



OPEN ACCESS

EDITED BY

Deepak Y. Patil,
National Institute for One Health, India

REVIEWED BY

Srivastava Vartika,
Cleveland Clinic, United States
Saeed Soleiman-Meigooni,
Aja University of Medical Sciences, Iran

*CORRESPONDENCE

David Chisompola
✉ d.chisompola@gmail.com

RECEIVED 02 October 2025

REVISED 04 December 2025

ACCEPTED 22 December 2025

PUBLISHED 12 January 2026

CITATION

Chisompola D, Luwaya E, Nzobokela J,
Mwansa P and Chakulya M (2026)
AI-powered analysis of viral metagenomic
sequencing data for rapid outbreak
investigation and novel pathogen discovery.
Front. Microbiol. 16:1717859.
doi: 10.3389/fmicb.2025.1717859

COPYRIGHT

© 2026 Chisompola, Luwaya, Nzobokela,
Mwansa and Chakulya. This is an
open-access article distributed under the
terms of the [Creative Commons Attribution
License \(CC BY\)](#). The use, distribution or
reproduction in other forums is permitted,
provided the original author(s) and the
copyright owner(s) are credited and that the
original publication in this journal is cited, in
accordance with accepted academic
practice. No use, distribution or reproduction
is permitted which does not comply with
these terms.

AI-powered analysis of viral metagenomic sequencing data for rapid outbreak investigation and novel pathogen discovery

David Chisompola*, Emmanuel Luwaya, John Nzobokela,
Phinnoty Mwansa and Martin Chakulya

School of Medicine and Health Sciences, Mulungushi University, Livingstone, Zambia

Emerging viral outbreaks continue to pose a persistent global health threat, underscoring the urgent need for a shift from reactive to proactive health security strategies. Viral metagenomic next-generation sequencing (mNGS) offers an unbiased, powerful approach to pathogen detection and discovery, yet its utility has been constrained by the computational complexity and slow turnaround time of data analysis during outbreak crises. The integration of artificial intelligence (AI) and mNGS is dismantling these barriers, enabling faster, more scalable outbreak response. This review synthesizes how AI-driven analytics are transforming mNGS applications, from genome assembly to sequence classification, using advanced architectures such as convolutional neural networks, recurrent neural networks, and transformers. Beyond accelerating workflows, AI's capacity for pattern recognition outperforms traditional homology-based methods, facilitating the discovery of novel viral families and tracing hidden transmission chains through anomaly detection. Nonetheless, critical challenges remain, including limited training data, the interpretability of AI models, and resource-intensive computational demands that risk widening an "AI divide" in global health. We evaluate these obstacles and highlight forward-looking strategies, including federated learning for privacy-preserving data sharing and explainable AI for improving trust and biological insight. Looking ahead, we envision an "AI-first" paradigm for outbreak preparedness, anchored in integrated "Digital Immune Systems" for continuous, global-scale surveillance. By framing the synergy between mNGS and AI as a transformative leap, this review underscores its potential to strengthen resilience against future pandemics.

KEYWORDS

artificial intelligence, outbreak investigation, pandemic preparedness, pathogen discovery, viral metagenomics

1 Introduction

Emerging and re-emerging viral pathogens continue to threaten global health security, as evidenced by outbreaks of Ebola virus, Zika virus, SARS-CoV-2, and, most recently, Mpox (Han et al., 2023). These events highlight how infectious diseases can trigger epidemics and pandemics that overwhelm healthcare systems and cause widespread societal and economic disruption (Peters et al., 2020; Bavinger et al., 2020). The frequency of such outbreaks, particularly those of zoonotic origin, is increasing, driven by factors including climate change, ecological disruption, and intensified global connectivity (Wilder-Smith, 2021). A critical determinant in mitigating the impact of these events is the speed of the public health response. The rapid and accurate identification of the causative pathogen is the essential first step for

initiating effective containment, guiding therapeutic development, and deploying public health interventions (Yimer et al., 2024).

To identify pathogens, public health laboratories employ a variety of testing methods. Traditional assays include microscopy, culture-based analyses, and immunoassays that detect either pathogen antigens or host immune responses (Miller et al., 2013; Roux et al., 2021). While highly specific, these methods often require prior knowledge of the pathogen and can be slow. The adoption of nucleic acid amplification tests (NAATs), such as (Polymerase chain reaction) PCR, marked a significant advancement in speed and sensitivity but remains inherently targeted (Miller et al., 2013; Khan et al., 2024). Despite the availability of conventional testing approaches, many samples submitted to public health laboratories during outbreaks remain undiagnosed, leaving critical questions unanswered and exposing the limitations of standard diagnostic methods. This diagnostic gap has positioned metagenomic next-generation sequencing (mNGS) as a pivotal frontier for novel viral discovery.

Viral metagenomic next-generation sequencing, which enables the analysis of DNA and/or RNA from a sample (Roux et al., 2021), has emerged as a powerful tool for pathogen detection (Khan et al., 2024). By comprehensively interrogating nucleic acids in clinical and environmental samples, mNGS can identify known and novel viruses without prior knowledge of a causative agent (Roux et al., 2021; Khan et al., 2024; Mokili et al., 2012). This unbiased nature makes it indispensable for investigating unknown outbreaks. Despite its potential, the vast quantity and complexity of metagenomic datasets pose significant analytical challenges, especially when timely interpretation is critical during an emerging outbreak. However, analyzing mNGS data for novel viral discovery is challenging, as it demands specialized expertise in bioinformatics and data analysis. The introduction of artificial intelligence (AI) tools in mNGS is reshaping this landscape by enabling faster and more accurate interpretation of sequencing data, thereby accelerating the identification of novel pathogens.

Artificial intelligence such as, deep learning (DL) and machine learning (ML) have increasingly been recognized as transformative tools in biomedical data science, offering novel solutions for pattern recognition, anomaly detection, and predictive modeling (Hanna et al., 2025; Srivastava et al., 2025). In viral metagenomics, AI-powered methods have the potential to accelerate genome assembly, improve classification of unknown sequences, and uncover features that may otherwise be overlooked by conventional bioinformatics pipelines (Dhaarani and Reddy, 2025). The integration of AI into mNGS analysis promises not only to enhance outbreak investigations (Haug and Drazen, 2023), but also to expand the discovery of novel viruses with epidemic or pandemic potential (Liu, 2025).

Despite these advances, the application of AI in viral metagenomics is still in its early stages. Challenges such as data quality, model interpretability, computational resource demands, and ethical considerations hinder widespread adoption, particularly in resource-limited settings (Chong et al., 2025). Although AI has shown impressive advances in medicine, its application has faced significant obstacles especially when dealing with sensitive patient data. As global health systems seek faster and more reliable outbreak detection methods, understanding the opportunities and limitations of AI-driven metagenomic analysis becomes increasingly important.

While previous reviews have examined viral metagenomics and AI separately, few have explored their integration in outbreak

investigation. This review provides a comprehensive overview of how the synergy between mNGS, and AI transforms rapid outbreak response and the systematic exploration of novel viruses. We outline current sequencing platforms, analytical frameworks, and AI applications for pathogen detection and discovery, while addressing key challenges, ethical considerations, and future directions. By consolidating recent evidence, we emphasize the potential of AI-powered metagenomics to advance outbreak investigation and novel pathogen discovery.

2 Methods for literature review

To ensure a comprehensive and systematic synthesis of current knowledge, this review was conducted according to a structured methodological framework.

2.1 Information sources and search strategy

A systematic literature search was performed across three major bibliographic databases: PubMed, Scopus, and Web of Science. To capture the most recent advancements, the search was limited to articles published between January 2014 and September 2025. The search strategy employed a combination of keywords and Medical Subject Headings (MeSH) terms related to the core concepts. The primary search string was: (“viral metagenomic*” OR “metagenomic next-generation sequencing” OR “mNGS”) AND (“artificial intelligence” OR “machine learning” OR “deep learning”) AND (“outbreak investigation” OR “pathogen discovery” OR “pandemic preparedness”). This string was adapted to the syntax requirements of each database.

2.2 Inclusion and exclusion criteria

The inclusion criteria prioritized peer-reviewed original research articles, case studies, reviews, and seminal methodological papers. Eligible studies were those published in English and focused on the application of AI/ML to mNGS data for viral detection, classification, or outbreak analytics. Studies addressing the use of mNGS in outbreak investigations or novel pathogen discovery were also included. However, studies were excluded if they focused exclusively on bacterial, fungal, or parasitic pathogens without a viral component. Articles were also excluded if AI/ML was applied solely to non-metagenomic data, such as medical imaging or electronic health records, without a direct connection to genomic sequence analysis. Conference abstracts, editorials, non-peer-reviewed commentaries, and studies lacking full-text availability were similarly omitted.

2.3 Study selection process

The search results from all databases were consolidated, and duplicate records were removed using reference management Zotero version 7.0.30. The selection process adhered to a two-stage screening protocol to ensure rigor and minimize bias. First, two independent

reviewers screened the titles and abstracts of all retrieved records against the predefined eligibility criteria. Any discrepancies or conflicts regarding inclusion at this stage were resolved through consensus discussion or, when necessary, arbitration by a third senior reviewer. Second, the full-text articles of all records deemed potentially relevant during the initial screening were retrieved and subjected to a comprehensive eligibility assessment by the same two reviewers. Final inclusion decisions were made based on strict application of the criteria. The results of this systematic selection process are detailed in the PRISMA-style flow diagram (Figure 1).

2.4 Data extraction and synthesis

Data from included studies were extracted into a standardized form, capturing information on study objectives, AI/ML methodologies, sequencing platforms, key findings, and identified challenges. Given the diverse and rapidly evolving nature of the field, a narrative synthesis approach was employed. Findings were

thematically organized to construct a coherent overview of current applications, comparative advantages, persistent challenges, and future directions of AI-powered viral metagenomics.

3 An integrated framework for outbreak response

The rapid characterization of a viral pathogen is a critical determinant in the successful containment of an outbreak and the mitigation of its public health impact. No single technology operates in a vacuum; rather, a synergistic combination of tools creates a powerful, multi-layered defense system (Thomas et al., 2012). This integrated framework strategically leverages the unique strengths of rapid antigen tests, portable sequencing, high-throughput genomics, and artificial intelligence to create a cohesive pipeline from initial suspicion to definitive public health action (Linares et al., 2020).

Rapid antigen tests (RATs), serve as the crucial first line of defense (Cantón Cruz et al., 2025). These lateral flow immunoassays detect the

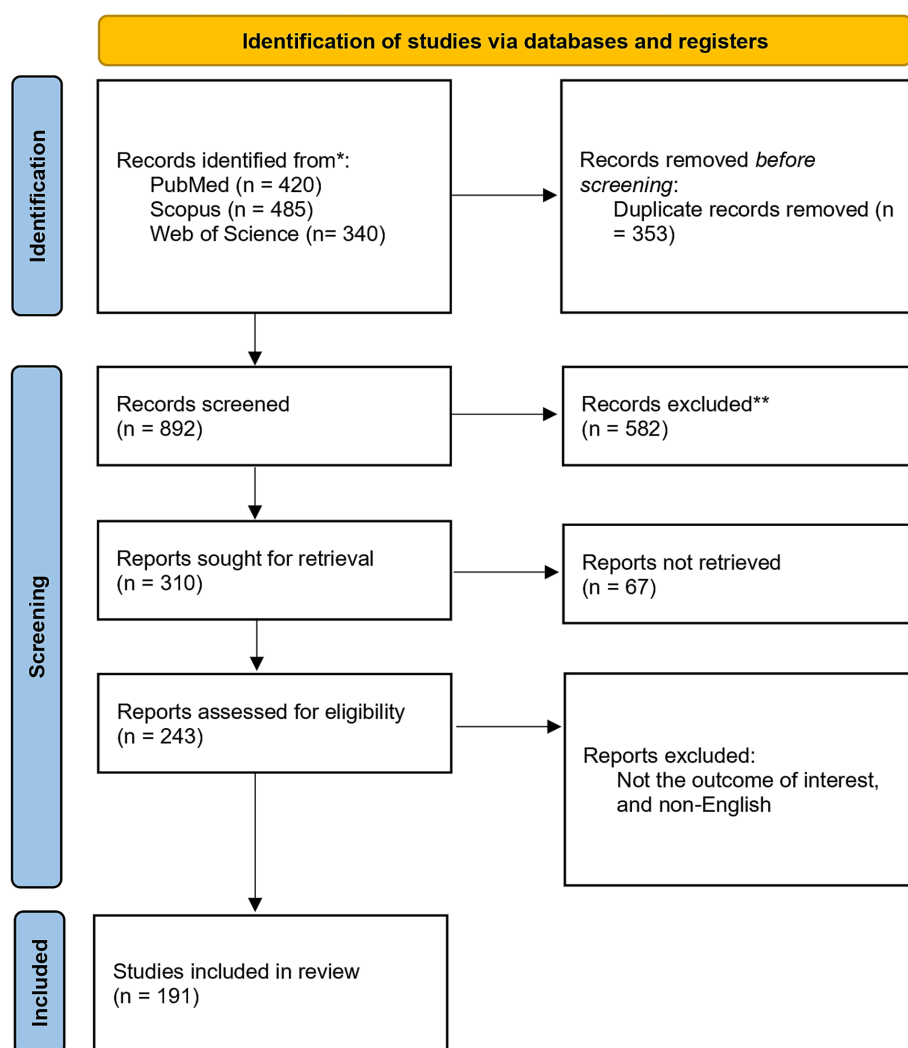


FIGURE 1
PRISMA flow diagram of literature selection.

presence of viral proteins, providing results in 15–30 min at the point of care. Their primary role is rapid case identification and triage, enabling immediate isolation and the prompt initiation of contact tracing to break chains of transmission in an outbreak's early stages (Yimer et al., 2024). One of their key strengths is that RATs can be deployed without the need for specialized infrastructure, requiring only minimal tools, which has made them particularly valuable in resource-limited settings. However, their analytical performance, particularly sensitivity and specificity, has shown considerable variability, necessitating confirmation with PCR-based methods (Yimer et al., 2024; Hirabayashi et al., 2024).

