Seed
Sap inoc
Water
Water-borne
Pre (plants)
Veg-prop
Trans-egg (vert)
Indirect cont w bf
Cohabitation
Pre (egg)
Transovarial
Trans-egg (invert)
Post (mam)
Semi-vertical
Post
Oral-Oral
Blood
Bodily-fluids
Saliva
Direct contact
Ingestion
Food-water
Faeces
Urine
Respiratory
Midge
Environmental
Env/fomite
Pollen
Whitefly
Air (dry)
Passive diff
Air
Air (wet)
Fomite
Direct
Non-sexual
DC (plants)
Plant contact
Germline intg
Nematode
Thysanoptera
Thrip
Soil
Mealybug
Diptera
Nematocera
Endopterygota
Mosquito
Oral-skin/bloodstream
Contact w bf
Other vectors
Arachnid
Sternorrhyncha
Aphid
Faecal-oral
Vector
Arthropod
Fungi/plasmod
Sandfly
Mite
Semi-persistent
Insect
Coleoptera
Beetle
Auchenorrhyncha
Leafhopper
Network] MN4D
Hosts] Hosts
Dinucleotide] CT p2r
RSCU] ACC 450
RSCU] AGC 300
Dinucleotide] AC p1
Dinucleotide] AC
Amino acids] Aliphatic
Dinucleotide] CC
Dinucleotide] TA
RSCU] ATG 450
RSCU] ATG 300
Dinucleotide] AA pr
Dinucleotide] AG
Amino acids] Large
Dinucleotide] CT p1
Amino acids] Aromatic
Dinucleotide] CA
RSCU] CTA 450
RSCU] GAA 300
RSCU] ACG
RSCU] CCT
RSCU] CCT
RSCU] CCT
RSCU] CAC
Dinucleotide] TT p3r
RSCU] GCT
RSCU] TGC 300
RSCU] CAG 450
RSCU] ATT 450
RSCU] AGG 300
RSCU] CTG 300
RSCU] ACC 300
Amino acids] Hydrogen donor 450
Dinucleotide] TG p3
RSCU] ATT
Amino acids] Hydrogen donor
RSCU] TTT 300
RSCU] TTG 450
Dinucleotide] AA p1r
Nucleotide] G
ORF coverage] Coverage sense 300
Amino acids] Small 300
RSCU] TTG 300
Dinucleotide] TC
Dinucleotide] AA
RSCU] GGA 300
Dinucleotide] GC
RSCU] CTT 300
RSCU] GAC 450
RSCU] ACG 450
RSCU] CGA 300
RSCU] TGA 450
ORF overlap] T0 450
RSCU] GTT 300
ORF composition] ORFs sense
Dinucleotide] AC pr
Dinucleotide] CT 300
ORF overlap] T0 300
Dinucleotide] GG p2
Dinucleotide] CC p1r
NA] Sulfur 450
NA] Sulfur
Amino acids] (+) charge
Amino acids] Polar
Amino acids] Hydroxyl
RSCU] CGA
Dinucleotide] TT
RSCU] CCA 300
Genome] RT
ORF overlap] T0
Amino acids] (-) charge
Amino acids] Pleomorphic
Morphology] Pleomorphic
RSCU] GGT 300
Dinucleotide] GG p3
Dinucleotide] GG
RSCU] TAA 300
Amino acids] Very small 450
Dinucleotide] CC p3
Amino acids] Polar 300
Dinucleotide] GA
Dinucleotide] CC p3r
ORF composition] ORFs sense 450
Dinucleotide] AG pr
RSCU] GTA
RSCU] GTA 450
RSCU] CTG 450
Amino acids] Aliphatic 450
NA] Sulfur 300
Amino acids] Small
RSCU] GGG 450
RSCU] TGG 450
ORF composition] ORFs sense 300
RSCU] CGA 450
Amino acids] Neutral
RSCU] TAG 300
Amino acids] (+) charge 450
RSCU] CTT 450
Morphology] Icosahedral
Genome] Segmented
Dinucleotide] CA pr
Dinucleotide] TG pr
Dinucleotide] TT p1r
Morphology] Rod
Morphology] Enveloped
Dinucleotide] AC p2r
Dinucleotide] CT p2
Morphology] Spherical
Genome] (-) Sense
Network] MN4C
