Available online at www.sciencedirect.com

ScienceDirect

Current Opinion in Virology

Predicting zoonotic potential of viruses: where are we?

Nardus Mollentze[1,2] and Daniel G Streicker[1,2]

The prospect of identifying high-risk viruses and designing interventions to pre-empt their emergence into human populations is enticing, but controversial, particularly when used to justify large-scale virus discovery initiatives. We review the current state of these efforts, identifying three broad classes of predictive models that have differences in data inputs that define their potential utility for triaging newly discovered viruses for further investigation. Prospects for model predictions of public health risk to guide preparedness depend not only on computational improvements to algorithms, but also on more efficient data generation in laboratory, field and clinical settings. Beyond public health applications, efforts to predict zoonoses provide unique research value by creating generalisable understanding of the ecological and evolutionary factors that promote viral emergence.

Addresses
[1] School of Biodiversity, One Health and Veterinary Medicine, University of Glasgow, Glasgow G12 8QQ, United Kingdom
[2] MRC-University of Glasgow Centre for Virus Research, G61 1QH, United Kingdom

Corresponding author: Streicker, Daniel G (daniel.streicker@glasgow.ac.uk)

Current Opinion in Virology 2023, **61**:101346

This review comes from a themed issue on **Adaptation of viruses to new hosts**

Edited by **Silke Stertz** and **Xander de Haan**

For complete overview about the section, refer "Adaptation of viruses to new host (2023)"

Available online 27 July 2023

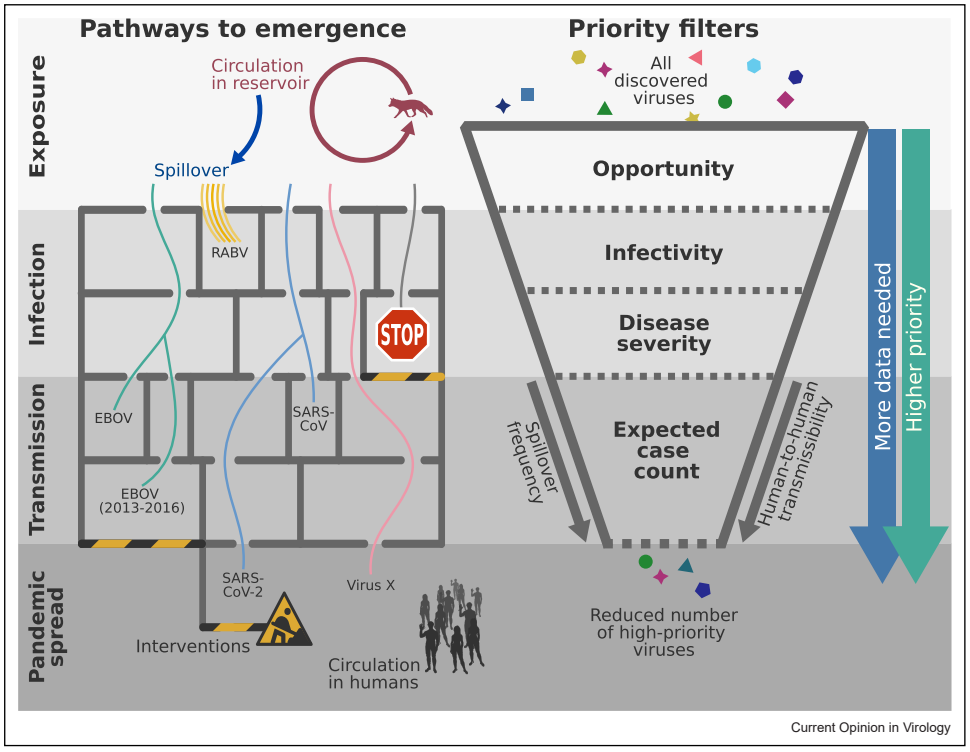
https://doi.org/10.1016/j.coviro.2023.101346

Introduction

Most emerging infectious diseases in humans are caused by viruses that originate from other animal species [1]. As such, genomic surveillance and virus discovery in non-human animals have been proposed to form important components of preparing for future zoonotic threats [2–4]. While virus discovery has unquestionable value in filling gaps in the evolutionary history of viruses and may enhance our ability to pinpoint the animal origins of novel zoonoses, the value of these data for pandemic or indeed spillover prevention remains controversial. In principle, applying carefully designed experiments to newly discovered viruses might provide proxies for human infectivity [5] (particularly when interpreted with expert opinion); indeed, this approach is already used to risk-assess influenza lineages [6]. However, when such experiments are practically feasible, they remain laborious and cannot currently keep pace with the dramatic acceleration of virus discovery empowered by advances in genomics and computational biology [7–9]. The result is an ever-growing backlog of viruses that cannot be comprehensively assessed as part of efforts to prevent emergence. Quantitative triage systems are now being developed to allow systematic, evidence-led prioritisation of newly detected viruses for downstream research and surveillance.