PCR based techniques remain the gold standard for detecting viral infections, including SARS-CoV-2 (Cantón Cruz et al., 2025), and mpox in recent outbreaks (da Silva et al., 2023; Li et al., 2010). Despite their high analytical performance, particularly in terms of sensitivity, specificity, and throughput capacity for testing large volumes of suspected cases within a limited time, PCR methods are constrained by high costs, long turnaround times, reliance on skilled personnel, and potential exposure risks at testing sites. Nonetheless, PCR continues to be the preferred method at many facilities, while RATs are increasingly adopted as complementary tools in outbreak settings (Cantón Cruz et al., 2025).

During outbreaks, neither RATs nor PCR techniques have demonstrated the capacity to identify novel viral pathogens, as both depend on prior knowledge of existing viruses for their design and clinical utility. This limitation underscores the need for advanced sequencing technologies to enable novel pathogen discovery. Several sequencing platforms are available, including first-generation methods such as Sanger sequencing, second-generation platforms like Illumina, and third-generation technologies such as PacBio (Heather and Chain, 2016).

Sanger sequencing, also known as the chain-termination method, was developed by Frederick Sanger in 1977 and is considered the first-generation DNA sequencing technology (Eren et al., 2022). It relies on the selective incorporation of chain-terminating dideoxynucleotides (ddNTPs) during DNA synthesis, producing fragments of varying lengths that can be resolved by capillary electrophoresis to determine the nucleotide sequence (Heather and Chain, 2016; Eren et al., 2022). Although limited by its application in large-scale outbreak settings, relatively low throughput, short read lengths, and higher costs compared to next-generation sequencing (NGS) methods, Sanger sequencing remains widely used due to its high accuracy, reliability, and suitability for small-scale projects such as gene validation, clinical diagnostics, and confirmatory sequencing. Its precision in detecting single nucleotide variants continues to make it a valuable reference standard, even in the era of high-throughput sequencing technologies.

Illumina sequencing delivers high-fidelity data for definitive analysis. Typically deployed on PCR or RAT-positive samples, Illumina's high accuracy and throughput are the cornerstone of genomic epidemiology (Huang et al., 2019). It enables precise reconstruction of transmission chains through whole-genome sequencing, distinguishes between multiple introductions of a virus, and powers large-scale surveillance to monitor for variants of concern. These attributes make Illumina particularly well-suited for outbreak investigations and the detection of novel viral pathogens.

Oxford Nanopore Technologies (ONT) sequencing provides real-time genomic intelligence. The portability of devices like the MinION allows for sequencing to be deployed directly in the field or in regional

laboratories (Lu et al., 2016). This facilitates rapid initial characterization of an outbreak, enabling immediate detection of genetic drift or the emergence of a novel variant. Most significantly, its capacity for long-read, unbiased metagenomic sequencing makes it a powerful tool for *de novo* viral discovery when targeted tests are negative.

Sequencing platforms generate massive datasets that necessitate robust bioinformatic analysis. The sheer volume of data makes timely sequence interpretation challenging, and bioinformatic tools are essential for identifying novel viral pathogens. However, traditional bioinformatics is often constrained by the computational resources, costs, and specialized expertise required, limitations that are especially pronounced in resource-limited settings. Since the emergence of artificial intelligence (AI), there has been a growing effort to develop and validate automated AI-driven tools capable of analyzing sequencing data rapidly, enabling near real-time insights to support public health responses during outbreaks.

The full potential of this approach is realized through integration (Figure 2). In this framework, samples collected from humans, animals, or environmental sources are routed for sequencing. ONT provides rapid, near-source intelligence for initial outbreak characterization, while Illumina supplies high-fidelity data for definitive reconstruction and long-term surveillance. This combination of speed, portability, and accuracy forms a robust system for mitigating the impact of viral outbreaks and accelerating the discovery of emerging pathogens.

4 Traditional outbreak screening methods and their limitations

In the recent past, outbreak screening has facilitated the development and refinement of a wide range of diagnostic methods for the early detection of infectious diseases (Watkins et al., 2006). These include culture-based techniques, direct microscopy, immunoassays (antigen and antibody detection), and targeted NAATs such as PCR (Ieven, 2007). Traditional methods have been instrumental not only in guiding clinical diagnosis but also in paving the way for more advanced techniques. However, evidence suggests that there is considerable variability in their application during outbreak investigations, with no single method consistently preferred across different pathogens or public health settings. This inconsistency stems largely from differences in analytical performance, the availability of technical expertise, infrastructure demands, turnaround times, and associated costs.

Traditional outbreak screening methodologies have primarily relied on a combination of clinical suspicion and targeted laboratory testing (Abat et al., 2016). Symptoms alone are rarely sufficient for accurate identification, especially given that many pathogens produce overlapping clinical syndromes. As a result, clinicians, including physicians, veterinarians, nurse practitioners, and pathologists, play a central role in identifying suspected cases by screening patients presenting with compatible symptoms, collecting appropriate specimens, and initiating laboratory confirmation (Wagner et al., 2006). In this way, clinical suspicion provides a gateway for standardized case definitions and subsequent confirmatory laboratory testing. Conventional diagnostic tools such as real-time RT-PCR and

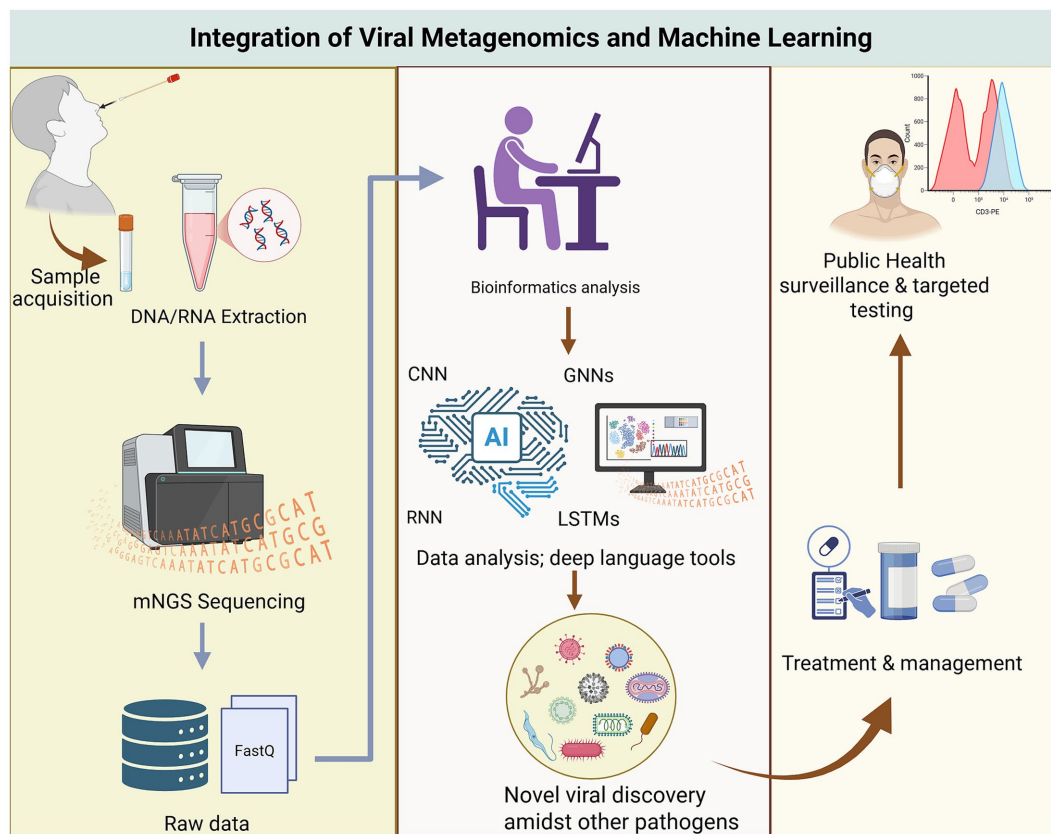


FIGURE 2
Integration of viral metagenomics with AI intelligence, machine learning and deep learning.

viral culture have thus been central to outbreak detection (Reintjes and Zanuzdana, 2009). Nonetheless, despite their long history of use, these traditional approaches face significant challenges related to speed, accuracy, cost, and scalability in the context of modern outbreaks.

4.1 Direct microscopy

Direct microscopy, including electron microscopy (EM), remains a rapid tool for the presumptive identification of pathogens, allowing direct visualization of viral particles without prior knowledge (Apollon et al., 2022). Historically, EM played a decisive role in virus discovery, such as the 1948 differentiation of smallpox from chickenpox (Goldsmith and Miller, 2009). advances like transmission electron microscopy (TEM) and cryogenic electron microscopy (cryo-EM) (Richert-Pöggeler et al., 2019), sensitivity is generally lower than culture or PCR, and diagnostic error is a risk, as seen during coronavirus outbreaks (Bullock et al., 2022; Curry, 2003). Furthermore, in the case of SARS-CoV-2, EM confirmation of viral presence has been limited to select tissues (lung, heart, olfactory mucosa, and placenta), with inconsistent findings elsewhere (Birkhead et al., 2021). Such limitations highlight that, while EM remains invaluable for the discovery and confirmation of novel or unusual pathogens, it is not suitable for routine or large-scale outbreak diagnostics.

4.2 Polymerase chain reaction methods

The advent of nucleic acid amplification techniques, particularly PCR, transformed diagnostic virology beginning in the 1970s (Leland and Ginocchio, 2007). PCR enables sensitive and highly specific detection of viral nucleic acids without the need for viral propagation in culture, making it a faster and more versatile tool than traditional methods. Modern real-time PCR and multiplex NAAT platforms now allow simultaneous detection of up to 15 viruses and 4 bacteria in a single assay, representing a major advancement in outbreak screening, particularly for respiratory infections (Das et al., 2015).

PCR methods typically achieve sensitivity near 95% and specificity approaching 100%, which has established them as the gold standard in many diagnostic contexts. Their ability to detect low viral loads early in infection is particularly advantageous during outbreak investigations, where rapid case identification is critical. However, PCR is not without limitations. False positives may occur due to sample contamination, while false negatives may result from poor sample quality, inhibitors, or genetic mutations in the target region (Ieven, 2007; Iwata, 2020). Furthermore, high costs of reagents, consumables, and equipment, as well as the requirement for reliable electricity and trained staff, limit widespread accessibility in many resource-constrained settings. Despite these barriers, PCR remains the most widely adopted tool for outbreak detection, bridging the gap between clinical suspicion and definitive laboratory confirmation.

4.3 Culture-based techniques

For much of the 20th century, culture-based methods were the cornerstone of viral diagnostics and were long regarded as the gold standard for pathogen identification (Leland and Ginocchio, 2007). Viruses such as vaccinia, smallpox, and yellow fever were among the earliest to be propagated in culture between 1913 and the 1950s, with subsequent breakthroughs following the discovery that poliovirus could grow in non-neural cell lines. Viral culture remains unmatched in its ability to generate live isolates for further characterization, including drug susceptibility testing, antigenic typing, and vaccine development.

However, culture-based techniques are inherently slow, often requiring days to weeks to yield results, an unacceptable delay in the context of rapidly evolving outbreaks. They are also technically demanding, requiring specialized laboratory facilities, strict biosafety protocols, and highly trained personnel. Contamination risks further complicated interpretation, sometimes necessitating repeat cultures and extending diagnostic timelines. In many low- and middle-income countries, inadequate infrastructure and resource constraints have restricted the use of viral culture in routine outbreak surveillance.

While culture retains value for research, reference laboratories, and vaccine development, its role in frontline outbreak detection has largely been superseded by faster and more sensitive molecular methods.

4.4 Immunoassays (antigen and antibody detection)

Immunoassays remain a cornerstone in viral discovery due to their specificity, sensitivity, and relative ease of implementation (Wang et al., 2023). These assays leverage highly selective interactions between viral antigens and host antibodies or between antibodies and viral antigens, enabling both direct and indirect detection of viral pathogens. The two primary categories—antigen detection and antibody detection—serve complementary roles in uncovering known and novel viruses (Pavia and Plummer, 2021).

Antigen-based immunoassays detect viral proteins directly in clinical or environmental samples, providing evidence of active infection (Louten, 2016). Techniques such as enzyme-linked immunosorbent assays (ELISAs), lateral flow assays, and chemiluminescent immunoassays utilize monoclonal or polyclonal antibodies to capture and quantify viral antigens. In the context of viral discovery, these assays can rapidly screen large sample sets, flagging potential cases for more detailed molecular characterization. For example, during early outbreak investigations, antigen assays have been critical in identifying emerging influenza strains or novel coronaviruses (Cantón Cruz et al., 2025; Hirabayashi et al., 2024; Pavia and Plummer, 2021), often preceding nucleic acid-based confirmation.