Genome] Length
Network] MN3H


Observed
In-sample
Predicted
% Observed
% In-sample
% Predicted
A
Veg-prop
True-vertical
Transovarial
Seed
Pre (plants)
Pre (mam)
Pre (egg)
Pre (animals)
Pollen
Germline intg
Transstadial
Trans-egg (vert)
Trans-egg (invert)
Trans-egg
Semi-vertical
Post (mam)
Post
During (mam)
During
Direct
Sexual
Semen/sperm
Genital cont
Genital
Vector feeding
Urine
Skin-skin/eye
Saliva
Respiratory
Pred/cannibalism
Plant contact
Oral-skin/bloodstream
Oral-Oral
Non-sexual
Feeding contact
Faeces
Direct contact
DC (plants)
DC (animals)
Contact w bf
Bodily-fluids
Blood
Indirect
Ingestion
Food-water
Faecal-oral
Water-borne
Water
Soil
Sap inoc
Passive diff
Indirect cont w bf
Fomite
Environmental
Env/fomite
Cohabitation
Air (wet)
Air (dry)
Air
Whitefly
Vector
Tick
Thysanoptera
Thrip
Sternorrhyncha
Sandfly
Planthopper
Paraneoptera
Other vectors
Nematode
Nematocera
Mosquito
Mite
Midge
Mealybug
Leafhopper
Insect
Hemiptera
Fungi/plasmod
Endopterygota (other)
Endopterygota
Diptera
Coleoptera
Brachycera
Beetle
Auchenorrhyncha
Arthropod
Arachnid
Aphid
Semi-persistent
Propagative
Persistent
Non-propagative
Non-persistent
Non-circulative
Circulative
0 10 25 50 75 100 125 150
Unique pathways
% associations
B
Observed
21172
(84.8%)
317
(1.3%)
3464
(13.9%)
In-sample
74
(11.0%)
597
(89.0%)
Unknown
2777
(91.7%)
54
(1.8%)
199
(6.6%)
Horizontal
Vertical
Both
C
D
E
Host
Virus
Circulative
Non-circulative
Other vectors
Insect-borne
Arachnid-borne
Arthropod-borne
Vector-borne
Ingestion
Soil
Water
Air
Env/Fomite
Environmental
Indirect
Feeding contact
Bodily-fluids
Direct contact
Non-sexual
Sexual
Direct
Horizontal
Vertical (during/post)
Vertical (pre)
Vertical
Humans
Mammalia
Aves
Amphibia
Fish
Reptiles
Magnoliopsida
Pinopsida
Other plants
Insecta
Arachnida
Asfuvirales
Baculoviridae
Chitovirales
Herpesvirales
Hytrosaviridae
Nudiviridae
Pimascovirales
Polydnaviridae
Rowavirales
Sepolyvirales
Zurhausenvirales
Anelloviridae
Cirlivirales
Cremevirales
Geplafuvirales
Mulpavirales
Piccovirales
Polivirales
Birnaviridae
Durnavirales
Ghabrivirales
Reovirales
Amarillovirales
Hepelivirales
Martellivirales
Negevirus
Nidovirales
Nodamuvirales
Ourlivirales
Patatavirales
Permutotetraviridae
Picornavirales
Sobelivirales
Stellavirales
Tolivirales
Tymovirales
Wolframvirales
Articulavirales
Bunyavirales
Mononegavirales
Serpentovirales
Ortervirales
Blubervirales


Vertical
Vertical (pre)
Vertical (during/post)
Horizontal
Direct
Sexual
Non-sexual
Direct contact
Bodily-fluids
Feeding contact
Indirect
Environmental
Ingestion
Vector-borne
Arthropod-borne
Arachnid-borne
Insect-borne
Other vectors
Vectoring mechanism
Not modelled
A
Planthopper
Treehopper
Auchenorrhyncha
Leafhopper
Bug
