

SPATIAL, TAXONOMIC, AND PREDICTIVE PATTERNS OF BAT-ASSOCIATED VIRAL FAMILIES ACROSS EURASIA FROM A DBATVIR/ZOVER2.0-DERIVED DATASET

Y. Hodzhev¹, M. Zhelyazkova²,
B. Tsafarova^{3,4}, V. Tolchkov^{3,4},
M. Petrova¹, A. Generalova^{3,4},
P. Stoev⁵, N. Simov⁵, S. Panaiotov^{3,4}

¹ New Bulgarian University, Sofia, Bulgaria

² University of Mining and Geology “St. Ivan Rilski”, Sofia, Bulgaria

³ National Centre of Infectious and Parasitic Diseases, Sofia, Bulgaria

⁴ Center of Competence “Immunopathogen”, Y. Sakazov 26, 1504 Sofia, Bulgaria

⁵ National Museum of Natural History, Bulgarian Academy of Sciences, Sofia, Bulgaria

ABSTRACT

Background: Bats and the viruses they host are central to understanding zoonotic disease emergence. Eurasia forms a vast, interconnected habitat for many bat species, and extensive migrations suggest that these populations function as linked transmission networks rather than isolated groups.

Material and methods: Public databases such as DBatVir/ZOVER2.0 enable large-scale comparative analyses across bat hosts and viruses. In this study, we examined the Eurasian subset to describe geographic and taxonomic patterns in bat-associated viruses and compare machine-learning approaches (decision tree, linear support vector machine - SVM, random forest) with zero-inflated count models.

ADDRESS FOR CORRESPONDENCE:

Yordan Hodzhev

Department of Natural Sciences,

New Bulgarian University

email: yhodzhev@nbu.bg

Stefan Panaiotov

Department of Microbiology,

National Center of Infectious and Parasitic Diseases

email: spanaiotov@yahoo.com

Results: The dataset included 9,442 cleaned records, with 7,839 from Asia and 1,603 from Europe. These records represented 28 viral families and 153 distinct bat species after taxonomic harmonization. Data distribution was highly uneven, with the ten most frequently reported species accounting for 47.8% of all records. Frequently recorded hosts included *Scotophilus kuhlii*, *Rhinolophus sinicus*, and *Miniopterus schreibersii*. *Coronaviridae* was the most commonly reported viral family (5,526 records), followed by *Astroviridae* and *Rhabdoviridae*. Asia showed a strong dominance of coronavirus records, whereas Europe exhibited a more diverse viral profile with relatively more rhabdoviruses. Machine-learning models achieved moderate accuracy (0.73-0.76), but macro-F1 scores remained low (0.37-0.42), indicating weaker performance for rare viral families. A zero-inflated Poisson model estimated significantly lower expected counts in Europe compared to Asia, yielding an incidence rate ratio of 0.245 (95% CI 0.232-0.259). **Conclusions:** Overall, the dataset is suitable for identifying broad geographic trends but less reliable for detailed analysis of rare viruses. These differences likely reflect ecological variation, host susceptibility, exposure risk, and uneven surveillance effort across regions and time.

Keywords: bats, Eurasia, *Coronaviridae*, zero-inflated Poisson, machine learning

BACKGROUND

Bats are recognized as important hosts for a wide range of viruses, including species with documented or suspected spillover potential (1,2). Their long evolutionary history, social behavior, mobility, and use of roosting networks make them especially informative for studies of wildlife viromes and zoonotic risk. DBatVir (<https://www.mgc.ac.cn/DBatVir/>) was developed as a sequence-centered database for bat-associated viruses, and ZOVER (<https://www.mgc.ac.cn/cgi-bin/ZOVER/main.cgi>) subsequently integrated DBatVir with other zoonotic and vector-borne virus resources to support broader comparative analyses (3). The updated ZOVER2.0 platform extends this virome-centered approach and illustrates the growing demand for large-scale, database-driven surveillance analytics (4,5). Such resources are valuable for descriptive synthesis and hypothesis generation, but discovery bias and incomplete metadata can strongly influence apparent host and geographic patterns.

Bat-associated viruses in Eurasia have attracted growing interest because of their diversity and zoonotic relevance (6,7). The present study examines geographic distribution, host taxonomy, and viral-family structure in a curated Eurasian dataset. The specific objectives were to describe the distribution of detection records from the ZOVER 2.0 database by bat species, bat family, viral family, and country, to explore host-viral profile structure with clustering and principal component analysis, to evaluate simple multiclass classifiers, and to estimate adjusted count contrasts with a zero-inflated Poisson model.

MATERIALS AND METHODS

Data source and curation

Data were downloaded in CSV format from ZOVER 2.0. DBatVir-derived source metadata, where bat-related information such as diet type and geographic distribution were integrated into the database. Missing Bat family and Viral family were treated as hard exclusions. Among the 3,488 Asia-associated records not retained, 1,259 had missing core metadata, mainly Bat diet type (1,063), Bat family (960), and Viral family (224), with overlap between categories. The remaining 2,229 records from Asia-associated records were removed during harmonization of the final Eurasian analytical table. These non-retained records also included a subset of bat taxa characteristic of tropical and subtropical Asia, especially frugivorous species, which were not prioritized for the final comparative Eurasian subset when they did not satisfy the study design criterion of Europe–Asia continuity. After restricting the cleaned file to the intended Europe and Asia continent groups, 9,442 records remained (7,839 from Asia and 1,603 from Europe). No record-level deduplication was attempted because repeated metadata rows may represent distinct sequences or detections in the source database export (8). Analyses treated each row as a database detection record. Counts, therefore, reflect reporting frequency in the supplied export rather than standardized viral prevalence. Species-level figures were limited to binomial host labels to avoid mixing named species with genus-only or family-level host annotations (9). The cleaned dataset included 28 viral families, 174 host labels, and 51 countries. Because this is a secondary analysis of database records, the manuscript structure follows the logic of STROBE reporting for observational studies, while the predictive com-