Serological immunoassays detect host antibodies generated in response to viral infection, offering insights into exposure history and immune response dynamics. ELISA, Western blotting, and multiplex immunoassays allow high-throughput screening for IgM, IgG, or IgA antibodies against viral antigens. In viral discovery, antibody detection is particularly valuable for identifying past or subclinical infections, uncovering viruses that may evade direct detection (Louten, 2016). Serology can also guide

epidemiological investigations, revealing the prevalence and distribution of previously unrecognized viral pathogens in populations.

Immunoassays are often used in tandem with molecular techniques to increase detection sensitivity and validate findings (Wang et al., 2023). For emerging viruses with limited genomic information, immunoassays can provide the first clues of viral presence by recognizing conserved structural proteins or cross-reactive epitopes. Furthermore, advances in recombinant antigen production, high-affinity antibody engineering, and multiplexed assay platforms have expanded the ability to detect multiple viral targets simultaneously, accelerating the pace of discovery (Matsunaga and Tsumoto, 2025).

Despite their utility, immunoassays face several limitations. Cross-reactivity with related viruses may produce false positives, particularly in antibody-based assays (Luvira et al., 2022). Antigen assays may have reduced sensitivity in low-viral-load samples, while serology is limited by the window period between infection and detectable antibody production. Nevertheless, when carefully designed and interpreted in conjunction with complementary techniques such as metagenomic sequencing, immunoassays provide a rapid, cost-effective, and scalable approach for identifying novel viruses and monitoring emerging infectious threats.

4.5 DNA microarray

DNA microarray technology emerged as an advanced diagnostic tool for infectious diseases, designed to enable the simultaneous and specific detection of a wide range of pathogens (Asmare and Erkihun, 2023). The principle of detection relies on solid-phase hybridization, where pathogen-specific oligonucleotide probes are immobilized on a solid surface and hybridize with complementary sequences from a mixture of fluorescently labeled nucleic acids (Martínez et al., 2014). Over time, diverse microarray platforms were developed to target pathogens associated with respiratory, hemorrhagic, blood-borne, and central nervous system syndromes (Martínez et al., 2014), while broader-spectrum microarrays were designed for virus discovery and surveillance (Wang et al., 2002).

This technology represented a pivotal step forward in molecular diagnostics, as it enabled the parallel screening of thousands of predefined viral sequences on a single chip through probe–target hybridization (Wang et al., 2003). Compared with single-plex PCR, microarrays offered a much wider scope of detection. However, they remained fundamentally targeted: probe design required prior knowledge of pathogen genomic sequences, restricting their ability to identify novel or highly divergent agents. Additional challenges included cross-hybridization artifacts, which could compromise specificity, and a generally lower sensitivity compared to amplification-based methods.

As a result, while DNA microarrays played an important transitional role in broad-spectrum pathogen detection, they were eventually surpassed in pathogen discovery by metagenomic next-generation sequencing. Unlike microarrays, metagenomic next-generation sequencing provides a truly unbiased approach, capable of identifying both known and previously uncharacterized pathogens, making it the focus of current and future innovations in infectious disease diagnostics.

5 Viral metagenomic next-generation sequencing

Viral metagenomics next-generation sequencing provides a fast, sensitive, and robust approach for detecting viruses, including those that remain undetectable by traditional culture techniques and sequence-dependent assays (Mokili et al., 2012). This unique capability has firmly established mNGS as a leading tool in the discovery of novel viruses. Unlike conventional diagnostic assays, which rely on prior knowledge of target sequences, mNGS employs an unbiased strategy that enables the simultaneous detection of both known and unknown viral pathogens. This makes it invaluable in situations where the causative agent of an outbreak is unknown, representing a critical first step in mounting an effective and timely outbreak response (Greninger, 2018). When both culture-based methods and advanced molecular assays fail to detect a pathogen, mNGS has often served as the ultimate diagnostic approach, leading to the identification of novel viruses (Figure 2; Roux et al., 2021).

Since the first viral genome was sequenced using metagenomic methods in 2002, the pace of virus discovery has accelerated dramatically (Mokili et al., 2012; Dutilh et al., 2017). A landmark example is the identification of SARS-CoV-2 in 2019, where mNGS enabled rapid characterization of the novel coronavirus and provided genomic data that informed early diagnostic test development, epidemiological modeling, and vaccine design. Two decades later, mNGS remains at the forefront of virology, now enhanced by the integration of artificial intelligence (AI), machine learning (ML), and deep learning tools, which increase the speed, accuracy, and interpretability of vast sequencing datasets.

The strength of mNGS lies in its flexibility and broad applicability. Unlike targeted assays such as PCR, which require specific primers, mNGS can be applied to a wide variety of sample types, including blood, respiratory swabs, stool, plant tissues, and environmental reservoirs such as wastewater, while still generating high-quality sequence data (Roux et al., 2021). This versatility is particularly significant in today's context of frequent human-animal-environment interactions, where zoonotic spillover events have led to the emergence of high-impact pathogens such as Ebola virus, mpox virus, and coronaviruses. Beyond pathogen discovery, mNGS has proven valuable in detecting co-infections, characterizing viral diversity within hosts, and monitoring viral evolution. For example, it has successfully identified viral-bacterial co-infections such as varicella zoster virus with herpes simplex virus-2 (Schuele et al., 2025). Similarly, Slavov, reported that clinically important viruses, including measles virus, SARS-CoV-2, hepatitis B virus, parvovirus B19, adenovirus, and human herpesviruses, were detected alongside commensal members of the blood virome such as anelloviruses (Slavov, 2025). These findings highlight mNGS's ability not only to detect pathogens but also to provide insights into the broader viral ecosystem associated with human health and disease.

The workflow of mNGS begins with specimen collection, which can include clinical samples (e.g., blood, cerebrospinal fluid, nasopharyngeal swabs), animal reservoirs, or environmental sources (e.g., soil and water). Viral nucleic acids (DNA and/or RNA) are then extracted, followed by random or targeted amplification and library preparation (Morgan et al., 2010). Sequencing is performed using high-throughput platforms such as Illumina, Oxford Nanopore Technologies, or Pacific Biosciences (PacBio), each offering distinct

advantages in terms of read length, throughput, and error profile. The raw sequence data undergoes comprehensive bioinformatics processing, including quality filtering, host read subtraction, *de novo* assembly, and taxonomic classification, to achieve viral identification and characterization (Slavov, 2025; Alcolea-Medina et al., 2024; Chiu and Miller, 2019).

Despite its transformative power, the application of mNGS requires specialized expertise, sophisticated equipment, significant computational resources, and robust laboratory and bioinformatics infrastructure. These requirements currently limit their widespread use in low- and middle-income countries where outbreak risks are often highest. Nevertheless, the unbiased nature of mNGS makes it uniquely suited for outbreak investigations and pathogen surveillance. Its real-world impact has been repeatedly demonstrated, such as in the rapid identification of SARS-CoV-2 in Wuhan in 2019, the genomic characterization of mpox outbreaks, and the elucidation of viral genomes during Zika and Ebola epidemics. Moving forward, continued improvements in sequencing technology, data analysis pipelines, and cost reduction, alongside integration with AI-driven analytics, will further strengthen the role of mNGS in global health security and pandemic preparedness.

5.1 Metagenomic next generation sequencing technologies

Metagenomic next-generation sequencing platforms are now widely employed not only for targeted sequencing of specific genes or genomic regions but also for comprehensive, sequence-based association analyses that drive pathogen discovery and characterization (Harismendy et al., 2009). Their growing adoption is largely fueled by the urgent need for faster, more accurate, and versatile diagnostic tools in infectious disease management. Beyond diagnostics, mNGS has opened entirely new avenues for research by allowing scientists to interrogate genetic information at an unprecedented scale and resolution, thereby advancing our understanding of microbial diversity, host-pathogen interactions, and evolutionary dynamics (Goodwin et al., 2016).

Technology's integration of high-throughput performance with steadily improving affordability has solidified its role as a cornerstone in fields spanning from fundamental biology and epidemiology to precision medicine and clinical diagnostics. While debates persist regarding their overall cost-effectiveness, particularly in low-resource settings, the practical utility of mNGS in accelerating discovery and improving diagnostic sensitivity has proven invaluable. Importantly, one of its defining advantages lies in its ability to generate vast volumes of sequence data, typically ranging from 300 to 500 cycles per run, enabling deep coverage and comprehensive genomic profiling (Kozich et al., 2013).

Nevertheless, sequencing performance varies considerably across platforms, with notable differences in read length, accuracy, throughput, and error profiles that directly impact downstream analyses and clinical applications. Some technologies are optimized for generating short, highly accurate reads, whereas others prioritize long-read sequencing, which is advantageous for genome assembly and structural variant detection. Historically, Illumina's short-read (second-generation) sequencing has dominated mNGS because of its high accuracy and throughput. In contrast, the advent of

third-generation platforms such as Oxford Nanopore Technologies (ONT) and Pacific Biosciences (PacBio) has revolutionized the field by enabling long-read sequencing, often spanning thousands of bases, with steadily improving accuracy (Han et al., 2024). While short reads (<300 bp) can lead to fragmented assemblies and may overlook structural variants, long reads provide the ability to resolve repetitive regions and capture full-length genes, thereby offering significant advantages in metagenomic applications (Han et al., 2024). Below we compare Illumina, ONT, and PacBio across key metrics for metagenomics (Table 1).

5.1.1 Illumina sequencing platforms

Illumina-based sequencing technology is one of the most widely adopted platforms in metagenomics, providing high-throughput and cost-effective sequencing of DNA and RNA (Xia et al., 2023; Schirmer et al., 2016). It has become a cornerstone for microbial community analysis, particularly for taxonomic profiling, functional annotation, and pathogen detection (Elbehiry and Abalkhail, 2025). The platform's short-read approach delivers massive sequencing depth at relatively low cost, making it ideal for large-scale studies across clinical,

environmental, and engineered systems (Xia et al., 2023). Illumina short reads enable sensitive and robust detection of microbial community composition, antimicrobial resistance genes, and metabolic pathways, provided sufficient coverage is achieved. Furthermore, metagenomic Illumina tags (*mi*Tags) can mitigate PCR amplification biases, offering more accurate estimates of microbial richness and evenness compared with traditional amplicon-based methods (Logares et al., 2014). Nonetheless, with modern instruments such as the NovaSeq 6000 and NovaSeq X series, Illumina can yield terabase-scale data in a single run, supporting large-scale environmental surveys and deep sequencing for rare pathogen detection.

Despite these strengths, Illumina sequencing introduces several challenges and biases. Position- and motif-specific errors, most commonly substitution errors linked to specific nucleotide motifs, may persist even after quality filtering and affect downstream analyses (Schirmer et al., 2016). The reliance on short reads (<300 bp) can lead to fragmented assemblies, particularly in complex communities or genomes with high sequence similarity, thereby limiting recovery of complete genomes and structural variants (González et al., 2025).

TABLE 1 Comparison of major high-throughput sequencing platforms for viral metagenomics.

Feature	Illumina (short-read)	Oxford nanopore technologies (ONT) (long-read)	Pacific biosciences (PacBio) (long-read, HiFi)
Key technology	Sequencing-by-synthesis with reversible dye-terminators (Chen et al., 2013).	Nanopore-based electronic sensing of DNA/RNA strands (Ni et al., 2019).	Single Molecule, Real-Time (SMRT) sequencing with circular consensus sequencing (CCS) (Zhang et al., 2025).
Typical read length	50–300 bp (short) (Kozich et al., 2013).	10 kb–100+ kb (ultra-long) (Moustakli et al., 2025; Shafin et al., 2020).	Long & Accurate (15–25 kb) (HiFi reads) (Wen and Tang, 2025; Travers et al., 2010; Wenger et al., 2019).
Throughput and speed	Very High throughput (Terabases/run). Run time: hours to days (Liu et al., 2021).	Scalable & Real-Time. From portable (MinION) to high throughput (PromethION). Data is available immediately. Scalable (MinION: ~30 Gb; PromethION: ~13 Tb) (Chen and Xu, 2023).	High (Revio: ~100–120 Gb per SMRT Cell) (Zhang et al., 2025).
Accuracy profile	Very High (>99.9%); substitution errors (Schirmer et al., 2016).	Moderate-High (Raw: ~99%; with Q20+: >99.9%); indel errors in homopolymers (Sereika et al., 2022; Chen et al., 2024).	Extremely High (>99.9% with HiFi mode) (Wen and Tang, 2025; Travers et al., 2010; Wenger et al., 2019).
Primary strengths	- Gold standard for cost-effective, high-depth sequencing. - Excellent for SNP/variant calling and quantitative abundance (Xia et al., 2023; Logares et al., 2014).	- Real-time analysis for immediate insights. - Extreme portability for field use. - Best for detecting structural variants and <i>de novo</i> assembly. - Direct RNA sequencing (De Coster et al., 2019).	- Unparalleled combination of long reads and high accuracy. - Ideal for resolving complex regions, haplotype phasing, and high-quality <i>de novo</i> assembly (Slavov, 2025; Alcolea-Medina et al., 2024; Chiu and Miller, 2019).
Primary limitations	- Short reads fail to resolve repeats, leading to fragmented assemblies. - Cannot phase haplotypes or detect large Structural Variants easily (González et al., 2025).	- Historically higher error rate, though improving rapidly. - Requires substantial bioinformatics for basecalling and error correction (Sereika et al., 2022; Chen et al., 2024).	- Highest per-sample cost. - Requires high-quality, high molecular weight DNA input. - Less suitable for rapid, real-time applications (Han et al., 2024).
Best suited for mNGS application	- Large-scale surveillance and detection: Sensitive pathogen identification in complex samples (e.g., microbiome, plasma). - High-resolution SNP analysis for outbreak tracing (Xia et al., 2023).	- Rapid Real-time pathogen detection and outbreak response in the field. <i>De novo</i> assembly of unknown pathogens. - Epigenetic modification detection (Marić et al., 2024).	- Gold-standard <i>de novo</i> genome assembly of novel viruses/bacteria. - Resolving complex viral communities and strain variants. - Full-length 16S/18S sequencing without amplification bias (Slavov, 2025; Alcolea-Medina et al., 2024; Chiu and Miller, 2019).