One way viruses might be prioritised is by quantifying their relative risk to public health, which is governed by key viral characteristics (Figure 1). First, there must be opportunities for transmission, which are shaped by both the amount of contact between humans and the virus' natural reservoir (or any potential intermediate hosts) and the nature of that contact ('opportunity'). Second, the virus must be capable of infecting humans, which requires successful interactions with multiple host effectors to enter cells, replicate, spread through the body and evade or suppress immune responses ('infectivity'). Third, high-risk viruses must cause large numbers of cases, either through repeated spillover (e.g. rabies virus) or through successful transmission between humans, which may be modulated by human population connectivity in the geographic location of spillover or cross-reactive immunity from vaccination against or exposure to related viruses ('case count'). Finally, while any virus causing high numbers of infections will likely threaten at least part of the population, the level of public health threat will be determined by the symptoms caused (i.e. 'disease severity'). The continuous nature of risk factors (e.g. the level of exposure to humans, the probability of infection upon exposure, etc.), implies a wide range within which viruses may be stratified (Figure 1). In specific contexts and depending on the aim of the risk-ranking exercise, other risk factors may also be appropriate (e.g. potential economic damage or the lack of effective diagnostic tests, treatments or vaccines) [10,11]. Classically, infectivity, expected case counts and disease severity can only be estimated after human infections have been reported. Anticipating opportunities for zoonotic emergence before human cases is similarly challenging, requiring knowledge of animal reservoirs and transmission routes that typically remain elusive for years after a virus emerges [12]. A growing

Figure 1



Prioritising viruses based on public health risk. Zoonotic viruses of major public health concern overcome a range of barriers to emergence (left), including ecological opportunity for spillover to humans and navigation of a range of within- and between-host barriers to successful infection and onward transmission. The relative ability of different viruses to overcome these barriers, along with the consequences of such infections in humans, could be used to stratify them by the risk posed (right), providing potential opportunities for earlier intervention (traffic signs/striped barriers) than currently possible. Recent advances in predicting human infection suggest it may be possible to develop quantitative filters to supplement the limited data on risk factors that are available for most viruses. Since earlier filters require fewer new data but can remove many potentially low-risk viruses from further consideration, the more laborious downstream characterisation needed to measure other risk factors becomes more focused and feasible. RABV: *Lyssavirus rabies*; EBOV: *Orthoebolavirus zairense*; SARS-CoV: *Severe acute respiratory syndrome-related coronavirus*.

number of computational approaches therefore attempt to predict factors contributing to public health risk from more available data sources.