ponents were described in the spirit of TRIPOD+AI transparency recommendations (8,9). The viral detections were reported from different biological materials collected from bats, such as guano, swabs, blood, and tissues, depending on the original study design. Sample type was not analysed separately because this information was not uniformly standardized across the exported records (3,4).

Statistics and exploratory machine learning

Descriptive summaries were generated for host species, bat family, viral family, country, and continent. Unsupervised analysis used species-by-continent viral-family profiles (9). Profiles with fewer than 20 total records were excluded to reduce extreme sparsity. The resulting count matrix was transformed with $\log(1 + \text{count})$, standardized, and examined with principal component analysis. K-means clustering was evaluated for k values from 2 to 6, with the best silhouette score observed at $k = 4$. The first two principal components were used for visualization in Figure 4. Supervised analysis focused on the seven most frequent viral families. Categorical predictors included bat family, continent, sample type, and bat diet. Records missing one or more of these predictors were excluded from the classification step. The dataset was split into 75% training and 25% test partitions with stratification by viral family. One-hot encoded categorical features were used to train a decision tree, a linear support vector machine, and a random forest. Performance was summarized by accuracy, balanced accuracy, and macro-F1.

Zero-inflated Poisson model

A natural way to model count data with an excessive number of zero counts in the data is to put a point mass p at 0. That is, with probability p , we sample a degenerate distribution at 0 and with probability $(1-p)$ we sample, say, a Poisson distribution. Such models are called zero-inflated Poisson (ZIP) models (12). To model sparse count data with many zeros, a zero-inflated Poisson regression was fitted to a complete bat-family by viral-family by continent count table that explicitly included zero cells (12, 13,14). The response variable was the number of records in each cell. The count component included continent, bat family, and viral family as predictors. The inflation component included an intercept only. Reference levels were Asia for the continent, *Vespertilionidae* for the bat family, and *Coronaviridae* for the viral family. Exponentiated coefficients were reported as

incidence-rate ratios with 95% confidence intervals. As a simple sensitivity check, standard Poisson and negative binomial generalized linear models were also fitted to the same aggregated table.

Software and reporting

Analyses were carried out in Python using pandas for data handling, scikit-learn for clustering and classification, statsmodels for count regression, matplotlib and geopandas for visualization, and python-docx for manuscript generation.

RESULTS

Dataset structure and concentration of records

The cleaned Eurasia subset contained 9,442 records, of which 7,839 (83.0%) were from Asia and 1,603 (17.0%) were from Europe (Table 1). China alone contributed 5,437 records, or 57.6% of the complete dataset. The ten most represented bat species accounted for 47.8% of all records. The most frequently represented species were *Scotophilus kuhlii*, *Rhinolophus sinicus*, *Miniopterus schreibersii*, *Rhinolophus affinis*, and *Rhinolophus ferrumequinum* (Table 2). Notably, four bat species showed a common population for Europe and Asia *Miniopterus schreibersii*,

Rhinolophus ferrumequinum, *Eptesicus serotinus*, *Myotis daubentonii*, other species were exclusively in Asia (Fig. 1).

Table 1. Overview of the cleaned Eurasia subset.

Characteristic	Value
Raw rows	9445
Rows retained after continent harmonization	9442
Records from Asia	7839
Records from Europe	1603
Viral families represented	28
Host labels represented	174
Species-level host labels	153
Countries represented	51

Table 2. Ten most represented bat species in the cleaned workbook.

Species	Records	Share of all records (%)
<i>Scotophilus kuhlii</i>	936	9.9
<i>Rhinolophus sinicus</i>	688	7.3
<i>Miniopterus schreibersii</i>	521	5.5
<i>Rhinolophus affinis</i>	477	5.1
<i>Rhinolophus ferrumequinum</i>	423	4.5
<i>Eptesicus serotinus</i>	408	4.3
<i>Myotis ricketti</i>	343	3.6
<i>Rhinolophus pusillus</i>	270	2.9
<i>Myotis daubentonii</i>	231	2.4
<i>Pipistrellus abramus</i>	220	2.3