Housefly
Beetle
Whitefly
Prosorrhyncha
Mealybug
Cimicoidea
Hemiptera
Sternorrhyncha
Mosquito
Sandfly
Brachycera
Weevil
Coleoptera
Coccoidea
Aphid
Nematocera
Diptera
Endopterygota
Paraneoptera
Thysanoptera
Blackfly
Midge
Louse
Endopterygota (other)
Insect
Tick
Thrip
Mayfly
Other insects
Arachnid
Arthropod
Mite
Nematode
Grasshopper
Leafminer
Vector
Other
Fungi/plasmod
Te (vert)
During (mam)
Propagative
Non-perst
Leech
Te (invert) Trans-egg
Non-propg
Indirect insect
Semi-persit
During
Persistent
Cuscuta
Food-water
Transstadial
Indirect cont
Post (mam)
Post
Semi
Circulative
Non-circl
Ingestion
Pollen (ingst)
Indirect
Faecal-oral
Vectoring mech
Air (wet)
Soil
Veg-prop
Environmetnal
Air
Vertical
Root
Horizontal
Air (dry)
Cofeed
Water
Fomite
Water-borne
Pre (plants)
True
Perinatal
Env/fomite
Sap inoc
Seed
Pollen
Pre (animals)
Transovarial
Direct
Passive diff
Sexual
Indirect cont w bf
Cohabitation
Pre (mam)
Germline
Pre (egg)
Genital
Blood
Faeces
Contact w bf
Semen/sperm
Genital cont
Oral-genital
Non-sexual
Bodily-fluids
Urine
Saliva
Direct contact
Feeding contact
Skin-skin/eye
Vector feeding
DC (plants)
DC (animals)
Pred/cannibalism
Respiratory
Plant contact
Oral-skin/bloodstream
Oral-Oral
100%
50%
10%
1%
% of route/modes
% of parent associations
B
Food-water
Sap inoc
Fomite
Aphid
Veg-prop
Faecal-oral
Oral-skin/bloodstream
Mosquito
Respiratory
Contact w bf
Faeces
Saliva
Skin-skin/eye
Pred/cannibalism
Air (wet)
Whitefly
Thrip
Tick
Water-borne
Urine
Propagative
Vector feeding
Air (dry)
Plant contact
Blood
Nematode
Indirect cont w bf
Mite
Post (mam)
Cohabitation
Pollen
Seed
Pre (mam)
Sandfly
Beetle
Midge
Leafhopper
Non-persistent
Genital cont
Oral-Oral
Pre (egg)
Transstadial
Transovarial
Semen/sperm
Pollen (ingestion)
Trans-egg (vert)
Mealybug
Passive diff
Semi-persistent
Non-propagative
Blackfly
Soil
Fungi/plasmod
Louse
Brachycera
During (mam)
Germline intg
Trans-egg (invert)
Perinatal
Leech
Mechanical
Planthopper
Other insects
Cuscuta
Endopterygota (other)
Mayfly
Prosorrhyncha
Housefly
Treehopper
Oral-genital
Weevil
Bug
Cimicoidea
Cofeeding
Coccoidea
Indirect contact w insect
Leafminer
Grasshopper
Unique hosts (logged)
Unique viruses (logged)
C
Horizontal
Env/fomite
Environmental
Indirect
Vector
Arthropod
Direct
Insect
Non-sexual
Vertical
Hemiptera
Paraneoptera
Direct contact
Endopterygota
Sternorrhyncha
DC (animals)
Ingestion
Diptera
Bodily-fluids
True-vertical
Feeding contact
Nematocera
Pre (plants)
Arachnid
Air
Circulative
Semi-vertical
Persistent
Thysanoptera
Water
Post
Pre (animals)
Other vectors
Sexual
DC (plants)
Auchenorrhyncha
Non-circulative
Coleoptera
During
Genital
Trans-egg
Indirect contact
Unique hosts (logged)
Unique viruses (logged)