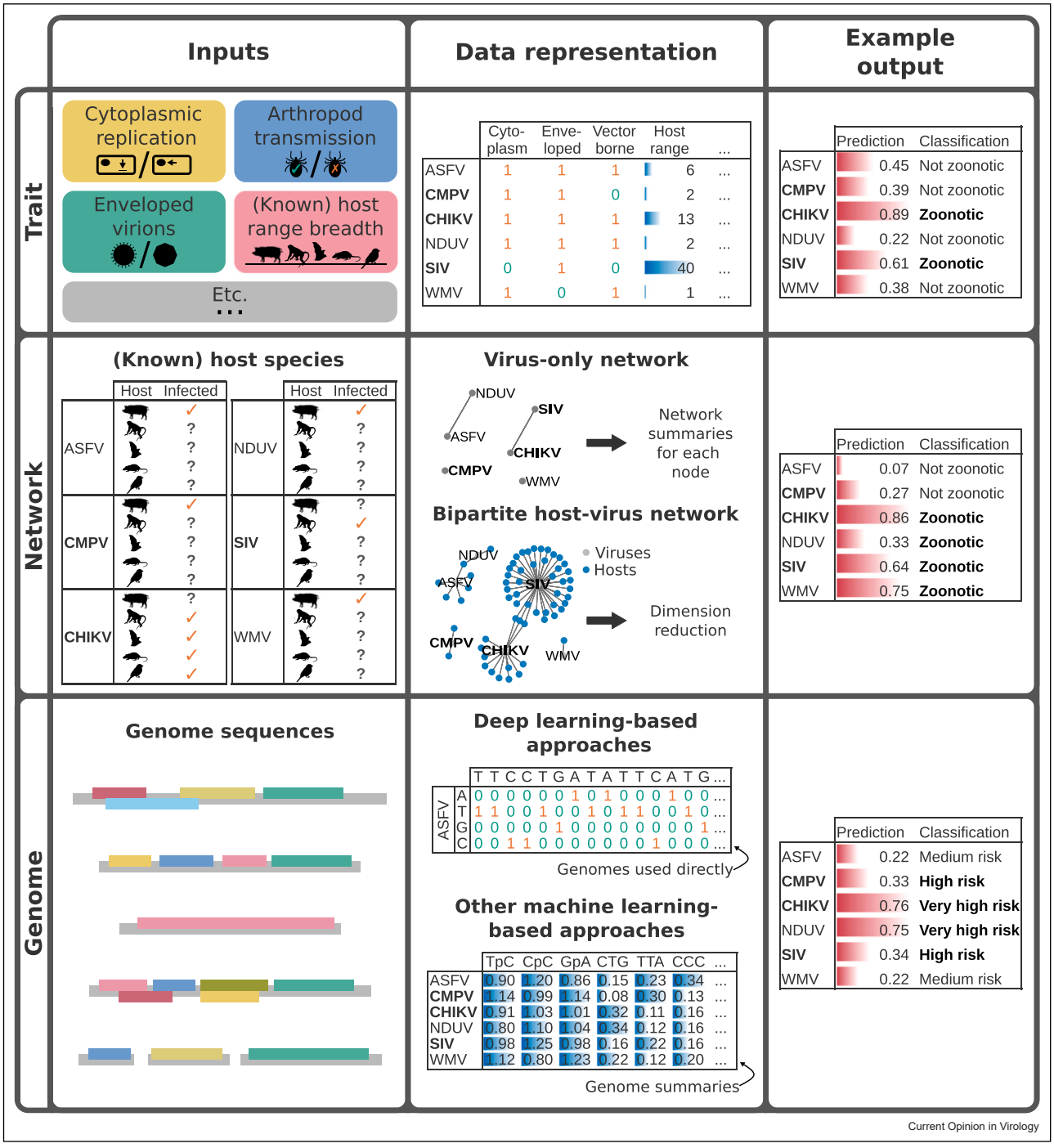
Where we are now

Current approaches to prioritise viruses focus primarily on the ability to infect humans (often termed 'zoonotic potential') but differ in their data inputs and in the range of viruses targeted. Approaches targeting individual virus families (e.g. [13]) have the advantage of being able to rely on more targeted data (e.g. commonly sequenced genomic fragments), but do not provide general insights into zoonotic risk across viruses. Here, we focus on more broadly applicable approaches, which can be classified into three groups based on the types of predictors used (Figure 2).

Trait-based

Trait-based approaches focus on identifying virus, host or ecological traits correlated with reports of human infection. A wide range of such traits has been identified (e.g. the breadth of known host range, cytoplasmic replication, enveloped virions and transmission by arthropod vectors) [14–16]. These and other risk factors have been combined in an expert opinion-based risk assessment, but formal assessments of its predictive power are currently unavailable [17]. On the one hand, trait-based models are desirable since they are easy to interpret and provide testable hypotheses about infection or emergence risk. On the other hand, they suffer from similar issues to the experimental approaches they are meant to supplement. Data for the most informative traits (e.g. host range breadth) are often unavailable and laborious to collect, and when available may reflect ascertainment biases. Whether routinely used corrections based on the number of publications associated with each virus species adequately address such biases remains unclear. One reason to doubt these corrections is that the search effort represented by individual publications is likely to vary systematically according to the

Figure 2



Current approaches to predicting viral zoonotic risk. The three broad classes of models differ in inputs and internal data representation but provide qualitatively similar outputs. Combining input classes may help improve the robustness of models when faced with novel viruses; however, the input data required also determine when and where models can be applied, since some data (e.g. host range) are generally unavailable for recently discovered viruses. Viruses currently known to infect humans are indicated in bold. ASFV: *African swine fever virus*; CMPV: *Camelpox virus*; CHIKV: *Chikungunya virus*; NDUV: *Ndumu virus*; SIV: *Simian immunodeficiency virus*; WMV: *Wad Medani virus*.
Data and example outputs are derived from [14,19,26,46].

traits of the host species involved (e.g. studies involving endangered host species or larger animals will sample fewer individuals, while species that live in large groups that are accessible to researchers may be more heavily sampled). More widely available risk factors (e.g. cytoplasmic replication) tend to be too phylogenetically conserved to make predictions at useful taxonomic scales. As such, trait-based approaches may provide useful biological insights but offer inherently limited prospects to narrow the growing chasm between virus discovery and zoonosis prevention.