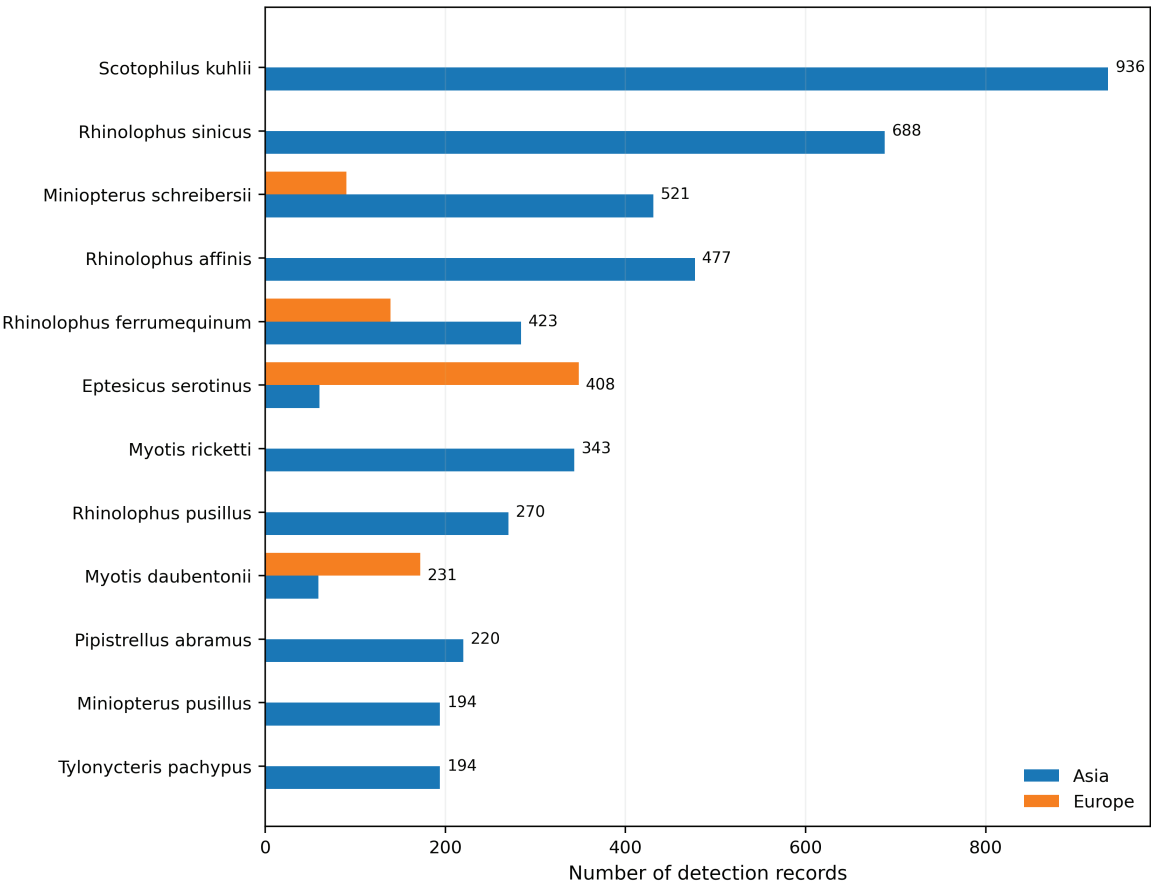


Fig. 1 Number of detection records for the twelve most represented species in the cleaned Eurasia workbook. Counts are shown separately for Asia and Europe, and species are ordered by total record count.