Low-abundance organisms or genes may also be difficult to detect without very deep sequencing (e.g., ~30 million reads for 1% abundance) (Rooney et al., 2022).

To overcome these limitations, hybrid strategies that combine Illumina with long-read platforms such as Oxford Nanopore Technologies (ONT) or Pacific Biosciences (PacBio) are increasingly employed (Xia et al., 2023; Sevim et al., 2019). These approaches improve assembly contiguity, genome completeness, and the resolution of structural variants, while retaining Illumina's advantage in low error rates and reliable genome recovery. Ultimately, Illumina remains foundational in metagenomics for its accuracy, throughput, and affordability, but optimal study design requires careful consideration of its error profiles and assembly constraints. In many cases, hybrid or long-read approaches provide a more comprehensive view of microbial diversity and genome structure.

5.1.2 Oxford nanopore technologies

Oxford Nanopore Technologies (ONT) has transformed metagenomics by enabling real-time, portable, and long-read sequencing of complex microbial communities (Ni et al., 2019). ONT devices employ protein nanopores embedded in membranes to directly read native DNA or RNA molecules as they translocate through the pore, generating reads of virtually unlimited length, routinely exceeding 100 kilobases (kb) and occasionally surpassing 1 megabase (Mb) (Moustakli et al., 2025; Shafin et al., 2020). This ability to generate ultra-long reads is ONT's defining feature, enabling more contiguous genome assemblies, improved detection of structural variations, and strain-level resolution in metagenomic samples. Compared with short-read platforms, ONT excels at resolving repetitive regions and reconstructing complex genomes.

ONT platforms, including the portable MinION and high-throughput PromethION, are increasingly applied across clinical diagnostics, outbreak investigations, environmental monitoring, and food safety (De Coster et al., 2019). Technology is particularly valuable for rapid pathogen detection, often providing actionable results within hours. For example, ONT sequencing has been deployed in real-time outbreak response and field-based surveillance due to its portability and minimal infrastructure requirements (Oehler et al., 2023). Output scales from ~30–50 Gb on a MinION flow cell to up to ~13 Tb on a PromethION run (48 flow cells, 72 h), making it suitable for both targeted and large-scale applications (Chen and Xu, 2023).

Despite these advantages, ONT historically faced limitations due to higher raw error rates (5–10%), predominantly indels in homopolymeric regions. Although consensus polishing with short reads was often required, recent advances, including R10.4.1 nanopores and Q20+ chemistry—have markedly improved basecalling performance, achieving raw-read accuracies exceeding 99% (Sereika et al., 2022; Chen et al., 2024). Nevertheless, homopolymer-associated errors remain a challenge, and robust bioinformatics workflows are required for error correction, host DNA contamination filtering, and metagenome-assembled genome (MAG) reconstruction. Hybrid approaches that integrate ONT with Illumina sequencing remain the gold standard for producing highly complete and accurate assemblies.

Specialized tools (e.g., Pike for OTU-level analysis) and the development of field-adapted extraction protocols continue to expand ONT's utility, offering flexible and cost-effective solutions for microbial surveillance and biodiversity studies (Krivoson et al., 2025).

Overall, ONT's key strength lies in its real-time sequencing capability, enabling rapid clinical diagnostics, novel pathogen surveillance, and metagenomic assemblies requiring ultra-long reads. However, its per-base cost remains higher than Illumina for very large-scale projects, and careful consideration of study goals is necessary when selecting ONT as a primary sequencing platform.

5.1.3 Pacific biosciences (PacBio)

Pacific Biosciences has established itself as the leader in high-fidelity long-read sequencing through its circular consensus sequencing (CCS) strategy, yielding HiFi reads that combine long-read lengths (typically 15–25 kb) with accuracies exceeding 99.9% (Q30 or higher) (Wen and Tang, 2025; Travers et al., 2010; Wenger et al., 2019). The latest Revio platform generates ~100–120 Gb per SMRT Cell, with up to four cells running in parallel, enabling hundreds of gigabases of highly accurate long-read data per day (Zhang et al., 2025). PacBio's requirement for high molecular weight DNA and relatively complex library preparation workflows pose technical challenges, but the resulting data are uniquely powerful. HiFi reads preserve single-nucleotide accuracy while spanning long genomic regions, enabling recovery of complete metagenome-assembled genomes (MAGs), structural variant resolution, and improved taxonomic resolution for rare or novel taxa. Comparative studies have shown that PacBio recovers more low-abundance lineages than Illumina or ONT, due to its combination of read length and accuracy. The high capital cost of PacBio systems and consumables may restrict adoption in resource-limited settings, but for high-resolution metagenomics, especially in research on complex microbial communities, PacBio HiFi data are considered the gold standard.

6 Current approaches to data analysis

Traditional alignment-assembly-annotation pipelines remain the backbone of viral metagenomic sequencing for outbreak investigation (Yang et al., 2024). They provide interpretable and clinically actionable results: read classification enables rapid confirmation of suspected pathogens, genome assemblies allow high-resolution phylogenetic analyses, and annotation facilitates detection of resistance or virulence markers (Song et al., 2021; Rosenboom et al., 2022). Benchmarking studies of clinical metagenomic pipelines confirm that these workflows are highly specific for known pathogens and reproducible across laboratories, supporting their integration into surveillance and public health responses (de Vries et al., 2021). The assembled genomes were central in tracking SARS-CoV-2 lineage dynamics during the COVID-19 pandemic, enabling timely insights into transmission, mutation hotspots, and global spread (Saravanan et al., 2022). The maturity of these pipelines, coupled with standardized platforms such as Nextflow and Snakemake, ensures reproducibility and traceability, critical strengths during time-sensitive outbreak responses (Langer et al., 2025).

On the other hand, recent studies confirm important limitations when applying traditional pipelines for outbreak investigation and novel pathogen discovery. Wet-lab protocol comparisons have demonstrated that enrichment via capture panels dramatically increases sensitivity in low viral load samples: for respiratory pathogens such as SARS-CoV-2 or influenza A, target capture sequencing yielded 180–2,000-fold higher viral read counts compared

to untargeted metagenomics in some clinical specimens (Takemae et al., 2024). Benchmarking of virus identification tools using real-world metagenomic datasets found that while some tools perform well under default settings, there is a trade-off between sensitivity and specificity and many tools fail to recover virus contigs when genomes are fragmented or diverged (Takemae et al., 2024). Therefore, reliance solely on traditional pipelines can delay detection of novel or low-abundance pathogens in outbreak settings, unless supplemented with optimized sample preparation, high sequencing depth, and tools designed to handle divergent sequences (Wu et al., 2024).

7 AI and machine learning applications

The rapid expansion of metagenomic sequencing has generated unprecedented volumes of complex and heterogeneous data, necessitating advanced analytical frameworks beyond conventional bioinformatics (Pita-Galeana et al., 2025). AI, particularly ML and DL, has emerged as a transformative tool for extracting biologically meaningful insights from metagenomic datasets. Its applications span the entire analytical pipeline, from raw data preprocessing to functional inference and clinical translation.

One of the central challenges in viral metagenomics is accurate classification of sequences, especially when viral genomes exhibit high mutation rates or when reference databases are incomplete. Traditional bioinformatics pipelines rely on alignment-based methods (e.g., BLAST, Bowtie) or k-mer frequency approaches (Pita-Galeana et al., 2025; Wu et al., 2021). These approaches are limited when viral sequences diverge significantly from known references.

Deep learning algorithms, particularly convolutional neural networks (CNNs) (de Souza et al., 2023), and recurrent neural networks (RNNs) (Deif et al., 2021), have been developed to overcome these limitations. By learning hierarchical sequence features directly from raw data, deep learning models can classify viral sequences with higher accuracy and generalize better to novel or divergent genomes. For instance, CNN-based models have been applied to detect viral families from short sequencing reads without requiring genome assembly (Tampuu et al., 2019).

Autoencoders and attention-based models (e.g., transformers) further extend classification performance by capturing long-range dependencies and sequence motifs relevant to viral taxonomy (Mswahili and Jeong, 2024). Importantly, these models can recognize “novelty signatures,” allowing for classification at higher taxonomic ranks when species-level resolution is not possible (Table 2). This feature is critical in outbreak investigations where the pathogen may belong to an underrepresented or previously unknown viral group.

7.1 Key architectures and applications

7.1.1 Convolutional neural networks (CNNs)

Convolutional neural networks, initially developed for computer vision, have proven highly effective in biological sequence analysis due to their ability to detect local, position-invariant patterns, a defining property of genomic data (Rives et al., 2021; Rives et al., 2021). By applying learnable filters across nucleotide or amino acid sequences, CNNs act as motif discovery engines, identifying conserved short patterns such as transcription factor binding sites or protease cleavage

motifs (Alley et al., 2019; Jumper et al., 2021; Senior et al., 2020). Sequences are typically represented numerically through one-hot encoding, where nucleotides or amino acids are mapped into binary vector space, forming input matrices analogous to images (Jumper et al., 2021). Convolutional and pooling layers then generate feature maps that summarize motif occurrence and position, conferring robustness to sequence variability (Unsal et al., 2022; Bileschi et al., 2022).

This hierarchical feature extraction enables CNNs to detect both simple motifs and complex higher-order structures, such as protein domains or viral polymerases (Hie et al., 2021; Ming et al., 2023; Lee, 2023). Consequently, CNNs can classify viral from non-viral sequences, annotate regulatory regions, and identify taxonomic signatures, even in cases of low sequence similarity to reference genomes (104). This capability to detect motifs without high-sequence homology is particularly advantageous over alignment-based methods like BLAST during the investigation of a novel outbreak. For instance, a CNN can enable family-level classification of an unknown virus based on conserved polymerase motifs, providing a crucial first clue for public health responders within hours, even when the virus shares less than 50% sequence similarity to any known reference. Tools like DeepVirFinder leverage CNN architectures to uncover viral sequences in metagenomic assemblies by integrating k-mer compositions with contextual genomic signals, outperforming alignment-based methods in novel virus discovery (Ren et al., 2020; Ren et al., 2020). As sensitive pattern detectors, CNNs thus provide a scalable solution for viral genome annotation and pathogen discovery in metagenomics.

7.1.2 Recurrent neural networks (RNNs)

Recurrent Neural Networks are designed specifically for sequential data and thus provide a natural framework for nucleotide and amino acid sequence analysis (Chandra et al., 2023). Unlike CNNs, which specialize in local motif detection, RNNs capture dependencies across positions by processing sequences element-by-element while maintaining a hidden state that reflects prior context (Graves, 2012). This allows the model to incorporate the meaning of a nucleotide or codon in relation to its surrounding sequence, which is critical for recognizing reading frames, splice sites, and regulatory elements spanning long genomic distances (Rumelhart et al., 1986; Auslander et al., 2021).

However, conventional RNNs are hindered by the vanishing gradient problem, limiting their ability to learn long-range dependencies (Bengio et al., 1994). Despite this, they remain valuable for tasks requiring short- to medium-range contextual modeling, including identifying conserved sequence patterns, functional annotation of genes, and detecting short-range evolutionary constraints.

7.1.3 Long short-term memory networks (LSTMs)

Long short-term memory networks extend RNNs by incorporating gating mechanisms, input, forget, and output gates, that regulate the retention and flow of information (Gers et al., 2000). This architecture mitigates vanishing gradients, enabling learning across long genomic regions where distant interactions carry biological significance.

In viral metagenomics, LSTMs are particularly effective in modeling genome-wide signals such as codon usage bias, oligonucleotide frequencies, and co-evolutionary patterns between

TABLE 2 Key artificial intelligence architectures and their applications in viral metagenomics.