Network-based

A second class of 'network-based' approaches aims to predict human infection based on the set of host species a virus is currently known to infect as well as host range similarities between viruses. Host range is expressed either by connecting viruses that share at least one known host (virus-only network) or by connecting viruses and their known hosts (a bipartite host-virus network, Figure 2) [18,19]. Virus-only networks cluster viruses with similar evolutionary history, traits and/or ecology, all of which influence opportunities for host-sharing. Similarly, host-virus networks capture established predictors of human infection, including host range and ecological opportunity (assuming that presence in hosts already known to share other viruses with humans indicates a viable contact route for exposure). This implicit information has been used to predict human infection either by summarising network properties directly, or by creating a low-dimension embedding representing the overall network structure [18,19]. The ability of network-based approaches to infer plausible network links involving hosts other than humans may allow application earlier than possible from observed host range data alone [19]. However, as with trait-based approaches, the effects of biases inherent to the datasets used to train and evaluate these models remain poorly explored. Human-associated viruses tend to be better-studied, which may make them stand out in unexpected ways. For example, human-associated viruses will have more recorded connections to non-human host species, but models implicitly relying on this feature may fail to recognise novel human-infecting viruses that have not yet received the same level of research attention. Consequently, it remains unclear how much host range data will be required for a given virus before network-based approaches can reliably be applied.

Genome-based

Given the data availability issues plaguing other methods, a third set of approaches has focused on predicting ability to infect humans directly from virus genome sequences. Genome sequences are advantageous as they are generally the first (and often only) data available for newly discovered viruses. Viral genomes supply information about virus phylogeny (helpful for prediction since related viruses often have related hosts [20]) as well as more poorly understood signals of host range encoded in genome compositional biases [21–23]. Current approaches rely on artificial intelligence but differ in how they represent sequence data. Deep-learning approaches can automatically extract useful representations from sequences and show impressive performance, detecting sequence reads associated with human-infecting viruses without the need for sequence alignment [24]. However, the data-hungry nature of these models, requiring thousands of virus observations, has required training models using large numbers of closely related genomes (e.g. strains of the same virus species). The resulting pseudoreplication is likely to generate models that overestimate their own predictive performance: models have effectively already 'seen' the human infection status of the new viruses they are challenged to predict. Novel virus species, which can represent large fractions of taxa in modern virus discovery efforts, may be poorly predicted because model training has inadvertently deprioritised identification of generalisable signals of human infection. Further, identifying instances of misleading performance is hampered by the limited interpretability of deep-learning models. Alternative approaches that use less data-demanding machine- learning algorithms can reduce pseudoreplication by building models from smaller datasets (i.e. hundreds, rather than thousands of viruses) at coarser taxonomic resolution (e.g. species), but require human-designed representations of viral genomes (Figure 2). Current models use compositional biases (e.g. codon usage, dinucleotide or amino acid biases, etc.) or the presence or absence of individual substrings (k-mers) and have been used to predict both host range generally [25] and ability to infect humans specifically [26], performing equivalently to or better than trait-based approaches.

Reliance on viral genomes opens the risk that models succeed by recreating viral evolutionary relationships [27], with zoonoses tending to be taxonomically clustered. In fact, genomic models outperform explicit approximations of viral taxonomy and appear to uncover signals that predict human infection across unrelated viral species or even families [26]. While interpretation of these putatively general signals of human infection remains challenging, methods to understand the underpinnings of why a virus is predicted to be zoonotic (or not) are now emerging [28]. On a more practical level, generalisability across viruses enables models to predict the risk posed by novel viruses at the time of their (genomic) discovery [26]. The low cost and data requirements of genome-based models also enable repeat applications to identify biologically important variation within species, including temporal variation from continued evolution.

Opportunities for improvement

Current approaches to predict zoonotic potential are promising but require further improvement to guide

preparedness. Improvements are likely to arise from advances in the design and implementation of computational algorithms, innovations that identify more informative indicators of human infection from widely available data and growth and refinement of datasets. Ongoing computational efforts to combine the trait, network and genomic approaches mentioned above may aid generalisability by reducing the reliance on any single data type [19,29]. New indicators of human infection are also likely to arise. For example, Mollentze et al. [26] improved the performance of a viral genomic predictor of human infection by re-expressing genomic biases relative to those in the human genome. Cutting-edge developments in natural language processing, recently applied in host range prediction, may also play a role in feature engineering [30], and new ways of incorporating predicted protein structures unavailable to models to date may further improve performance [9]. Technological developments in laboratory science also promise improved predictive performance. For example, a major limitation of the datasets used to train and evaluate all models developed to date is our inability to distinguish viruses that are capable of human infection, but have not yet been observed to infect humans, from viruses that are genuinely incapable of human infection. By redefining the status of true zoonoses, new methods in high-throughput, massively multiplexed serology (e.g. Phage Immunoprecipitation Sequencing, PhIP-Seq) may resolve some incorrect labelling of viruses and therefore improve model specificity [31].

