JEROME FAVRE/BLOOMBERG VIA GETTY

Live chickens on sale at a market in Hong Kong in 2013, during an outbreak of avian influenza in people.

Predicting a pandemic

Machine learning could help to identify the viruses most likely to spill over from animals to people and cause future pandemics. **By Simon Makin**

In February 2021, seven Russian poultry-farm workers were reported to have been infected with H5N8 avian influenza. This subtype of bird flu had never been known to infect people before, and the virus's genetic sequence was quickly uploaded to the genetic data repository GISAID. For Colin Carlson, a biologist at Georgetown University in Washington DC, it presented an opportunity. "I immediately thought, 'I want to run this through FluLeap'," he says.

FluLeap is a machine-learning algorithm that uses sequence data to classify influenza viruses as either avian or human. The model had been trained on a huge number of influenza genomes — including examples of H5N8 — to learn the differences between those that infect people and those that infect birds. But the model had never seen an H5N8 virus categorized as human, and Carlson was curious to see what it made of this new subtype.

Somewhat surprisingly, the model identified it as human with 99.7% confidence. Rather than simply reiterating patterns in its training data, such as the fact that H5N8 viruses do not typically infect people, the model seemed to have inferred some biological signature of compatibility with humans. "It's stunning that the model worked," says Carlson. "But it's one data point; it would be more stunning if I could do it a thousand more times."

The zoonotic process of viruses jumping from wildlife to people causes most pandemics. As climate change and human encroachment on animal habitats increase the frequency of these events, understanding zoonoses is crucial to efforts to prevent pandemics, or at least to be better prepared.

Researchers estimate that around 1% of the mammalian viruses on the planet have been identified[1], so some scientists have attempted to expand our knowledge of this global virome by sampling wildlife. This is a huge task, but over the past decade or so, a new discipline has emerged — one in which researchers use statistical models and machine learning to predict aspects of disease emergence, such as global hotspots, likely animal hosts or the ability of a particular virus to infect humans. Advocates of such 'zoonotic risk prediction' technology argue that it will allow us to better target surveillance to the right areas and situations, and guide the development of vaccines and therapeutics that are most likely to be needed.

However, some researchers are sceptical of the ability of predictive technology to cope with the scale and ever-changing nature of the virome. Efforts to improve the models and the data they rely on are under way, but these tools will need to be a part of a broader effort if they are to mitigate future pandemics.

Virus hunting

Some researchers have long argued that expanding our knowledge of viral diversity will help to manage pandemic threats. PREDICT, a US$200-million project funded by the US Agency for International Development (USAID), spent around a decade looking for animal viruses. By the time it ended in 2020, it had identified 949 new viruses in samples from wildlife, livestock and people, in 34 countries.

Some of PREDICT's findings might seem prescient, in hindsight. A 2017 study[2] estimated that there are thousands of undiscovered coronaviruses in bats (widely thought to be the source of the virus SARS-CoV-2), and predicted that southeast Asia would be home to the greatest number of viruses in the family to which SARS-CoV-2 belongs. It also associated activities that involve high levels of human–wildlife contact, such as wildlife markets, with a higher prevalence of coronaviruses.

Another 2017 study[3] collected data on which viruses infect which mammals, creating a database of virus–host associations. "The goal was to understand which viruses are capable of infecting people, what animals we're most often getting new viruses from and the underlying factors that drive those patterns," says ecologist and study leader Kevin Olival at the EcoHealth Alliance in New York City, a non-profit body focused on bio-surveillance and conservation. The team's analysis showed that the proportion of viruses in a given host species that can infect humans is affected by how closely related humans are to that species, as well as factors that influence human–wildlife contact, such as the human population density and the degree of urbanization in that species' geographical range. The team used statistical modelling to predict animal groups and regions that were likely to harbour a large number of undiscovered viruses — bats featured prominently, along with rodents and primates, in regions including South America, Africa and southeast Asia. The researchers also found traits associated with a virus being zoonotic, such as the range of species it can infect.