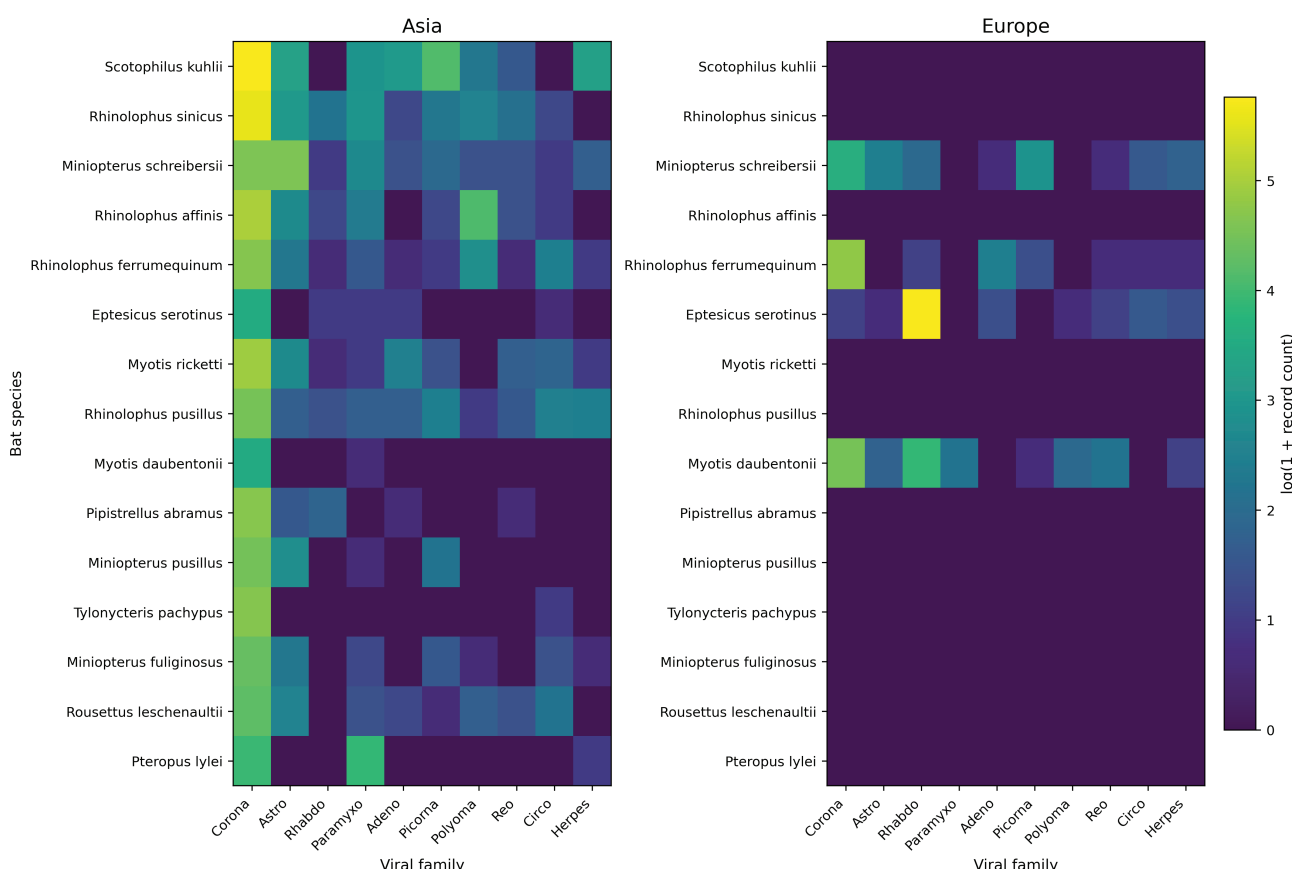


Fig. 2 Log-scaled heatmaps of record counts for the fifteen most frequently reported species and the ten most common viral families in Asia and Europe. The heatmap highlights extensive coronavirus representation in many Asian species and a concentrated *Rhabdoviridae* signal among a smaller set of European species. (Virus family abbreviations: Corona – *Coronaviridae*; Astro – *Astroviridae*; Rhabdo – *Rhabdoviridae*; Paramyxo – *Paramyxoviridae*; Adeno – *Adenoviridae*; Picorna – *Picornaviridae*; Polyoma – *Polyomaviridae*; Reo – *Reoviridae*; Circo – *Circoviridae*; Herpes – *Herpesviridae*).

Viral-family patterns across continents

Coronaviridae dominated the dataset with 5,526 records, followed by *Astroviridae* (927) and *Rhabdoviridae* (514). The continental structure was asymmetric. In Asia, *Coronaviridae* accounted for 4,930 records, or 62.9% of all Asian records. Europe was more heterogeneous. *Coronaviridae* remained the largest category in Europe with 596 records, but *Rhabdoviridae* contributed 438 records and represented 27.3% of all European records. *Adenoviridae* were also comparatively enriched in Europe, with 153 European records versus 141 Asian records, despite the much smaller European sample size. These patterns suggest that the Europe–Asia contrast observed in the supplied workbook cannot be explained solely by lower sampling volume in Europe, but also reflects differences in viral-family composition (Fig. 2 and Fig. 3).

Exploratory machine learning

For unsupervised analysis, 66 species-by-continent profiles with at least 20 records were retained. The first two principal components explained 20.8% and

9.1% of the variance, respectively. Silhouette scores were modest across the tested range, with the best value at $k = 4$ (0.265), so the clustering results should be interpreted as exploratory structure rather than sharply separated ecological groups (Fig. 4). The PCA plot nonetheless showed a consistent cluster of coronavirus-rich Asian profiles that included several *Rhinolophus* species, *Scotophilus kuhlii*, and *Miniopterus schreibersii*. In supervised multiclass classification restricted to the seven most common viral families, all three models achieved similar overall accuracy. The random forest performed best in raw accuracy at 0.755, but balanced accuracy and macro-F1 remained low for all models (Table 3). This indicates that broad host and sample metadata recover dominant families reasonably well, but are insufficient for robust discrimination of minority classes in this strongly imbalanced dataset. The discrepancy between overall accuracy and class-balanced metrics suggests that model performance was driven largely by the most prevalent viral families, especially *Coro-*