Beyond infection, approaches to predict other components of public health risk are urgently needed (Figure 1). Trait-based approaches have shown some success at identifying correlates of exposure risk, virulence and human-to-human transmissibility [32–35], but again rely on broad or generally unavailable variables. It may be more feasible to subdivide these problems, focusing on more readily measured features of virus-host interactions that nevertheless give some information about each virus' capabilities. For example, predicting cell-type or tissue tropism in humans may give indirect evidence of potential differences in transmissibility and virulence, while benefiting from large bodies of experimental data available to train models. Such detailed predictions would also be more readily verifiable in laboratory experiments. Indeed, in parallel to development of predictive models, there has been a growing body of literature seeking to infer zoonotic risk directly from scalable, modular experimental assays [5,36,37]. Integrating predictive models with these approaches would create verifiable methods for converting laboratory results into relative measures of risk, while keeping models grounded in virus biology. Such interpretability is key to both improving trust in predictive models and for anticipating the conditions under which models fail.

Attempts to predict components of public health risk are unlikely to cover all possible contributing factors. For example, the virus-focused nature of predictions means they tend to ignore external influences on risk, such as spatiotemporal variation in the likelihood of spillover. Since the reservoirs, potential intermediate hosts and true spatial distribution of most viruses remains unknown, this will be difficult to address analytically. Further, human population connectivity and cross-protective immunity at the location of emergence may alter outcomes of spillover, but their influence remains difficult to predict for individual viruses. Nonetheless, predictions of the inherent ability of different viruses to infect and transmit among humans allow viruses to be compared relative to each other. This can provide valuable information that could be considered alongside projections from virus species-agnostic epidemiological models and expert opinion-based risk assessments focused on specific regions.

Controversy

Although the prospect of combining increasingly low-cost sequencing, publicly available data and computational models to identify and pre-empt zoonotic risks is enticing, the viability of this approach for generating actionable recommendations remains contentious. A fundamental challenge is the vast number and diversity of viruses in nature. Simple calculations based on current rates of virus discovery point to as many as 1.7 million species infecting mammals and birds [2]. Models accounting for repeat discoveries in multiple host species still suggest at least 40 878 species infecting mammals alone [38]. Only a fraction of these viruses is predicted to be capable of infecting humans (e.g. N = 9787, 23.9% of mammalian viruses [38]) and a smaller, unknown proportion would cause sufficient public health risk to merit intervention. This diversity creates a distinct challenge for actionable zoonosis prediction: even reasonably accurate models will produce a considerable number of false positives. The challenge of viral diversity is compounded by continued viral evolution in reservoir hosts, which may necessitate periodic monitoring to update models and predictions. This criticism in principle could be mitigated by greater capacity for high-throughput laboratory screening to flag false positives, making the development of such systems a priority.

Zoonosis prediction is also stymied by data availability. In addition to being incomplete, current knowledge of viral diversity is biased towards animal taxa that have been linked to human disease and towards the evolutionary relatives of established zoonoses [39]. The extent to which different model types will be influenced by these biases is likely to vary according to model dependence on viral host range and network structure. Greater emphasis on untargeted, metagenomic virus

discovery in taxa without clear links to human health will gradually reduce the magnitude of these potential biases [8,9,40]. However, filling gaps in the viral evolutionary tree increases uncertainty in the taxonomic resolution at which viruses should be modelled (e.g. species, lineage, or isolate) and in which traits of viruses (e.g. host range) can be assumed to be constant across related taxonomic units. Scrutiny of model- training inputs and outputs will be vital to understand the basis of predictions and their credibility, particularly as declining computational and technical barriers make sophisticated algorithms widely accessible.

Given the diminishing technical challenge of viral synthesis and the push for public availability of viral genome sequences, a forward-looking criticism of zoonosis prediction is that pre-identification of viruses with pandemic potential might have dual-use applications in bioterrorism [41]. At present, the inability of models to accurately predict viral characteristics relevant to a bioweapon (e.g. disease severity or transmissibility) would make predicted zoonoses a poor blueprint for creating a pandemic compared with historical or contemporary viruses with verified pandemic capability. However, as recent years have shown, technological advances powered by artificial intelligence may not be incremental (e.g. AlphaFold [42]). Model developers should therefore be cognizant of potential dual-use risks and consider the appropriateness of restricting access to data and/or model predictions.