The team says this information can help to guide surveillance efforts. "It allows us to forecast areas most at risk," says Jonna Mazet, an epidemiologist at the University of California, Davis, who directed PREDICT. Identifying specific threats also allows local researchers and health-care workers to tailor mitigation and response capabilities. "It allows communities to say 'we have this, this and this, and we can reduce our risk in these ways'," says Mazet.

"With a trillion points, you could predict spillover like the weather."

PREDICT was intended to be just a pilot project. "It generated a lot of data, but it was a drop in the bucket," says Olival. "We need something bigger." Researchers therefore proposed the Global Virome Project (GVP) in 2016, seen as a global partnership of government agencies, non-governmental organizations and researchers, with the aim of discovering most of the viruses in mammals and birds (from which most zoonotic viruses originate). However, in the face of criticism from some researchers, it has never been funded. It exists today as a non-profit organization, aiming to provide countries with the knowledge required to carry out their own viral surveys, Mazet says. A smaller, much less costly project called Discovery and Exploration of Emerging Pathogens — Viral Zoonoses (DEEP VZ) was launched by USAID in October 2021.

One criticism of the GVP is that the scale of the task is simply unmanageable. PREDICT researchers estimate[4] that there are 1.67 million unknown viruses in mammals and birds, and although this figure is contested, there is no doubt that the virome is vast. It is also constantly changing, so one-off discovery efforts would not be enough. "RNA viruses evolve at a hefty rate," says Edward Holmes, a virologist at the University of Sydney in Australia. "So you'd have to keep doing it."

There is also scepticism that the project would have identified potential pandemics. "I have no problem with it in terms of understanding virus evolution and ecology," Holmes says. "But as a predictive tool to understand what comes next, it's a non-starter." One issue is that some host species and viral families have been intensively studied, but others have hardly been touched. Existing data are also skewed towards viruses that have already spilled over[5]. As a result, most predictions so far have been based on "completely biased data", says Jemma Geoghegan, a virologist at the University of Otago in New Zealand. Moreover, even when a virus is discovered and its genome is sequenced, many factors that can influence its potential to spark a pandemic, such as its ability to infect humans and be transmitted from person to person, will still be unclear. "You've then got to do all these experiments, which will take years and cost a fortune," says Holmes.

This is where machine learning might provide a short cut. Rather than attempting to fully characterize every new virus, models could be used to flag high-priority targets for further investigation. "What we need is a triaging system downstream, so we know which viruses need to be characterized with in-depth virology studies," says Sara Sawyer, a virologist at the University of Colorado, Boulder.

Inside the models

When a virus is discovered, often little is known about it other than its genetic sequence. Models that can triage viruses using only their genomes would therefore be particularly useful. Nardus Mollentze, a computational virologist at the University of Glasgow, UK, and his colleagues have developed one such model, which assesses viruses in part by using a measure of their genetic similarity to parts of the human genome[6]. Evolutionary pressure on viruses can result in genetic segments that resemble those in the host's genome — either to evade the innate immune system or to aid replication. When tested on a library of 861 known viruses, the algorithm could classify them as zoonotic or not with 70% accuracy.

Mollentze has since joined the Viral Emergence Research Initiative (Verena), a consortium of researchers seeking to develop and improve zoonotic prediction models. Mollentze collaborated with Verena researchers to combine his algorithm with techniques that exploit knowledge of which viruses infect which hosts, including methods for inferring unknown host–virus associations. This combined approach raised performance by roughly ten percentage points[7]. In future, knowledge of how viruses interact with hosts on a molecular level could be incorporated. "It's going to be all about proteins and biochemistry," says Carlson, who directs Verena. "That's the future of this."

ATTILABARSAN/ GETTY

Bats harbour many unknown coronaviruses.

An important goal is to learn which models work well, and why. There are models that merely classify according to patterns in the data, and those that infer the reasons for those patterns, but it can be difficult to tell them apart. "There's this question: are we just teaching machines to reiterate things they already know, or are they learning principles that carry into new space?" says Carlson.

To make progress, the process of validating models will be crucial. For instance, several studies have tried to predict which species host zoonotic viruses, with mixed results, but there has been little systematic comparison, making it difficult to know which approaches work. To address this, in early 2020, Verena researchers used predictions of which bat species might host betacoronaviruses as a case study[8]. They created eight statistical models and used them to generate a list of suspected hosts. In the following 16 months, 47 new bat hosts were discovered. When the researchers compared these with their predictions, they found that half of the models performed significantly better than chance. These models included traits such as the species' lifespan or size. The other four models did not take such features into account and performed poorly.