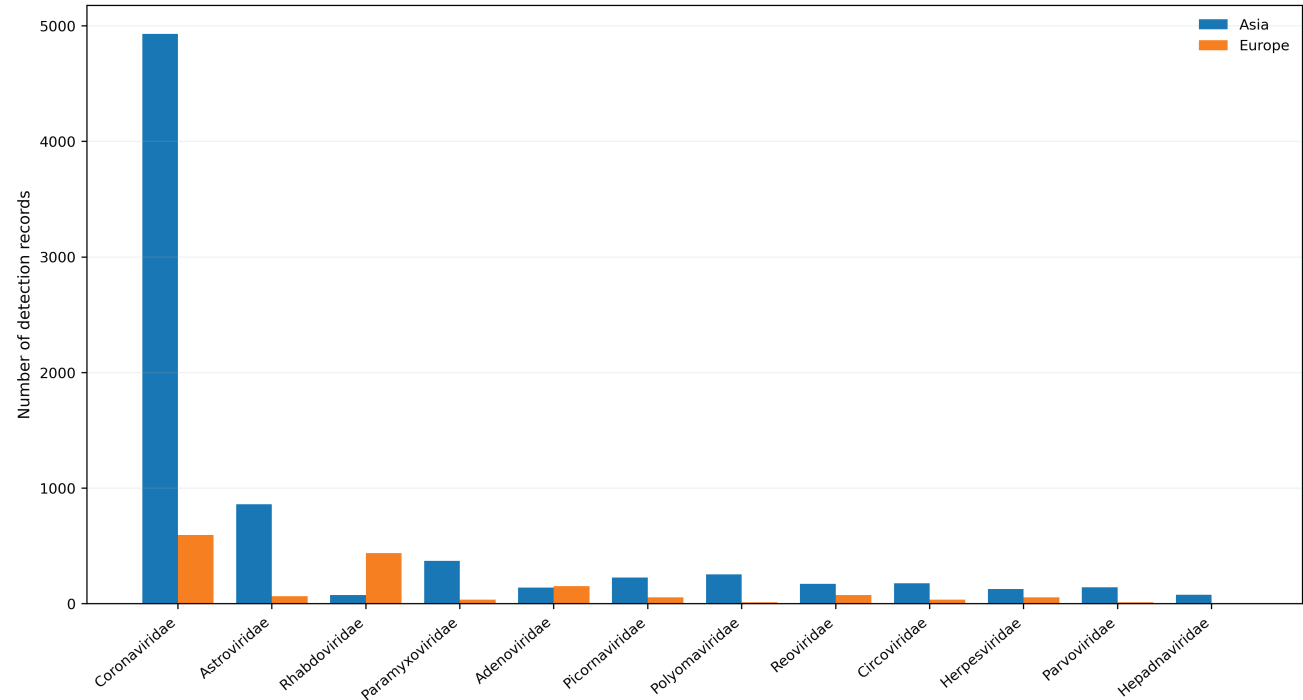


Fig. 3. Comparison of viral-family record counts between Asia and Europe for the twelve most frequent viral families. *Coronaviridae* dominate overall and in Asia, whereas *Rhabdoviridae* and *Adenoviridae* make a proportionally larger contribution in Europe.

naviridae, whereas rarer classes were classified poorly. The fact that this pattern was consistent across all three algorithms indicates that the principal limitation lies in the severe class imbalance of the underlying database rather than in model choice. Accordingly, these predictive results support the recovery of major database-level signals, but not reliable fine-scale discrimination across sparsely represented viral families.

Table 3. Test-set performance of exploratory multi-class classifiers for the seven most frequent viral families.

Model	Accuracy	Balanced accuracy	Macro-F1
Decision tree	0.751	0.369	0.419
Linear SVM	0.734	0.355	0.367
Random forest	0.755	0.366	0.419

Zero-inflated Poisson regression

The zero-inflated Poisson model captured the dominant structure of the aggregated count table and estimated a marked continental contrast. Table 4 showed that after adjustment for bat family and viral family, expected counts in Europe were roughly one quarter of those in Asia (IRR 0.245, 95% CI 0.232–0.259, $p < 0.001$). Relative to the reference family *Vespertilionidae*, adjusted expected counts were lower for *Rhinolophidae* (IRR 0.660, 95% CI 0.631–0.690) and much

lower for sparsely represented families such as *Emballonuridae*, *Molossidae*, and *Rhinopomatidae* (Fig. 5). These coefficients reflect total aggregated counts across all viral families and do not imply absence of important family-specific associations. The accompanying sensitivity comparison suggested residual overdispersion beyond zero inflation, because the simple negative binomial model achieved a lower AIC than either the standard Poisson or the ZIP model. For this reason, the IRRs reported from the ZIP model in Table 4 and Fig. 5 should be viewed as descriptive rather than fully definitive inferential estimates. They are presented because the ZIP framework provides a convenient and interpretable representation of zero inflation in the aggregated dataset, but the better AIC of the negative binomial model indicates that extra-Poisson variation was not fully captured.

Table 4. Selected incidence-rate ratios from the zero-inflated Poisson model.

Comparison	IRR (95% CI)	p
Europe vs Asia	0.245 (0.232–0.259)	<0.001
<i>Emballonuridae</i>	0.037 (0.031–0.044)	<0.001
<i>Molossidae</i>	0.027 (0.021–0.034)	<0.001
<i>Nycteridae</i>	0.000 (0.000–0.002)	<0.001
<i>Pteropodidae</i>	0.257 (0.240–0.275)	<0.001
<i>Rhinolophidae</i>	0.660 (0.631–0.690)	<0.001
<i>Rhinopomatidae</i>	0.007 (0.005–0.010)	<0.001