Conclusions

It remains undetermined whether zoonosis prediction will ever generate virus-specific insights sufficient for prevention. We nevertheless argue for continued research in this area. First, large-scale comparative analyses of viruses are uniquely able to understand risk factors for emergence that generalise across viruses. Characterising viral communities in units of predicted zoonotic risk rather than viral species richness or phylogenetic diversity could identify high-risk interfaces for surveillance. Researchers focused on how environmental change affects emergence could similarly study net zoonotic risk at the viral community level or choose to study focal, 'model' viruses that have predicted risk over subjective alternatives [43]. Similarly, if genomic risk factors for human infection that span viral groups exist, as suggested by results from [26], this would drive fundamental and applied research in virology in directions that could not have otherwise been easily identified. Second, predictive models add value to surveillance studies at the human–animal interface by providing a rational evidence base to select which viruses to monitor or study in laboratories [44]. Importantly, the alternative approach — focusing on close phylogenetic relatives of zoonoses — has demonstrated potential to misguide the allocation of limited resources [26]. Third, the only alternative to bolstering capacity for prevention via virus prospecting and risk triaging would be heightened monitoring of high-risk human populations to more rapidly extinguish nascent pandemics, but focusing mitigation efforts on only a single stage of the emergence pathway carries substantial risk (Figure 1). While identification of agents is faster now than ever, the timeline between detection and an effective public health response remains a formidable challenge, with questionable viability for pathogens with even moderate transmissibility in humans [45]. Given the expected variance in human disease severity, some fraction of early cases would go undetected by hospital-based surveillance and detection in asymptomatic individuals via active surveillance would be unlikely to trigger action in the absence of additional information, thus further extending the time lag between zoonotic transmission, detection and action.

Virus discovery has exploded in recent years with the advent of metagenomic sequencing and shows no sign of decelerating [7,39]. As discussed above, these efforts have scientific value unrelated to pandemic prevention, though the capacity for the latter is at times exaggerated unhelpfully. When interpreted and communicated appropriately, complementing ongoing virus discovery with the development and refinement of inexpensive quantitative frameworks has few downsides and provides biological insights while increasing the relevance and efficiency of research and surveillance. Further, the relatively low cost of model development means that investment in this area need not compromise alternative investments such as heightened surveillance at high-risk interfaces, broad-acting preventive measures (e.g. personal protective equipment) or accelerated response. Given our current inability to either prevent spillover or extinguish developing epidemics, all tools available should be used to aid understanding of the process of viral emergence.

Data Availability

No data were used for the research described in the article.

Declaration of Competing Interest

None.

Acknowledgements

DS and NM were funded by a Wellcome Trust, UK Senior Research Fellowship (217221/Z/19/Z). Additional funding to DS was provided by the

National Science Foundation, USA/Biotechnology and Biological Sciences Research Council, UK Ecology and Evolution of Infectious Diseases Program (DEB 2011069, BB/V003798/1), the Leverhulme Trust, UK (PLP-2020-362) and the Medical Research Council, UK (MC_UU_12014/12). We thank Matt Arnold and Liam Brierley for helpful suggestions on earlier versions of this paper.

References and recommended reading

Papers of particular interest, published within the period of review, have been highlighted as:

- • of special interest
- •• of outstanding interest

1. Woolhouse M, Gaunt E: **Ecological origins of novel human pathogens**. *Crit Rev Microbiol* 2007, **33**:231-242.
2. Carroll D, Daszak P, Wolfe ND, Gao GF, Morel CM, Morzaria S, Pablos-Méndez A, Tomori O, Mazet JAK: **The global virome project**. *Science* 2018, **359**:872-874.
3. Dobson AP, Pimm SL, Hannah L, Kaufman L, Ahumada JA, Ando AW, Bernstein A, Busch J, Daszak P, Engelmann J, *et al*.: **Ecology and economics for pandemic prevention**. *Science* 2020, **369**:379-381.
4. Epstein JH, Anthony SJ: **Viral discovery as a tool for pandemic preparedness**. *Rev Sci Tech OIE* 2017, **36**:499-512.
5. Warren CJ, Sawyer SL: **Identifying animal viruses in humans**. *Science* 2023, **379**:982-983.
6. Trock SC, Burke SA, Cox NJ: **Development of framework for assessing influenza virus pandemic risk**. *Emerg Infect Dis* 2015, **21**:1278-1372.
7. Gibb R, Albery GF, Mollentze N, Eskew EA, Brierley L, Ryan SJ, Seifert SN, Carlson CJ: **Mammal virus diversity estimates are unstable due to accelerating discovery effort**. *Biol Lett* 2022, **18**:20210427.
8. Edgar RC, Taylor J, Lin V, Altman T, Barbera P, Meleshko D, Lohr D, Novakovsky G, Buchfink B, Al-Shayeb B, *et al*.: **Petabase-scale sequence alignment catalyses viral discovery**. *Nature* 2022, **602**:142-147.
9. • Hou X, He Y, Fang P, Mei S-Q, Xu Z, Wu W-C, Zhang S, Zeng Z-Y, Gou Q-Y, Xin G-Y, *et al*.: **Artificial intelligence redefines RNA virus discovery**. *bioRxiv* 2023, https://doi.org/10.1101/2023.04.18.537342.

Discovered a massive diversity of 'dark' viral matter within meta-transcriptomes using a deep learning model trained to identify viral RNA-dependent RNA polymerase using amino acid and structural information.