Data developments

Any artificial intelligence (AI) algorithm is fundamentally limited by the data it is fed. "AI works when the algorithm is trained on large amounts of quality data," says Sawyer. "But only a small number of spillovers occur each year, and data on viruses tend to be dirty, with a lot of missing information." Most researchers agree that the data are currently insufficient. "We don't have enough high-quality data to do a good job at prediction," says Mazet.

To some extent, modelling relies on scientists gathering fresh data, but viral-discovery efforts so far have been motivated by considerations such as the highest-risk places and situations. What modellers actually need is sampling aimed at improving geographical and taxonomic coverage, Carlson says. Supplying models with more data of this kind changes the horizon of what questions can be asked. "With a million data points, you can show how deforestation increases viral prevalence in bats," Carlson says. "With a trillion points, you could predict spillover like the weather."

To get anywhere close to that would require global cooperation, with open data sharing as the norm and data standards that everyone adheres to. The obstacles to this are more political, cultural and ethical than scientific. Academic incentives around publications, for example, are an obstacle to rapid data sharing. Guaranteeing that countries that share genetic data benefit from doing so is also crucial. "That's the key issue and dealing with it involves building trust," says Olival. "Making sure you're giving back, not only with vaccines, but with training, capacity building and co-authorship on papers."

The Nagoya Protocol, an international treaty that came into effect in 2014, enshrines countries' sovereignty over natural resources, including biological samples, and allows them to require benefit-sharing agreements in return for access to such samples. However, some labs can now synthesize pathogens or begin to develop vaccines using just genetic sequencing data. "We don't have anything set up in international law that deals with sequence data," says Carlson. "Nagoya isn't made for that world." Similar issues might some day apply to zoonotic risk prediction. "We're using data collected by researchers in the global south," says Carlson. "There are legitimate questions about what it means to take that data and make a technology."

Predict and prepare

For modelling to have real-world impact, it must lead to publicly accessible tools that provide actionable, locally relevant information. Modelling also needs to be better integrated with experimental work to interrogate the characteristics of pathogens. Just as a model might flag candidate viruses for further study, so might those investigations produce information that can be used to validate and refine the models. However, interdisciplinary communication is currently limited. "These are communities that don't talk or even read each other's papers much," says Sawyer.

Modellers also need to clearly communicate the uncertainty inherent in their work, and what they mean by prediction so they do not oversell the benefits. "No one says we're going to have the exact time, place and species that will lead to the next pandemic," says Olival. Researchers are dealing with probabilities, and unexpected things can and do happen.

Even at their best, predictive tools are not going to be able to completely prevent outbreaks. "I absolutely do not think we should hinge the world's security on these models," says Carlson. But alongside improved global surveillance systems, targeted vaccine development and efforts to build health-care capacity worldwide, their value is clear. "They let us do two things: understand what's happening around us and prioritize," Carlson says. Ultimately, that might help to reduce the frequency of pandemics. "We can get better at preventing some of them," says Carlson. "But it requires us to get better at what we're doing."

Simon Makin is a freelance writer in Reading, UK.

1. Carlson, C. J. *et al. Phil. Trans. R. Soc. Lond. B* **376**, 20200358 (2021).
2. Anthony, S. J. *et al. Virus Evol.* **3**, vex012 (2017).
3. Olival, K. *et al. Nature* **546**, 646–650 (2017).
4. Carroll, D. *et al. Science* **359**, 872–874 (2018).
5. Wille, M., Geoghegan, J. L. & Holmes, E. C. *PLoS Biol.* **19**, e3001135 (2021).
6. Mollentze, N., Babayan, S. A. & Streicker, D. G. *PLoS Biol.* **19**, e3001390 (2021).
7. Poisot, T. *et al.* Preprint at https://arxiv.org/abs/2105.14973 (2022).
8. Becker, D. J. *Lancet Microbe* **3**, E625–E637 (2022).