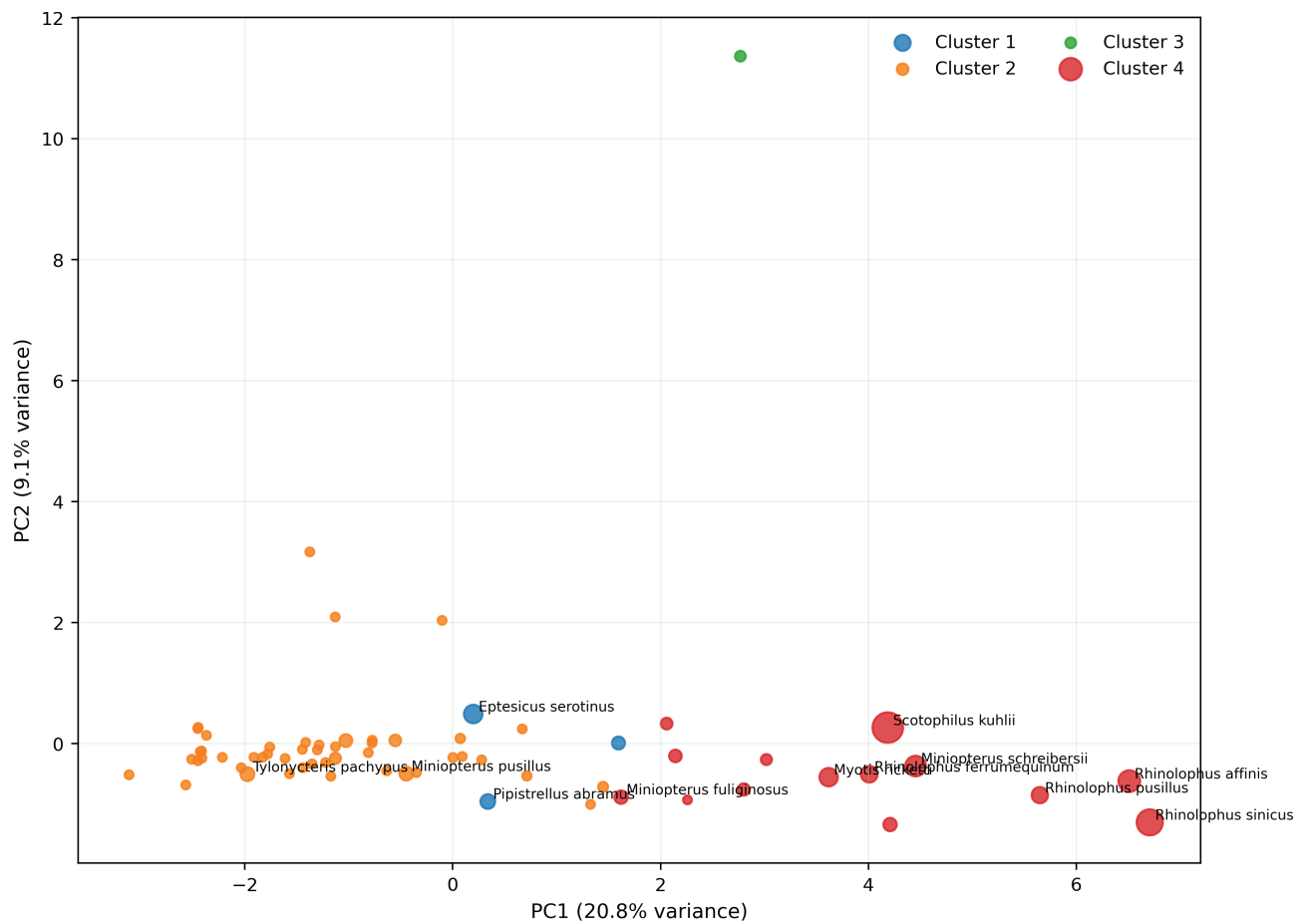


Fig 4. Principal component analysis of species-by-continent viral-family profiles after log transformation and standardization. Point size reflects the total number of records for each profile. Colors indicate k-means clusters with $k = 4$, the best silhouette score among the models tested from 2 to 6 clusters.