10. O'Brien EC, Taft R, Taft K, Ciotti M, Suk JE: **Best practices in ranking communicable disease threats: a literature review, 2015**. *Eurosurveillance* 2016, **21**:30212.
11. Mehand MS, Millett P, Al-Shorbaji F, Roth C, Kieny MP, Murgue B: **World Health Organization methodology to prioritize emerging infectious diseases in need of research and development**. *Emerg Infect Dis* 2018, **24**:e171427.
12. Viana M, Mancy R, Biek R, Cleaveland S, Cross PC, Lloyd-Smith JO, Haydon DT: **Assembling evidence for identifying reservoirs of infection**. *Trends Ecol Evol* 2014, **29**:270-279.
13. Brierley L, Fowler A: **Predicting the animal hosts of coronaviruses from compositional biases of spike protein and whole genome sequences through machine learning**. *PLoS Pathog* 2021, **17**:e1009149.
14. Olival KJ, Hosseini PR, Zambrana-Torrelio C, Ross N, Bogich TL, Daszak P: **Host and viral traits predict zoonotic spillover from mammals**. *Nature* 2017, **546**:646-650.
15. Woolhouse MEJ, Gowtage-Sequeria S: **Host range and emerging and reemerging pathogens**. *Emerg Infect Dis* 2005, **11**:1842-1847.
16. Valero-Rello A, Sanjuán R: **Enveloped viruses show increased propensity to cross-species transmission and zoonosis**. *Proc Natl Acad Sci* 2022, **119**:e2215600119.
17. Grange ZL, Goldstein T, Johnson CK, Anthony S, Gilardi K, Daszak P, Olival KJ, O'Rourke T, Murray S, Olson SH, *et al*.: **Ranking the risk of animal-to-human spillover for newly discovered viruses**. *Proc Natl Acad Sci* 2021, **118**:e2002324118.
18. Pandit PS, Anthony SJ, Goldstein T, Olival KJ, Doyle MM, Gardner NR, Bird B, Smith W, Wolking D, Gilardi K, *et al*.: **Predicting the potential for zoonotic transmission and host associations for novel viruses**. *Commun Biol* 2022, **5**:844.
19. Poisot T, Ouellet M-A, Mollentze N, Farrell MJ, Becker DJ, Brierley L, Albery GF, Gibb RJ, Seifert SN, Carlson CJ: **Network embedding unveils the hidden interactions in the mammalian virome**. *Patterns* 2023, **4**:100738.
20. Kitchen A, Shackelton LA, Holmes EC: **Family level phylogenies reveal modes of macroevolution in RNA viruses**. *Proc Natl Acad Sci* 2011, **108**:238-243.
21. Babayan SA, Orton RJ, Streicker DG: **Predicting reservoir hosts and arthropod vectors from evolutionary signatures in RNA virus genomes**. *Science* 2018, **362**:577-580.
22. Tang Q, Song Y, Shi M, Cheng Y, Zhang W, Xia X-Q: **Inferring the hosts of coronavirus using dual statistical models based on nucleotide composition**. *Sci Rep* 2015, **5**:17155.
23. Martínez MA, Jordan-Paiz A, Franco S, Nevot M: **Synonymous virus genome recoding as a tool to impact viral fitness**. *Trends Microbiol* 2016, **24**:134-147.
24. Bartoszewicz JM, Seidel A, Renard BY: **Interpretable detection of novel human viruses from genome sequencing data**. *NAR Genom Bioinform* 2021, **3**:lqab004.
25. Young F, Rogers S, Robertson DL: **Predicting host taxonomic information from viral genomes: a comparison of feature representations**. *PLoS Comput Biol* 2020, **16**:e1007894.
26. •• Mollentze N, Babayan SA, Streicker DG: **Identifying and prioritizing potential human-infecting viruses from their genome sequences**. *PLoS Biol* 2021, **19**:e3001390.

Developed machine learning models that identified candidate zoonoses based on putatively cross-viral species signals of preadaptation to infect humans in viral genomes.

27. Di Giallonardo F, Schlub TE, Shi M, Holmes EC: **Dinucleotide composition in animal RNA viruses is shaped more by virus family than by host species**. *J Virol* 2017, **91**:e02381-16.
28. Lundberg SM, Erion G, Chen H, DeGrave A, Prutkin JM, Nair B, Katz R, Himmelfarb J, Bansal N, Lee S-I: **From local explanations to global understanding with explainable AI for trees**. *Nat Mach Intell* 2020, **2**:56-67.
29. • Wardeh M, Blagrove MSC, Sharkey KJ, Baylis M: **Divide-and-conquer: machine-learning integrates mammalian and viral traits with network features to predict virus-mammal associations**. *Nat Commun* 2021, **12**:1-15 3954.