Architecture	Core mechanism	Key strengths in viral metagenomics	Primary applications and examples
Convolutional neural networks (CNNs)	Applies learnable filters to detect local, position-invariant patterns (Unsal et al., 2022; Bileschi et al., 2022).	Excellent motif discovery; robust to sequence variability; does not require high sequence homology (Ren et al., 2020; Ren et al., 2020).	- Viral sequence classification: Distinguishing viral from host sequences (e.g., DeepVirFinder) (Ren et al., 2020; Ren et al., 2020). - Family-level identification: Classifying unknown viruses based on conserved protein motifs (e.g., polymerase) (Mswahili and Jeong, 2024).
Recurrent neural networks (RNNs)/ Long short-term memory (LSTMs)	Processes sequences step-by-step, maintaining a “memory” of previous context via hidden states (Nayfach et al., 2021).	Models short-to-long-range dependencies; captures sequential context and temporal dynamics (Bengio et al., 1994).	- Modeling viral evolution: Tracking intra-host evolution and quasi-species dynamics during prolonged outbreaks (Singh et al., 2016). - Functional annotation: Identifying gene boundaries (start/stop codons) and regulatory elements (Pasolli et al., 2019).
Transformers and attention mechanisms	Uses self-attention to weigh the importance of all positions in a sequence simultaneously (Bigness et al., 2022; Choi and Lee, 2023).	Captures long-range, global dependencies efficiently; enables parallel processing; highly adaptable via pre-training (Bigness et al., 2022; Choi and Lee, 2023).	- De novo genome assembly: Improved assembly of novel viral genomes from complex metagenomic data (Slavov, 2025; Alcolea-Medina et al., 2024; Chiu and Miller, 2019). - Host prediction and pathogenicity: Learning generalizable representations for tasks like host prediction (e.g., ViralBERT, PathogenTransformer) (Abramoff et al., 2023; Vashisht et al., 2023; Choi and Lee, 2023).
Graph neural networks (GNNs)	Operates on graph structures where nodes (e.g., individuals, variants) are connected by edges (e.g., transmissions) (Bileschi et al., 2022; Liu et al., 2020).	Models relational and network data; integrates heterogeneous data types (genomic, mobility, contact) (Nasir et al., 2023).	- Transmission dynamics: Reconstructing transmission chains and identifying superspreader events (Choi et al., 2022). - Variant spread tracking: Integrating genomic and mobility data to project geographic spread of variants (Choi et al., 2022).
Autoencoders (for anomaly detection)	Compresses input data into a latent space and reconstructs it; high reconstruction error indicates anomalies (Marić et al., 2024; Zhao et al., 2023).	Unsupervised learning; does not require labeled data for novel pathogens; identifies deviations from known sequences (Barredo Arrieta et al., 2020).	- Novel pathogen discovery: Flagging unclassified or highly divergent genomic fragments as potential novel viruses (Barredo Arrieta et al., 2020). - Outbreak early warning: Detecting anomalous sequence clusters in surveillance data (Kuo and Ying, 2023).

viruses and hosts (Nayfach et al., 2021). They also excel at identifying start/stop codons, splice sites, and functional domains separated by introns or long intergenic regions (Pasolli et al., 2019). This is critical for modeling viral evolution within a host during a prolonged outbreak, such as an Ebola or SARS-CoV-2 epidemic, allowing researchers to track the emergence of quasi-species that may impact transmission or treatment efficacy. For phylogenetic classification, LSTMs capture patterns of mutation and conservation across entire genomes, producing robust evolutionary inferences beyond local homology (Singh et al., 2016). Although newer architectures such as Transformers (Tampuu et al., 2019), offer advantages in handling very long sequences with parallelization, LSTMs remain widely used due

to their strong performance in tasks where sequential order and contextual dependencies are biologically essential (Jurtz et al., 2017).

7.2 Transformers and attention mechanisms

Transformers mark a major advancement in sequence analysis, diverging from convolutional and recurrent architectures through their core innovation: the self-attention mechanism (Choi et al., 2022). Unlike CNNs, which emphasize local motifs, or RNNs, which process sequences sequentially, transformers assign weights to all positions

simultaneously, enabling direct modeling of long-range dependencies (Bigness et al., 2022; Choi and Lee, 2023). This “global receptive field” allows the model to capture distant but functionally linked genomic features, such as promoter–coding region interactions or co-evolutionary signals across protein domains (Marić et al., 2024; Zhao et al., 2023).

The multi-head attention framework further enhances representational capacity by attending to different subspaces in parallel, capturing both syntactic (e.g., reading frames) and semantic (e.g., protein domain function) dimensions of genomic sequences (Marić et al., 2024). By modeling entire genomes holistically, transformers facilitate more accurate *de novo* assembly of novel viral genomes from complex metagenomic samples, directly addressing the challenge of fragmented assemblies that can delay the development of confirmatory PCR tests. Pre-trained models such as ViralBERT and PathogenTransformer leverage massive viral sequence corpora to learn generalizable representations that can be fine-tuned for specific tasks, including host prediction, gene annotation, and pathogenicity classification (Abramoff et al., 2023; Vashisht et al., 2023; Choi and Lee, 2023). More recent architectures, such as MetaViT, extend transformer applications to metagenomics, effectively identifying novel viral sequences by recognizing global genomic signatures beyond local homology (Ji et al., 2021). By modeling genomes holistically, transformers advance viral genomics toward context-aware interpretation and accelerate novel virus discovery.

7.3 AI in anomaly and outlier detection

Early identification of novel pathogens is a critical application of AI in outbreak surveillance (Jurtz et al., 2017). Traditional approaches, which rely on known genetic signatures or symptom patterns, are limited in detecting truly novel threats (Tisza and Buck, 2021). In contrast, unsupervised and semi-supervised learning models excel at anomaly detection by learning baseline distributions of genomic or clinical data and flagging deviations (Barredo Arrieta et al., 2020). Within metagenomic sequencing datasets, algorithms such as isolation forests and autoencoders can detect unclassified genomic fragments as potential novel viruses (Kuo and Ying, 2023). For example, high reconstruction error in autoencoders indicates sequences that diverge from known distributions, serving as a quantifiable anomaly score (Marić et al., 2024; Zhao et al., 2023).

Beyond genomics, AI pipelines integrate clinical and epidemiological metadata, such as geographic location, travel history, and symptom onset, with sequence anomalies to detect clusters of unexplained infections (Edwards et al., 2016). This proactive detection framework can generate early outbreak warnings well before traditional confirmation methods, potentially reducing response delays by weeks (Willmington et al., 2022; Rudin, 2019).

7.4 Predictive models for transmission dynamics

Following pathogen identification, AI-driven models enhance prediction of transmission dynamics, informing timely public health interventions (Lee et al., 2023). Traditional SEIR (Susceptible–Exposed–Infectious–Recovered) frameworks provide a foundation but

rely on static parameters (Auslander et al., 2021; Cheohen et al., 2025). AI augments these models by integrating real-time data and adapting parameters dynamically. Time-series methods, particularly LSTMs, can incorporate case counts, human mobility, climate data, and social media signals to forecast short-term epidemic trends with greater accuracy (Choi and Lee, 2023; Kuo and Ying, 2023).

Graph neural networks (GNNs) extend predictive power by modeling transmission chains, representing individuals or communities as nodes and their interactions as edges (Bileschi et al., 2022; Liu et al., 2020). Such models can identify superspreader events, transmission hubs, and potential intervention points. Moreover, by incorporating genomic data, GNNs can track pathogen evolution alongside mobility-driven spread, enabling projections of both geographic expansion and variant dominance (Nasir et al., 2023). These integrative models support resource prioritization and targeted containment strategies, bridging epidemiological forecasting with genomic surveillance.

7.5 Case studies

7.5.1 SARS-CoV-2

The COVID-19 pandemic served as a large-scale proving ground for AI in virology (Rives et al., 2021; Shkoporov et al., 2022). Deep learning models, most notably AlphaFold2, accurately predicted the 3D structure of the SARS-CoV-2 spike protein, which significantly accelerated rational vaccine design and therapeutic development (Jumper et al., 2021). AI-driven genomic surveillance systems played a crucial role in monitoring viral evolution by classifying variants of concern (e.g., Alpha, Delta, Omicron) through detection of mutational signatures associated with increased transmissibility, pathogenicity, and immune escape. In parallel, natural language processing (NLP) tools enhanced real-time situational awareness by rapidly scanning global research articles, and news outlets to identify and synthesize emerging scientific insights (Jumper et al., 2021; Ren et al., 2020; Capponi et al., 2021). Collectively, these advances highlighted the transformative role of AI in outbreak response and set the stage for its broader integration into future pandemic preparedness strategies.

7.5.2 Ebola virus

During the 2018–2020 Kivu outbreak in the Democratic Republic of the Congo, AI facilitated predictive risk mapping by integrating satellite imagery, climate data, and animal habitat distributions to identify spillover hotspots (Willmington et al., 2022; Pigott et al., 2014). Machine learning models were employed to differentiate between local transmission chains and novel viral introductions, thereby informing containment strategies and resource allocation. Furthermore, AI-driven phylodynamic frameworks provided critical insights into the evolutionary dynamics and geographic spread of the virus. Complementarily, network-based analyses of contact-tracing data identified key transmission pathways, which guided targeted vaccination campaigns in resource-constrained and conflict-affected settings.

7.5.3 Mpox

The 2022 global Mpox outbreak highlighted AI's potential in detecting atypical transmission. Machine learning analyses of genomic data confirmed sustained human-to-human spread and revealed

hidden transmission chains beyond endemic regions (Gigante et al., 2022). Models integrating air travel and case data further predicted high-risk cities for importations, supporting proactive surveillance and public health messaging. In clinical diagnostics, CNN-based models distinguished Mpox from other skin lesions with accuracies ranging from 78 to 98.8% across multiple datasets and architectures (Chadaga et al., 2023), underscoring AI's potential for rapid and reliable Mpox detection.

7.5.4 Influenza

AI applications in influenza span routine forecasting and pandemic preparedness. In seasonal surveillance, U.S. CDC forecasts are augmented with models incorporating viral genomics, search engine data, and historical trends to predict epidemic timing and intensity (Reich et al., 2019). For pandemic risk assessment, AI evaluates avian influenza strains (e.g., H5N1, H7N9), predicting traits such as receptor binding specificity and antigenic drift to inform pre-pandemic vaccine libraries (Lou et al., 2024).

7.6 Advantages over traditional methods

The integration of AI into virology and epidemiology provides substantial advantages over conventional approaches, particularly in speed, scalability, and predictive power.

7.6.1 Speed and automation

Traditional sequence analyses, such as BLAST searches and phylogenetic reconstructions, are computationally intensive and require manual curation. In contrast, trained AI models can process millions of sequences in hours, enabling real-time surveillance. Automated pipelines convert raw sequencing reads into variant calls and lineage assignments with minimal human intervention, allowing experts to focus on interpretation rather than data processing (Lee, 2023).

7.6.2 Handling high-dimensional data

AI excels at integrating diverse datasets, including genomic sequences, protein structures, clinical outcomes, mobility patterns, and environmental variables, revealing complex, non-linear relationships that traditional methods cannot capture. Whereas logistic regression may detect a few predictors, machine learning models such as random forests or neural networks can uncover intricate interactions to predict patient severity or outbreak hotspots (Soenksen et al., 2022).

7.6.3 Discovery of novel patterns

Unlike hypothesis-driven methods, AI can identify previously unrecognized patterns via unsupervised learning. This has enabled the discovery of novel CRISPR systems and microbial defense mechanisms (Doron et al., 2018). In virology, such approaches facilitate the detection of novel viral families and unconventional pathogenic mechanisms that might be overlooked by conventional analyses.

7.6.4 Predictive accuracy and adaptability

Classical compartmental models, such as SEIR, rely on fixed parameters. AI-enhanced models continuously assimilate new data, adapting forecasts as outbreaks evolve. This adaptability improves

short-term predictive accuracy, as demonstrated by ensemble models that consistently outperformed traditional methods during the COVID-19 pandemic (Cramer et al., 2022).

8 Challenges and proposed solutions

The integration of AI into viral metagenomics offers transformative potential for outbreak response, yet its journey toward widespread, reliable, and equitable implementation is constrained by significant challenges. Critically examining these limitations is not intended to diminish the technology's promise, but rather to provide a roadmap for guiding its continued evolution. This section highlights the central barriers, ranging from data availability and quality to model interpretability, to infrastructural and resource constraints—and aligns them with emerging research directions and technological innovations that seek to address these gaps.

8.1 Data scarcity and labeling bottlenecks

A foundational challenge is the “data requirements” of deep learning models, which require vast quantities of high-quality, accurately labeled sequences for training (Ren et al., 2020; Greener et al., 2022). The performance and generalizability of models are directly correlated with the volume and quality of their training data. However, the ground truth in virology is often elusive; labeling sequences as “viral” or assigning taxonomy requires slow, manual experimental validation or high-confidence homology, creating a fundamental data bottleneck (Edwards et al., 2016; Roux et al., 2019). As a result, viral sequence datasets remain orders of magnitude smaller than those used to train foundational models in other domains. Researchers often resort to data augmentation, semi-supervised learning, or transfer learning to partially overcome these limitations (Li et al., 2021; Santiago-Rodriguez and Hollister, 2022). While useful, these approaches are ultimately stopgap solutions; they cannot fully replace large-scale, high-fidelity, experimentally validated data.

More pernicious than sheer quantity is the profound taxonomic bias embedded within existing genomic databases. Public repositories like GenBank and RefSeq are overwhelmingly skewed toward viruses of established clinical and agricultural importance (e.g., influenza, HIV, SARS-CoV-2) (Nayfach et al., 2021; Tisza and Buck, 2021; Shkoporov et al., 2022; Schulz et al., 2020). In contrast, viruses from environmental niches, extreme ecosystems, and non-model organisms are severely underrepresented, creating a vast “viral dark matter” (Li et al., 2021; Santiago-Rodriguez and Hollister, 2022). This imbalance creates a “long-tail” distribution problem where DL models become highly accurate at recognizing common human pathogens but fail to identify novel or underrepresented viral families from under-sampled ecosystems, potentially delaying the response to a novel zoonotic spillover event (Nazer et al., 2023; Rampelli et al., 2020).

8.1.1 Emerging solutions and research directions

A multi-faceted approach is being developed to combat data limitations:

- 1 Transfer learning and pre-trained models: Researchers are increasingly leveraging models pre-trained on massive,

general-purpose protein or nucleotide sequence databases (e.g., models inspired by AlphaFold, DNABERT) (Jumper et al., 2021; Capponi et al., 2021). These models learn fundamental biological “grammar” and can be fine-tuned for specific viral classification tasks with much smaller, viral-specific datasets, thereby reducing the burden of data scarcity (Camargo et al., 2023).