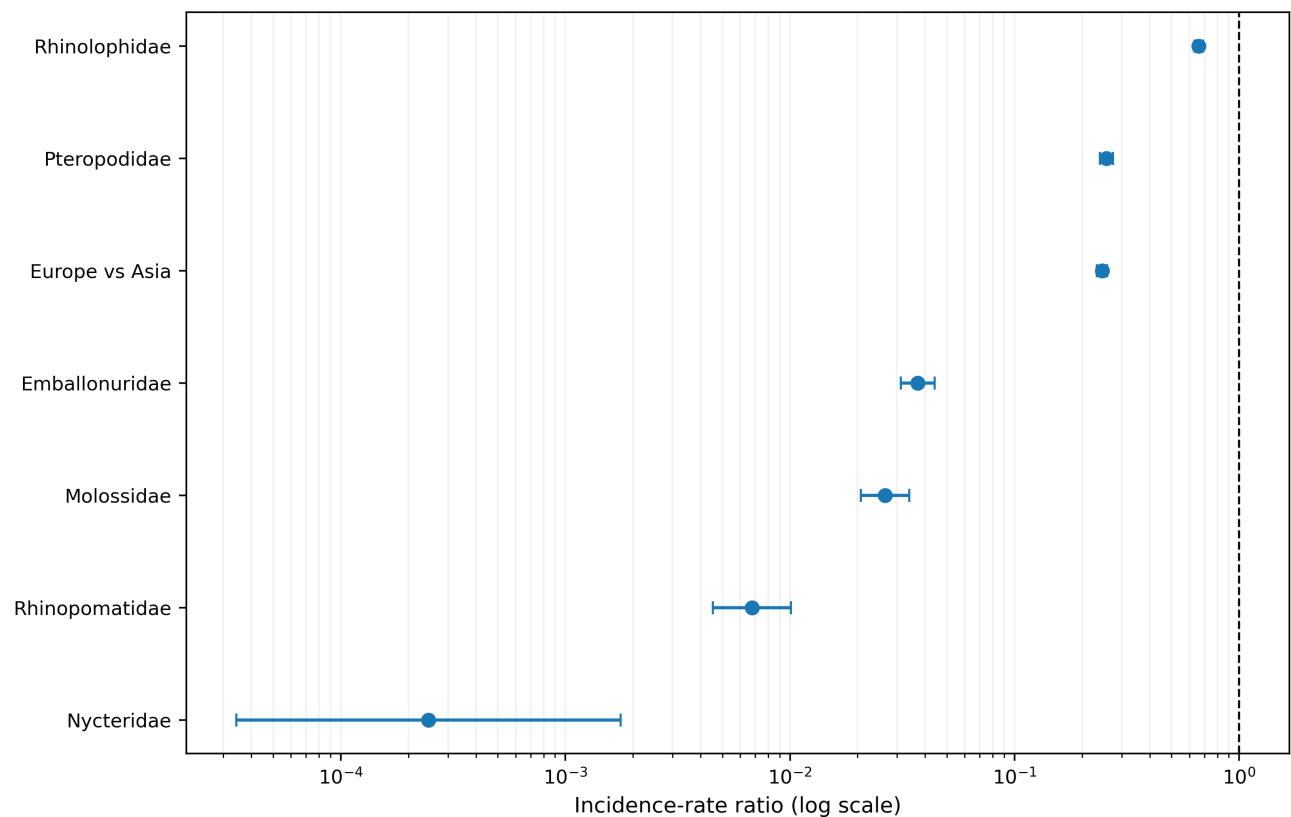


Fig. 5 Incidence-rate ratios from the zero-inflated Poisson model. The vertical dashed line marks an IRR of 1. Values below 1 indicate lower expected record counts relative to the reference level after adjustment for bat family, viral family, and continent.

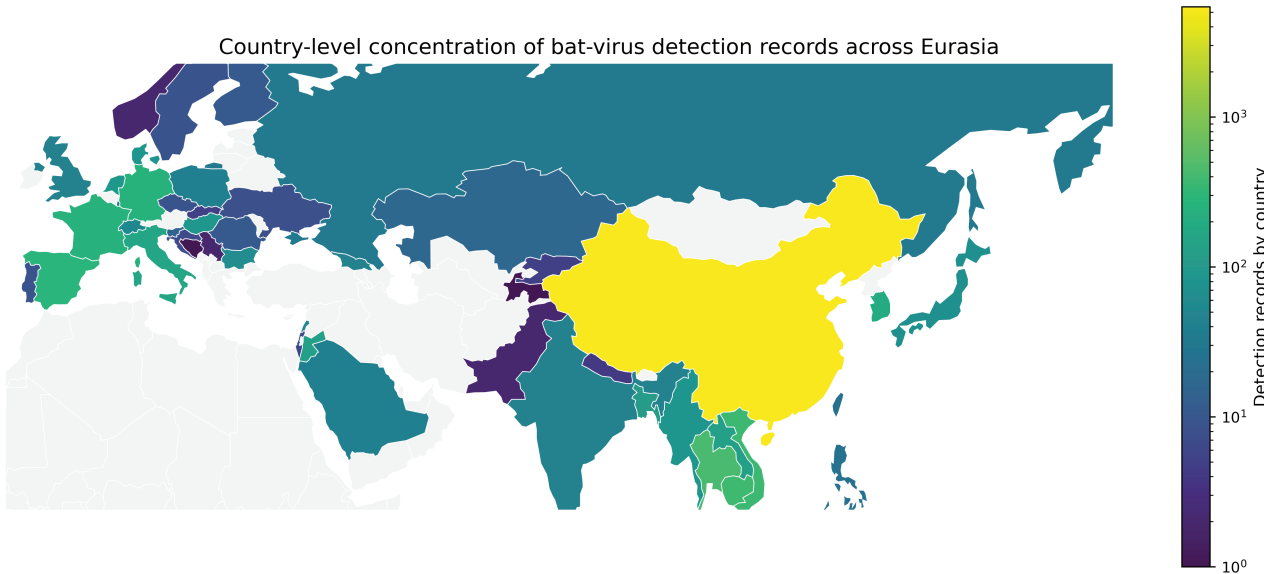


Fig. 6 Country-level concentration of record counts across the Eurasia subset. Colors correspond to the number of cleaned records mapped to each country on a logarithmic scale.

Geographic distribution

The country distribution map (Fig. 6) emphasizes the concentration of available records in East and South-east Asia and, to a lesser degree, in selected European countries such as Spain, Germany, and France.

DISCUSSION

The completed analysis confirms that the explored Eurasia subset was dominated by a small number of host species, countries, and viral families. This concentration is typical of wildlife virus databases, where observed diversity is shaped by a mixture of host biology and unequal discovery effort (10, 11). The prominence of coronavirus records in Asia is consistent with the strong focus on bat coronaviruses after SARS-related outbreaks and with the large number of coronavirus-focused field surveys conducted in China and neighboring countries (15,16). Europe showed a different profile. Although the European subset was much smaller, it contained a disproportionately large *Rhabdoviridae* component and a relatively stronger *Adenoviridae* signal. That pattern may reflect genuine ecological differences, but it is equally plausible that it arises from targeted surveillance programs, differences in sample type, or publication and database curation biases. Accordingly, the present results should be interpreted as structure in reported detections rather than direct measures of virome prevalence (9,10). The machine-learning results reinforce that point. Overall accuracy was driven by the dominant classes, whereas macro-F1

and balanced accuracy remained modest. In other words, broad categorical metadata can distinguish the most common viral families to some extent, but they do not provide enough information for balanced multiclass prediction across the whole dataset. The clustering results are best understood as exploratory summaries of profile similarity, not as definitive ecological partitions (11). The zero-inflated Poisson model provided a compact summary of continent and host-family contrasts in sparse count data, but the sensitivity analysis suggested substantial overdispersion that was not fully captured by zero inflation alone (14). Future work would benefit from explicitly modeling study effort, country-level variation, time, and host phylogeny, and from comparing ZIP, zero-inflated negative binomial, and hierarchical count models more formally. Where possible, database-derived analyses should also be complemented by standardized surveillance designs and by transparent treatment of missing metadata.