Used virus traits, host traits and host-virus association networks to predict missing components of the host range of mammal-infecting viruses.

30. •• Liu D, Young F, Robertson DL, Yuan K: **Prediction of virus-host association using protein language models and multiple instance learning**. *bioRxiv* 2023, https://doi.org/10.1101/2023.04.07.536023.

Showed that pre-trained large protein language models outperform traditional, manually-curated representations of viral proteins to predict viral host range while remaining interpretable.

31. Xu GJ, Kula T, Xu Q, Li MZ, Vernon SD, Ndung'u T, Ruxrungtham K, Sanchez J, Brander C, Chung RT, *et al*.: **Comprehensive serological profiling of human populations using a synthetic human virome**. *Science* 2015, **348**:aaa0698.
32. Allen T, Murray KA, Zambrana-Torrelio C, Morse SS, Rondinini C, Di Marco M, Breit N, Olival KJ, Daszak P: **Global hotspots and correlates of emerging zoonotic diseases**. *Nat Commun* 2017, **8**:1-10 1124.

33. Geoghegan JL, Senior AM, Giallonardo FD, Holmes EC: **Virological factors that increase the transmissibility of emerging human viruses**. *Proc Natl Acad Sci* 2016, **113**:4170-4175.

34. Walker JW, Han BA, Ott IM, Drake JM: **Transmissibility of emerging viral zoonoses**. *PLoS One* 2018, **13**:e0206926.

35. Brierley L, Pedersen AB, Woolhouse MEJ: **Tissue tropism and transmission ecology predict virulence of human RNA viruses**. *PLoS Biol* 2019, **17**:e3000206.

36. • Warren CJ, Yu S, Peters DK, Barbachano-Guerrero A, Yang Q, Burris BL, Worwa G, Huang I-C, Wilkerson GK, Goldberg TL, *et al*.: **Primate hemorrhagic fever-causing arteriviruses are poised for spillover to humans**. *Cell* 2022, **185**:3980-3991.e18.
Showed how applying laboratory assays to animal viruses can risk assess interactions with cellular and immunological barriers to human infection prior to zoonotic transmission and advocated using these results to motivate serological surveillance in humans.

37. Letko M, Marzi A, Munster V: **Functional assessment of cell entry and receptor usage for SARS-CoV-2 and other lineage B betacoronaviruses**. *Nat Microbiol* 2020, **5**:562-569.

38. Carlson CJ, Zipfel CM, Garnier R, Bansal S: **Global estimates of mammalian viral diversity accounting for host sharing**. *Nat Ecol Evol* 2019, **3**:1070-1075.

39. •• Wille M, Geoghegan JL, Holmes EC: **How accurately can we assess zoonotic risk?** *PLoS Biol* 2021, **19**:e3001135.
Described current trends in virus discovery and speculated that disproportionate study effort for viruses with human or animal health significance might undermine efforts to predict human infection.

40. Shi M, Lin XD, Chen X, Tian JH, Chen LJ, Li K, Wang W, Eden JS, Shen JJ, Liu L, *et al*.: **The evolutionary history of vertebrate RNA viruses**. *Nature* 2018, **556**:197-202.

41. Sandbrink J, Ahuja J, Swett J, Koblentz G, Standley C: **Mitigating biosecurity challenges of wildlife virus discovery**. *SSRN* 2022, https://doi.org/10.2139/SSRN.4035760

42. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, Tunyasuvunakool K, Bates R, Žídek A, Potapenko A, *et al*.: **Highly accurate protein structure prediction with AlphaFold**. *Nature* 2021, **596**:583-589.

43. Bergner LM, Mollentze N, Orton RJ, Tello C, Broos A, Biek R, Streicker DG: **Characterizing and evaluating the zoonotic potential of novel viruses discovered in vampire bats**. *Viruses* 2021, **13**:252.

44. Sun Y, Zhang K, Qi H, Zhang H, Zhang S, Bi Y, Wu L, Sun L, Qi J, Liu D, *et al*.: **Computational predicting the human infectivity of H7N9 influenza viruses isolated from avian hosts**. *Transbound Emerg Dis* 2021, **68**:846-856.

45. Ferguson NM, Cummings DAT, Cauchemez S, Fraser C, Riley S, Meeyai A, Iamsirithaworn S, Burke DS: **Strategies for containing an emerging influenza pandemic in Southeast Asia**. *Nature* 2005, **437**:209-214.

46. Gibb R, Albery GF, Becker DJ, Brierley L, Connor R, Dallas TA, Eskew EA, Farrell MJ, Rasmussen AL, Ryan SJ, *et al*.: **Data proliferation, reconciliation, and synthesis in viral ecology**. *BioScience* 2021, **71**:1148-1156.