- 2 Data augmentation and few-shot learning: Advanced techniques are being employed to artificially expand training datasets by generating realistic synthetic viral sequences (Ji et al., 2021). Furthermore, “few-shot learning” algorithms are being designed to learn effectively from a very small number of examples, which is critical for rare or novel viral families.
- 3 Global sequencing initiatives: Concerted efforts to systematically sequence diverse environments (e.g., the Global Virome Project, Earth Virome) are crucial for populating databases with novel viral sequences, thereby gradually correcting taxonomic biases and providing a more representative ground truth for model training (Rampelli et al., 2020).

8.2 The black box problem: interpretability and explainable AI

The predictive power of deep learning models is often tempered by their lack of interpretability, rendering them as inscrutable “black boxes” (Rudin, 2019; Greener et al., 2022). Although models may achieve high accuracy in distinguishing viral from host sequences, the underlying basis of their predictions often remains opaque. Identifying which specific nucleotides, motifs, or genomic structures drive a given decision is a persistent challenge. This lack of interpretability poses a critical barrier for virologists and public health officials, who need not only accurate classifications but also biologically meaningful and actionable insights to guide experimental validation and inform public health interventions (Barredo Arrieta et al., 2020; Samek et al., 2017).

8.2.1 Emerging solutions and research directions

The field of Explainable AI (XAI) is becoming indispensable for building trust and transforming predictions into scientific discovery.

- 1 Saliency maps and gradient-based techniques: Methods like Grad-CAM can highlight the nucleotides in an input sequence that most strongly influence the model’s output, creating a “heatmap” of importance across the genome (Singh et al., 2016). For instance, when a model classifies a sequence as a coronavirus, a saliency map might pinpoint the receptor-binding domain, providing immediate, biologically plausible validation.
- 2 Feature attribution methods: Frameworks like SHAP (SHapley Additive exPlanations) quantify the contribution of each input feature to the final prediction (Lundberg and Lee, 2017). In viral host prediction, SHAP can reveal if a model is relying on codon usage bias or specific promoter sequences, thereby uncovering genomic signatures of co-evolution and generating testable hypotheses (Auslander et al., 2021).
- 3 XAI for discovery: Crucially, XAI extends beyond model debugging to enable biological insight. For instance, if an XAI

model consistently highlights a non-structural protein gene in novel viruses associated with severe disease, it could point toward a previously uncharacterized virulence factor, guiding subsequent experimental research (Li et al., 2021; Choi et al., 2022). The development of domain-specific XAI tools is a critical research frontier for making AI a collaborative partner in virology.

8.3 Generalization and computational barriers

The development and training of state-of-the-art AI models require substantial GPU power and memory, creating a high financial and infrastructural barrier to entry for many academic and public health laboratories (Slavov, 2025). This is especially problematic in resource-limited settings where outbreak risks are often highest, threatening to create a new “AI divide” in global health security.

Furthermore, models trained on data from specific environments (e.g., human respiratory samples) often suffer from poor generalizability when applied to new contexts (e.g., seawater or animal vectors), a phenomenon known as overfitting (Li et al., 2023). A model that excels at identifying respiratory viruses in Illumina data from a US hospital may perform poorly on Nanopore data from bat samples in Southeast Asia, limiting its utility for proactive surveillance at the human-animal interface.

8.3.1 Emerging solutions and research directions

Innovations in computational infrastructure and model design are beginning to mitigate these barriers:

- 1 Cloud-based platforms and pre-trained models: The growth of cloud computing allows researchers to access high-performance computing on demand (Le Piane et al., 2024). More importantly, the sharing of pre-trained models means that end-users can fine-tune existing powerful models for their specific tasks, bypassing the immense cost of training from scratch.
- 2 Federated learning: This distributed approach enables AI models to be trained collaboratively across multiple decentralized datasets (e.g., from hospitals in different countries) without the raw data ever leaving its local environment (Nazer et al., 2023); (Yurdem et al., 2024). This preserves data privacy and sovereignty while allowing for the creation of more robust and generalizable models from diverse data sources, directly addressing the generalizability challenge.
- 3 Model optimization and lightweight architectures: Active research into model compression, quantization, and the development of more efficient neural network architectures aims to create powerful yet lean models that can be deployed on less powerful hardware, including at the point-of-care with portable sequencers (Dantas et al., 2024).

8.4 Ethical considerations and governance

Beyond technical and computational barriers, the deployment of AI-powered metagenomics raises profound ethical questions that

must be addressed to ensure responsible and equitable use. The capacity to sequence and analyze genetic material from any environment or patient with unprecedented speed creates new ethical dilemmas surrounding data privacy, bias, and the potential for unintended societal harm (Martinez-Martin and Magnus, 2019; Johnson et al., 2025). As NGS becomes increasingly integrated into clinical practice, the development of comprehensive, standardized regulations will be essential to effectively address its associated ethical challenges.

A primary concern is data privacy and consent. Clinical mNGS often sequences all nucleic acids in a sample, including the human host genome (Elbehiry and Abalkhail, 2025). This raises critical questions about patient autonomy and informed consent, as it is impossible to predict all pathogens that might be found. Furthermore, the integration of genomic data with clinical and mobility information in AI models creates rich datasets that are potentially re-identifiable, posing significant privacy risks if breached or misused.

Secondly, algorithmic bias and equity are major concerns (Joseph, 2025). As discussed in Section 6.1, models trained on biased data will perpetuate and potentially amplify these biases in their predictions. This can lead to systemic blind spots where pathogens circulating in under-surveilled regions are not detected, or where diagnostic AI tools perform poorly for certain populations. This could exacerbate global health inequities, directing resources and attention away from the most vulnerable communities.

Finally, the rapid identification of a novel pathogen with pandemic potential triggers complex questions about data sharing and dual-use risk. While rapid, open data sharing is crucial for a coordinated global response, it also creates a tension with national security and the risk of “dual use” research, where the same genomic information used to develop vaccines and diagnostics could theoretically be misused (Flores-Coronado et al., 2025). Establishing norms for the responsible communication of high-consequence findings to public health authorities without causing undue panic or stigma is a critical challenge.

8.4.1 Emerging solutions and governance frameworks

Developing robust ethical and governance structures is as important as advancing the technology itself.

- 1 Strengthening consent frameworks: Moving toward dynamic or tiered consent models for metagenomic testing, alongside the development and use of robust data anonymization and secure, federated learning techniques, can help protect individual privacy (Elbehiry and Abalkhail, 2025).
- 2 Bias audits and equity-focused design: Implementing mandatory algorithmic bias audits and actively promoting the sequencing of diverse viromes are essential to build fair and representative models. The “fairness” of AI models must be a key performance metric alongside accuracy (Chen et al., 2023).
- 3 International governance and policy: The establishment of clear international guidelines and agreements on the timely sharing of pathogen genomic data, coupled with frameworks to manage dual-use concerns, is urgently needed. Organizations like the WHO are pivotal in facilitating this dialog to ensure that these powerful tools serve global public health interests equitably and responsibly (Johnson et al., 2025).

Despite these challenges, the integration of deep learning, especially transformer-based models, into viral metagenomics is reshaping the field. By moving from reliance on known references toward data-driven discovery, these approaches are essential for both rapid outbreak characterization and systematic exploration of the global virosphere. Overcoming data scarcity, reducing bias, improving interpretability, ensuring generalizability and ethical consideration will be critical for unlocking their full potential.

By confronting these challenges with the outlined strategies, the field is moving steadily toward the development of robust, interpretable, and globally accessible AI-powered metagenomic tools. The goal is not to create perfect models, but to build resilient systems where the combined strengths of mNGS and AI can be reliably leveraged in the high-stakes, time-sensitive environment of an emerging outbreak.

9 Future perspectives

The trajectory of AI and viral metagenomics points toward a fundamental shift in how we monitor, detect, and respond to infectious disease threats. The future lies not merely in refining individual technologies, but in their deep integration into proactive, intelligent, and equitable global health systems. This section outlines several concrete paradigms and key milestones that will define the next decade of outbreak response.

The concept of an “AI-first outbreak response” heralds a paradigm shift wherein AI technologies transition from being supportive tools to leading players in epidemiological investigations and response strategies (Kaur and Butt, 2025). Traditionally, outbreak management has been largely human-driven, relying heavily on expert input for hypothesis generation, prioritizing laboratory testing, and coordinating contact tracing (Kaur and Butt, 2025). In contrast, AI-based systems can now autonomously generate hypotheses about pathogen origins and transmission pathways by integrating sequencing data with epidemiological and mobility information (Ye et al., 2025). They can also accelerate pathogen detection through rapid sequence classification and predict transmission hotspots for targeted interventions (Srivastava et al., 2025). By adopting an AI-first approach, public health systems can become significantly more proactive, adaptive, and scalable, often anticipating outbreaks before they fully manifest (CR et al., 2023). This strategy stands to drastically reduce the delay between pathogen identification and public health action, mitigating outbreak impacts and saving lives on a global scale (Villanueva-Miranda et al., 2025). As AI capabilities continue to advance, this proactive framework is poised to become a cornerstone of modern infectious disease control (Kaur and Butt, 2025).

The integration of AI with mNGS is poised to transform outbreak response from a largely reactive process into a proactive, predictive, and globally coordinated system. Several developments will shape this future trajectory. First, the concept of a “Digital Immune System” envisions an AI-driven global surveillance network capable of continuously analyzing metagenomic, clinical, environmental, and mobility data streams (Afshinnekoo et al., 2015). Such systems would detect anomalies and novel genomic signatures with sufficient accuracy to trigger automated early warnings, potentially identifying outbreaks weeks before traditional reporting. Future developments should focus on

creating AI systems that seamlessly mesh with existing public health surveillance infrastructures such as hospital reporting networks and environmental monitoring stations allowing for continuous, automated data ingestion and pathogen detection (Alwakeel, 2025; Kaur and Butt, 2025). This synergy will support the development of an early warning ecosystem capable of flagging viral emergence or mutations in real time, thereby allowing proactive measures to prevent large-scale spread (Villanueva-Miranda et al., 2025).

Second, advances in point-of-care metagenomics will enable rapid, field-deployable sequencing workflows, achieving “sample-to-answer” diagnostics in under 2 h (de Olazarra and Wang, 2023). Breakthroughs in portable hardware, lightweight AI algorithms, and curated “reference-on-a-chip” databases will make real-time sequencing feasible even in remote or resource-limited settings.

Third, federated learning frameworks will address data sovereignty and privacy concerns by enabling collaborative training of AI models without transferring raw genomic data across borders (Yurdem et al., 2024). This will foster equity, reduce taxonomic bias, and ensure that models remain generalizable across diverse populations and geographic regions. Cloud-based platforms combined with federated learning frameworks offer an innovative solution to the long-standing challenge of data sharing in pathogen surveillance (Aswini et al., 2025). Conventional centralized repositories often raise issues of privacy, sovereignty, and security, which discourage laboratories and countries from freely exchanging sensitive genomic data (Chourasia et al., 2024). Federated learning circumvents this by enabling AI models to be trained collaboratively across multiple decentralized datasets held by different organizations or countries, without the raw data ever leaving their local environments (Zwiers et al., 2024). This preserves patient confidentiality and national data ownership while harnessing the breadth of diverse, globally distributed datasets to produce more robust and generalizable AI models. Importantly, fostering such international collaboration is vital for pandemic preparedness and the detection of novel pathogens, as it enhances transparency, broadens surveillance capacity, and accelerates coordinated global responses (Zwiers et al., 2024; Calvino et al., 2024).

Fourth, the evolution of XAI will move AI beyond black-box predictions toward interpretable outputs that highlight key genomic features driving classification (Ali et al., 2023). This capability will enhance trust, accelerate biological discovery, and support evidence-based decision-making by public health authorities. The adoption of explainable AI will be critical for ensuring broad trust in AI-powered outbreak analytics among public health officials, researchers, and policymakers. Although AI can uncover intricate patterns and generate predictions from high-dimensional metagenomic data, its “black-box” nature risks undermining confidence if the reasoning behind outputs remains unclear (Giuste et al., 2023). To bridge this gap, explainability frameworks tailored for viral genomics and metagenomics will need to be developed, capable of providing domain-relevant insights into how models prioritize mutations, classify sequences, and generate risk assessments (Yagin et al., 2023). Such transparency not only enables experts to validate AI findings but also supports effective communication with non-specialist stakeholders, thereby strengthening decision-making and public

understanding (Msomi et al., 2025; Abe et al., 2023). As these explanation tools evolve, they will empower epidemiologists to better interpret AI-driven alerts, identify false positives or novel biology, and ultimately improve the overall accuracy and credibility of outbreak investigations.

Together, these innovations point toward an AI-first outbreak response paradigm, where intelligent systems autonomously analyze data, generate hypotheses, predict transmission dynamics, and guide interventions, while human experts provide oversight and strategy. By making surveillance faster, more interpretable, and globally inclusive, AI-powered mNGS could become the cornerstone of a resilient defense against future pandemics.

10 Conclusion

The integration of artificial intelligence with viral metagenomics marks a paradigm shift in outbreak response, moving us from reactive diagnostics to proactive pandemic preparedness. AI directly addresses the core bottleneck of mNGS—data complexity—by enabling rapid pathogen identification, novel virus discovery beyond traditional methods, and predictive modeling of outbreaks. While challenges of data scarcity, model interpretability, and equitable access remain, emerging solutions like explainable AI and federated learning provide a clear path forward. This powerful synergy is forging a new “AI-first” frontier in global health, paving the way for intelligent surveillance systems capable of defending against future viral threats.