LIMITATIONS

Several limitations should be stated explicitly. The analytical file was a provided export rather than a direct API query to the live database, so the manuscript cannot report an external extraction timestamp beyond the naming of the dataset. The workbook almost certainly mixes multiple kinds of records, including sequence-derived and literature-derived detections, and these units are not necessarily directly comparable as epidemiological observations. Some host

labels were genus-only or otherwise non-standard, and a small number of metadata fields required manual harmonization. Country and continent totals partly reflect where bat-virus surveillance has been concentrated, not only where viral diversity is greatest. The machine-learning analyses were strongly affected by class imbalance, so overall accuracy was higher than class-balanced performance. The count models showed residual overdispersion beyond zero inflation, meaning that the ZIP-based IRRs should be interpreted cautiously.

CONCLUSIONS

The uploaded DbatVir/ZOVER2.0-derived Eurasia dataset shows clear geographic and taxonomic structure in reported bat-virus detections. *Coronaviridae* dominate overall and are especially concentrated in Asia, whereas Europe displays a more mixed pattern with a pronounced rhabdovirus component. Records are concentrated in a small number of species and countries, and simple machine-learning models recover only part of the resulting structure. These counts summarize a discovery database, not a standardized prevalence.

Acknowledgements

This work was supported by the Bulgarian National Science Fund under grant KP-06-H51/9 of 16.11.2021.

Conflict of interest statement

The authors declare no conflicts of interest.

Ethics statement

This study used previously compiled database records and did not involve new sampling of animals or direct work with human participants.

Data and code availability

The analysis was performed on the dataset supplied with this project from ZOVER2.0. The completed manuscript and all figure files are provided with the present submission package.

REFERENCES

1. Calisher CH, Childs JE, Field HE, Holmes KV, Schountz T. Bats: important reservoir hosts of emerging viruses. *Clin Microbiol Rev.* 2006; 19(3):531-545. <https://doi.org/10.1128/CMR.00017-06>
2. Wang LF, Anderson DE. Viruses in bats and potential spillover to animals and humans. *Curr Opin Virol.* 2019; 34:79-89. <https://doi.org/10.1016/j.coviro.2018.12.007>
3. Chen L, Liu B, Yang J, Jin Q, Chen L. DBatVir: the database of bat-associated viruses. *Database.* 2014; 2014:bau021. <https://doi.org/10.1093/database/bau021>
4. Zhou S, Liu B, Han Y, Wang Y, Chen L, Wu Z, Yang J. ZOVER: the database of zoonotic and vector-borne viruses. *Nucleic Acids*

- Res.* 2022; 50(D1):D943-D949. <https://doi.org/10.1093/nar/gkab862>
5. Liu B, Zhu L, Zhou S, Li A, Xu P, Han Y, Shu Y, Chen L, Yang J, Wu Z. ZOVER 2.0: a virome-based platform for zoonotic and vector-borne viruses. *Nucleic Acids Res.* 2026; 54(D1):D925-D931. <https://doi.org/10.1093/nar/gkaf1243>
6. Fleming TH. Bat Migration. In: *Encyclopedia of Animal Behavior.* 2019; 605-610. <https://doi.org/10.1016/B978-0-12-809633-8.20764-4>
7. Phelps KL, Hamel L, Alhmoud N, Ali S, Bilgin R, Sidamonidze K, Urushadze L, Karesh W, Olival KJ. Bat Research Networks and Viral Surveillance: Gaps and Opportunities in Western Asia. *Viruses.* 2019; 11(3):240. <https://doi.org/10.3390/v11030240>
8. Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, et al. The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data.* 2016; 3:160018. <https://doi.org/10.1038/sdata.2016.18>
9. von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP. The STROBE statement: guidelines for reporting observational studies. *Lancet.* 2007; 370(9596):1453-1457. [https://doi.org/10.1016/S0140-6736\(07\)61602-X](https://doi.org/10.1016/S0140-6736(07)61602-X)
10. Collins GS, Moons KGM, Dhiman P, Riley RD, Beam AL, Van Calster B, et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models. *BMJ.* 2024; 385:e078378. <https://doi.org/10.1136/bmj-2023-078378>
11. Gibb R, Albery GF, Carlson CJ, Brierley L, Fletcher R, et al. Mammal virus diversity estimates are unstable due to accelerating discovery effort. *Biol Lett.* 2022; 18(1):20210427. <https://doi.org/10.1098/rsbl.2021.0427>
12. Agarwal DK, Gelfand AE, Citron-Pousty S. Zero inflated models with application to spatial count data. *Environ Ecol Stat.* 2002; 9:341-355.
13. Hall DB. Zero-inflated Poisson and binomial regression with random effects: a case study. *Biometrics.* 2000; 56(4):1030-1039. <https://doi.org/10.1111/j.0006-341X.2000.01030.x>
14. Lambert D. Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics.* 1992; 34(1):1-14. <https://doi.org/10.1080/00401706.1992.10485228>
15. Ruiz-Aravena M, McKee C, Gamble A, Lunn T, Morris A, Snedden CE, et al. Ecology, evolution and spillover of coronavirus-es from bats. *Nat Rev Microbiol.* 2022; 20:299-312. <https://doi.org/10.1038/s41579-021-00652-2>
16. Young CCW, Olival KJ. Optimizing viral discovery in bats. *PLoS One.* 2016; 11(2):e0149237. <https://doi.org/10.1371/journal.pone.0149237>