Author contributions

DC: Conceptualization, Project administration, Writing – original draft, Writing – review & editing. EL: Investigation, Writing – original draft, Writing – review & editing. JN: Investigation, Writing – original draft. PM: Data curation, Validation, Visualization, Writing – review & editing. MC: Data curation, Supervision, Validation, Visualization, Writing – review & editing.

Funding

The author(s) declared that financial support was not received for this work and/or its publication.

Conflict of interest

The author(s) declared that this work was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The author(s) declared that Generative AI was used in the creation of this manuscript. This manuscript was written by the author(s), and all scientific content, concepts, and interpretations are entirely original

and human generated. AI-based tools were used solely for language proofreading and rephrasing to enhance clarity and readability.

Any alternative text (alt text) provided alongside figures in this article has been generated by Frontiers with the support of artificial intelligence and reasonable efforts have been made to ensure accuracy, including review by the authors wherever possible. If you identify any issues, please contact us.

References

- Abat, C., Chaudet, H., Rolain, J.-M., Colson, P., and Raoult, D. (2016). Traditional and syndromic surveillance of infectious diseases and pathogens. *Int. J. Infect. Dis.* 48, 22–28. doi: 10.1016/j.ijid.2016.04.021
- Abe, S., Tago, S., Yokoyama, K., Ogawa, M., Takei, T., Imoto, S., et al. (2023). Explainable AI for estimating pathogenicity of genetic variants using large-scale knowledge graphs. *Cancers* 15:1118. doi: 10.3390/cancers15041118
- Abrahamoff, M. D., Tarver, M. E., Loyo-Berrios, N., Trujillo, S., Char, D., Obermeyer, Z., et al. (2023). Considerations for addressing bias in artificial intelligence for health equity. *NPJ Digit. Med.* 6:170. doi: 10.1038/s41746-023-00913-9
- Afshinnekoo, E., Meydan, C., Chowdhury, S., Jaroudi, D., Boyer, C., Bernstein, N., et al. (2015). Geospatial resolution of human and bacterial diversity with City-scale metagenomics. *Cell Systems* 1, 72–87. doi: 10.1016/j.cels.2015.01.001
- Alcolea-Medina, A., Alder, C., Snell, L. B., Charalampous, T., Aydin, A., Nebbia, G., et al. (2024). Unified metagenomic method for rapid detection of microorganisms in clinical samples. *Commun. Med.* 4:135. doi: 10.1038/s43856-024-00554-3
- Ali, S., Abuhmed, T., El-Sappagh, S., Muhammad, K., Alonso-Moral, J. M., Confalonieri, R., et al. (2023). Explainable artificial intelligence (XAI): what we know and what is left to attain trustworthy artificial intelligence. *Inf. Fusion* 99:101805. doi: 10.1016/j.inffus.2023.101805
- Alley, E. C., Khimulya, G., Biswas, S., AlQuraishi, M., and Church, G. M. (2019). Unified rational protein engineering with sequence-based deep representation learning. *Nat. Methods* 16, 1315–1322. doi: 10.1038/s41592-019-0598-1
- Alwakeel, M. M. (2025). AI-assisted real-time monitoring of infectious diseases in urban areas. *Mathematics* 13:1911. doi: 10.3390/math13121911
- Apollon, W., Kamaraj, S.-K., Vidales-Contreras, J. A., Rodríguez-Fuentes, H., Flores-Breceda, H., Arredondo-Valdez, J., et al. (2022). “A beginner’s guide to different types of microscopes” in *Microscopic techniques for the non-expert*. eds. S.-K. Kamaraj, A. Thirumurugan, S. S. Dhanabalan and S. A. Hevia (Cham: Springer International Publishing), 1–23.
- Asmare, Z., and Erkihun, M. (2023). Recent application of DNA microarray techniques to diagnose infectious disease. *PLMI* 15, 77–82. doi: 10.2147/PLMI.S424275
- Aswini, R., Saranya, B., Gayathri, K., and Karthikeyan, E. (2025). Revolutionizing infectious disease surveillance: multi-omics technologies and AI-driven integration. *Decoding Infect Trans* 3:100061. doi: 10.1016/j.dcit.2025.100061
- Auslander, N., Gussow, A. B., and Koonin, E. V. (2021). Incorporating machine learning into established bioinformatics frameworks. *Int. J. Mol. Sci.* 22:2903. doi: 10.3390/ijms22062903
- Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bannetot, A., Tabik, S., Barbedo, A., et al. (2020). Explainable artificial intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI. *Inf. Fusion* 58, 82–115. doi: 10.1016/j.inffus.2019.12.012
- Bavinger, J. C., Shantha, J. G., and Yeh, S. (2020). Ebola, COVID-19 and emerging infectious disease: lessons learned and future preparedness. *Curr. Opin. Ophthalmol.* 31, 416–422. doi: 10.1097/ICU.0000000000000683
- Bengio, Y., Simard, P., and Frasconi, P. (1994). Learning long-term dependencies with gradient descent is difficult. *IEEE Trans. Neural Netw.* 5, 157–166. doi: 10.1109/72.279181
- Bigness, J., Loinaz, X., Patel, S., Larschan, E., and Singh, R. (2022). Integrating long-range regulatory interactions to predict gene expression using graph convolutional networks. *J. Comput. Biol.* 29, 409–424. doi: 10.1089/cmb.2021.0316
- Bileschi, M. L., Belanger, D., Bryant, D. H., Sanderson, T., Carter, B., Sculley, D., et al. (2022). Using deep learning to annotate the protein universe. *Nat. Biotechnol.* 40, 932–937. doi: 10.1038/s41587-021-01179-w
- Birkhead, M., Glass, A. J., Allan-Gould, H., Goossens, C., and Wright, C. A. (2021). Ultrastructural evidence for vertical transmission of SARS-CoV-2. *Int. J. Infect. Dis.* 111, 10–11. doi: 10.1016/j.ijid.2021.08.020
- Bullock, H. A., Goldsmith, C. S., and Miller, S. E. (2022). Detection and identification of coronaviruses in human tissues using electron microscopy. *Microsc. Res. Tech.* 85, 2740–2747. doi: 10.1002/jemt.24115
- Calvino, G., Peconi, C., Strafella, C., Trastulli, G., Megalizzi, D., Andreucci, S., et al. (2024). Federated learning: breaking down barriers in global genomic research. *Genes* 15:1650. doi: 10.3390/genes15121650
- Camargo, A. P., Nayfach, S., Chen, I.-M. A., Palaniappan, K., Ratner, A., Chu, K., et al. (2023). IMG/VR v4: an expanded database of uncultivated virus genomes within a framework of extensive functional, taxonomic, and ecological metadata. *Nucleic Acids Res.* 51, D733–D743. doi: 10.1093/nar/gkac1037
- Cantón Cruz, K. A., Durán Barrón, M. A., Morales Lozada, I. A., Mujica Sánchez, M. A., Deloya Brito, G. G., García Colín, M. d. C., et al. (2025). Detection of SARS-CoV-2 using the Abbott™ PANBIO™ COVID-19 SELF-TEST rapid test in patients seen at INER. *Biomedicine* 13:1012. doi: 10.3390/biomedicine13051012
- Capponi, S., Wang, S., Navarro, E. J., and Bianco, S. (2021). AI-driven prediction of SARS-CoV-2 variant binding trends from atomistic simulations. *Eur. Phys. J. E Soft Matter* 44:123. doi: 10.1140/epje/s10189-021-00119-5
- Chadaga, K., Prabhu, S., Sampathila, N., Nireeshwalya, S., Katta, S. S., Tan, R.-S., et al. (2023). Application of artificial intelligence techniques for monkeypox: a systematic review. *Diagnostics (Basel)* 13:824. doi: 10.3390/diagnostics13050824
- Chandra, A., Tünnemann, L., Löfstedt, T., and Gratz, R. (2023). Transformer-based deep learning for predicting protein properties in the life sciences. *eLife* 12:e82819. doi: 10.7554/eLife.82819
- Chen, Y., Clayton, E. W., Novak, L. L., Anders, S., and Malin, B. (2023). Human-Centered design to address biases in artificial intelligence. *J. Med. Internet Res.* 25:e43251. doi: 10.2196/43251
- Chen, F., Dong, M., Ge, M., Zhu, L., Ren, L., Liu, G., et al. (2013). The history and advances of reversible terminators used in new generations of sequencing technology. *Genomics Proteomics Bioinformatics* 11, 34–40. doi: 10.1016/j.gpb.2013.01.003
- Chen, Z., Grim, C. J., Ramachandran, P., and Meng, J. (2024). Advancing metagenome-assembled genome-based pathogen identification: unraveling the power of long-read assembly algorithms in Oxford nanopore sequencing. *Microbiol. Spect.* 12, e0011724–e0011724. doi: 10.1128/spectrum.00117-24
- Chen, J., and Xu, F. (2023). Application of nanopore sequencing in the diagnosis and treatment of pulmonary infections. *Mol. Diagn. Ther.* 27, 685–701. doi: 10.1007/s40291-023-00669-8
- Cheohen, Caio, Gomes, Vinnicius Machado Schelk, and da Silva, Manuela Leal. CNN-LSTM hybrid model for AI-driven prediction of COVID-19 severity from spike sequences and clinical data. *arXiv* (2025). Available online at: <https://arxiv.org/html/2505.23879v1> (Accessed September 24, 2025).
- Chiu, C. Y., and Miller, S. A. (2019). Clinical metagenomics. *Nat. Rev. Genet.* 20, 341–355. doi: 10.1038/s41576-019-0113-7
- Choi, S. R., and Lee, M. (2023). Transformer architecture and attention mechanisms in genome data analysis: a comprehensive review. *Biology* 12:1033. doi: 10.3390/biology12071033
- Choi, T., Pyenson, B., Liebig, J., and Pavlic, T. P. (2022). Beyond tracking: using deep learning to discover novel interactions in biological swarms. *Artif. Life Robot.* 27, 393–400. doi: 10.1007/s10015-022-00753-y
- Chong, P. L., Vaigeshwari, V., Mohammed Reyesudin, B. K., Noor Hidayah Binti, R. A., Tatchanaamoorti, P., Yeow, J. A., et al. (2025). Integrating artificial intelligence in healthcare: applications, challenges, and future directions. *Future Sci. OA* 11:2527505. doi: 10.1080/20565623.2025.2527505
- Chourasia, P., Lonkar, H., Ali, S., and Patterson, M. (2024). EPIC: enhancing privacy through iterative collaboration. doi: 10.48550/arXiv.2411.05167
- CR, M. I., Chen, X., Kunasekaran, M., Quigley, A., Lim, S., Stone, H., et al. (2023). Artificial intelligence in public health: the potential of epidemic early warning systems. *J. Int. Med. Res.* 51:03000605231159335. doi: 10.1177/03000605231159335
- Cramer, E. Y., Ray, E. L., Lopez, V. K., Bracher, J., Brennan, A., Castro Rivadeneira, A. J., et al. (2022). Evaluation of individual and ensemble probabilistic forecasts of COVID-19 mortality in the United States. *Proc. Natl. Acad. Sci. USA* 119:e2113561119. doi: 10.1073/pnas.2113561119
- Curry, A. (2003). Electron microscopy and the investigation of new infectious diseases. *Int. J. Infect. Dis.* 7, 251–258. doi: 10.1016/S1201-9712(03)90103-2
- Silva, S. J. R. da, Kohl, A., Pena, L., and Pardee, K. Clinical and laboratory diagnosis of monkeypox (mpox): current status and future directions *iScience* 2023 26:106759 doi: 10.1016/j.isci.2023.106759 PMID: 37206155

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

- Dantas, P. V., da Sabino Silva, W., Cordeiro, L. C., and Carvalho, C. B. (2024). A comprehensive review of model compression techniques in machine learning. *Appl. Intell.* 54, 11804–11844. doi: 10.1007/s10489-024-05747-w
- Das, D., Floch, H. L., Houhou, N., Epelboin, L., Hausfater, P., Khalil, A., et al. (2015). Viruses detected by systematic multiplex polymerase chain reaction in adults with suspected community-acquired pneumonia attending emergency departments in France. *Clin. Microbiol. Infect.* 21:608.e1. doi: 10.1016/j.cmi.2015.02.014
- De Coster, W., De Rijk, P., De Roeck, A., De Pooter, T., D'Hert, S., Strazisar, M., et al. (2019). Structural variants identified by Oxford nanopore PromethION sequencing of the human genome. *Genome Res.* 29, 1178–1187. doi: 10.1101/gr.244939.118
- de Olazarra, A. S., and Wang, S. X. (2023). Advances in point-of-care genetic testing for personalized medicine applications. *Biomicrofluidics* 17:031501. doi: 10.1063/5.0143311
- de Souza, L. C., Azevedo, K. S., de Souza, J. G., Barbosa, R. d. M., and Fernandes, M. A. C. (2023). New proposal of viral genome representation applied in the classification of SARS-CoV-2 with deep learning. *BMC Bioinformatics* 24:92. doi: 10.1186/s12859-023-05188-1
- de Vries, J. J. C., Brown, J. R., Fischer, N., Sidorov, I. A., Morfopoulou, S., Huang, J., et al. (2021). Benchmark of thirteen bioinformatic pipelines for metagenomic virus diagnostics using datasets from clinical samples. *J. Clin. Virol.* 141:104908. doi: 10.1016/j.jcv.2021.104908
- Deif, M. A., Solymann, A. A., Kamarposhti, M. A., Band, S. S., Hammam, R. E., Deif, M. A., et al. (2021). A deep bidirectional recurrent neural network for identification of SARS-CoV-2 from viral genome sequences. *MBE* 18, 8933–8950. doi: 10.3934/mbe.2021440
- Dhaarani, R., and Reddy, M. K. (2025). Progressing microbial genomics: artificial intelligence and deep learning driven advances in genome analysis and therapeutics. *Intell. Based Med.* 11:100251. doi: 10.1016/j.ibmed.2025.100251
- Doron, S., Melamed, S., Ofir, G., Leavitt, A., Lopatina, A., Keren, M., et al. (2018). Systematic discovery of antiphage defense systems in the microbial pangenome. *Science* 359:eaar4120. doi: 10.1126/science.aar4120
- Dutilh, B. E., Reyes, A., Hall, R. J., and Whiteson, K. L. (2017). Editorial: virus discovery by metagenomics: the (Im)possibilities. *Front. Microbiol.* 8:1710. doi: 10.3389/fmicb.2017.01710
- Edwards, R. A., McNair, K., Faust, K., Raes, J., and Dutilh, B. E. (2016). Computational approaches to predict bacteriophage-host relationships. *FEMS Microbiol. Rev.* 40, 258–272. doi: 10.1093/femsre/fuv048
- Elbehiry, A., and Abalkhail, A. (2025). Metagenomic next-generation sequencing in infectious diseases: clinical applications, translational challenges, and future directions. *Diagnostics* 15:1991. doi: 10.3390/diagnostics15161991
- Eren, K., Taktakoğlu, N., and Pirim, I. (2022). DNA sequencing methods: from past to present. *Eurasian J. Med* 54, S47–S56. doi: 10.5152/eurasianjmed.2022.22280
- Flores-Coronado, J. A., Alanis-Valdez, A. Y., Herrera-Saldivar, M. F., Flores-Flores, A. S., Vazquez-Guillen, J. M., Tamez-Guerra, R. S., et al. (2025). Awareness of the dual-use dilemma in scientific research: reflections and challenges to Latin America. *Front. Bioeng. Biotechnol.* 13. doi: 10.3389/fbioe.2025.1649781
- Gers, F. A., Schmidhuber, J., and Cummins, F. (2000). Learning to forget: continual prediction with LSTM. *Neural Comput.* 12, 2451–2471. doi: 10.1162/089976600300015015
- Gigante, C. M., Korber, B., Seabolt, M. H., Wilkins, K., Davidson, W., Rao, A. K., et al. (2022). Multiple lineages of monkeypox virus detected in the United States, 2021–2022. *Science* 378, 560–565. doi: 10.1126/science.add4153
- Giuste, F., Shi, W., Zhu, Y., Naren, T., Isgut, M., Sha, Y., et al. (2023). Explainable artificial intelligence methods in combating pandemics: a systematic review. *IEEE Rev. Biomed. Eng.* 16, 5–21. doi: 10.1109/RBME.2022.3185953
- Goldsmith, C. S., and Miller, S. E. (2009). Modern uses of Electron microscopy for detection of viruses. *Clin. Microbiol. Rev.* 22, 552–563. doi: 10.1128/CMR.00027-09
- González, A., Fullaondo, A., and Odriozola, A. (2025). Why are long-read sequencing methods revolutionizing microbiome analysis? *Microorganisms* 13:1861. doi: 10.3390/microorganisms13081861
- Goodwin, S., McPherson, J. D., and McCombie, W. R. (2016). Coming of age: ten years of next-generation sequencing technologies. *Nat. Rev. Genet.* 17, 333–351. doi: 10.1038/nrg.2016.49
- Graves, A. (2012). Supervised sequence labelling with recurrent neural networks. Berlin, Heidelberg: Springer Berlin Heidelberg.
- Greener, J. G., Kandathil, S. M., Moffat, L., and Jones, D. T. (2022). A guide to machine learning for biologists. *Nat. Rev. Mol. Cell Biol.* 23, 40–55. doi: 10.1038/s41580-021-00407-0
- Greninger, A. L. (2018). A decade of RNA virus metagenomics is (not) enough. *Virus Res.* 244, 218–229. doi: 10.1016/j.virusres.2017.10.014
- Han, Y., He, J., Li, M., Peng, Y., Jiang, H., Zhao, J., et al. (2024). Unlocking the potential of metagenomics with the PacBio high-fidelity sequencing technology. *Microorganisms* 12:2482. doi: 10.3390/microorganisms12122482
- Han, J. J., Song, H. A., Pierson, S. L., Shen-Gunther, J., and Xia, Q. (2023). Emerging infectious diseases are virulent viruses—are we prepared? An overview. *Microorganisms* 11:2618. doi: 10.3390/microorganisms11112618
- Hanna, M. G., Pantanowitz, L., Dash, R., Harrison, J. H., Deebajah, M., Pantanowitz, J., et al. (2025). Future of artificial intelligence—machine learning trends in pathology and medicine. *Mod. Pathol.* 38:100705. doi: 10.1016/j.modpat.2025.100705
- Harismendy, O., Ng, P. C., Strausberg, R. L., Wang, X., Stockwell, T. B., Beeson, K. Y., et al. (2009). Evaluation of next generation sequencing platforms for population targeted sequencing studies. *Genome Biol.* 10:R32. doi: 10.1186/gb-2009-10-3-r32
- Haug, C. J., and Drazen, J. M. (2023). Artificial intelligence and machine learning in clinical medicine, 2023. *N. Engl. J. Med.* 388, 1201–1208. doi: 10.1056/NEJMr2302038
- Heather, J. M., and Chain, B. (2016). The sequence of sequencers: the history of sequencing DNA. *Genomics* 107, 1–8. doi: 10.1016/j.ygeno.2015.11.003
- Hie, B., Zhong, E. D., Berger, B., and Bryson, B. (2021). Learning the language of viral evolution and escape. *Science* 371, 284–288. doi: 10.1126/science.abd7331
- Hirabayashi, E., Mercado, G., Hull, B., Soin, S., Koshy-Chenthittayil, S., Raman, S., et al. (2024). Comparison of diagnostic accuracy of rapid antigen tests for COVID-19 compared to the viral genetic test in adults: a systematic review and meta-analysis. *JBI Evid Synth* 22, 1939–2002. doi: 10.111124/JBIES-23-00291
- Huang, B., Jennison, A., Whaley, D., McMahon, J., Hewitson, G., Graham, R., et al. (2019). Illumina sequencing of clinical samples for virus detection in a public health laboratory. *Sci. Rep.* 9:5409. doi: 10.1038/s41598-019-41830-w
- Ieven, M. (2007). Currently used nucleic acid amplification tests for the detection of viruses and atypicals in acute respiratory infections. *J. Clin. Virol.* 40, 259–276. doi: 10.1016/j.jcv.2007.08.012
- Iwata, T. (2020). PCR detection and new therapies for COVID-19. *J. Periodontal Implant Sci* 50, 133–134. doi: 10.5051/jpis.2020.50.3.133
- Ji, Y., Zhou, Z., Liu, H., and Davuluri, R. V. (2021). DNABERT: pre-trained bidirectional encoder representations from transformers model for DNA-language in genome. *Bioinformatics* 37, 2112–2120. doi: 10.1093/bioinformatics/btab083
- Johnson, T., Jamrozik, E., Ramachandran, P., and Johnson, S. (2025). Clinical metagenomics: ethical issues. *J. Med. Microbiol.* 74:001967. doi: 10.1099/jmm.0.001967
- Joseph, J. (2025). Algorithmic bias in public health AI: a silent threat to equity in low-resource settings. *Front. Public Health* 13. doi: 10.3389/fpubh.2025.1643180
- Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., et al. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature* 596, 583–589. doi: 10.1038/s41586-021-03819-2
- Jurtz, V. I., Johansen, A. R., Nielsen, M., Almagro Armenteros, J. J., Nielsen, H., Sønderby, C. K., et al. (2017). An introduction to deep learning on biological sequence data: examples and solutions. *Bioinformatics* 33, 3685–3690. doi: 10.1093/bioinformatics/btx531
- Kaur, J., and Butt, Z. A. (2025). AI-driven epidemic intelligence: the future of outbreak detection and response. *Front. Artif. Intell.* 8:1645467. doi: 10.3389/frai.2025.1645467
- Khan, S. F., Rathod, P., Gupta, V. K., Khedekar, P. B., and Chikhale, R. V. (2024). Evolution and impact of nucleic acid amplification test (NAAT) for diagnosis of coronavirus disease. *Anal. Chem.* 96, 8124–8146. doi: 10.1021/acs.analchem.3c05225
- Kozich, J. J., Westcott, S. L., Baxter, N. T., Highlander, S. K., and Schloss, P. D. (2013). Development of a dual-index sequencing strategy and curation pipeline for Analyzing amplicon sequence data on the MiSeq Illumina sequencing platform. *Appl. Environ. Microbiol.* 79, 5112–5120. doi: 10.1128/AEM.01043-13
- Krivonos, D. V., Fedorov, D. E., Konanov, D. N., Vvedensky, A. V., Sonets, I. V., Korneenko, E. V., et al. (2025). Pike: OTU-level analysis for Oxford nanopore amplicon metagenomics. *Int. J. Mol. Sci.* 26:4168. doi: 10.3390/ijms26094168
- Kuo, C.-W., and Ying, J. J.-C. (2023). “An unsupervised deep learning framework for anomaly detection” in Intelligent information and database systems. eds. N. T. Nguyen, S. Boonsang, H. Fujita, B. Hnatkowska, T.-P. Hong and K. Pasupaet al. (Singapore: Springer Nature), 284–295.
- Langer, B. E., Amaral, A., Baudement, M.-O., Bonath, F., Charles, M., Chitneedi, P. K., et al. (2025). Empowering bioinformatics communities with Nextflow and NF-core. *Genome Biol.* 26:228. doi: 10.1186/s13059-025-03673-9
- Le Piane, F., Voza, M., Baldoni, M., and Mercuri, F. (2024). Integrating high-performance computing, machine learning, data management workflows, and infrastructures for multiscale simulations and nanomaterials technologies. *Beilstein J. Nanotechnol.* 15, 1498–1521. doi: 10.3762/bjnano.15.119
- Lee, M. (2023). Deep learning techniques with genomic data in Cancer prognosis: a comprehensive review of the 2021–2023 literature. *Biology (Basel)* 12:893. doi: 10.3390/biology12070893
- Lee, J.-Y. (2023). The principles and applications of high-throughput sequencing technologies. *Dev. Reprod.* 27, 9–24. doi: 10.12717/DR.2023.27.1.9
- Lee, J. M., Jansen, R., Sanderson, K. E., Guerra, F., Keller-Olaman, S., Murti, M., et al. (2023). Public health emergency preparedness for infectious disease emergencies: a scoping review of recent evidence. *BMC Public Health* 23:420. doi: 10.1186/s12889-023-15313-7

- Leland, D. S., and Ginocchio, C. C. (2007). Role of cell culture for virus detection in the age of technology. *Clin. Microbiol. Rev.* 20, 49–78. doi: 10.1128/cmr.00002-06
- Li, K., DeCost, B., Choudhary, K., Greenwood, M., and Hattrick-Simpers, J. (2023). A critical examination of robustness and generalizability of machine learning prediction of materials properties. *NPJ Comput. Mater.* 9:55. doi: 10.1038/s41524-023-01012-9
- Li, M., Wang, Y., Li, F., Zhao, Y., Liu, M., Zhang, S., et al. (2021). A deep learning-based method for identification of bacteriophage-host interaction. *IEEE/ACM Trans. Comput. Biol. Bioinform.* 18, 1801–1810. doi: 10.1109/TCBB.2020.3017386
- Li, Y., Zhao, H., Wilkins, K., Hughes, C., and Damon, I. K. (2010). Real-time PCR assays for the specific detection of monkeypox virus west African and Congo Basin strain DNA. *J. Virol. Methods* 169, 223–227. doi: 10.1016/j.jviromet.2010.07.012
- Linares, M., Pérez-Tanoira, R., Carrero, A., Romanyk, J., Pérez-García, F., Gómez-Herruz, P., et al. (2020). Panbio antigen rapid test is reliable to diagnose SARS-CoV-2 infection in the first 7 days after the onset of symptoms. *J. Clin. Virol.* 133:104659. doi: 10.1016/j.jcv.2020.104659
- Liu, W. (2025). Bracing the artificial intelligence technology in viral infectious disease control. *Inf. Med.* 4:100186. doi: 10.1016/j.imj.2025.100186
- Liu, Y., Han, R., Zhou, L., Luo, M., Zeng, L., Zhao, X., et al. (2021). Comparative performance of the GenoLab M and NovaSeq 6000 sequencing platforms for transcriptome and LncRNA analysis. *BMC Genomics* 22:829. doi: 10.1186/s12864-021-08150-8
- Liu, J., Li, J., Wang, H., and Yan, J. (2020). Application of deep learning in genomics. *Sci. China Life Sci.* 63, 1860–1878. doi: 10.1007/s11427-020-1804-5
- Logares, R., Sunagawa, S., Salazar, G., Cornejo-Castillo, F. M., Ferrera, I., Sarmiento, H., et al. (2014). Metagenomic 16S rDNA Illumina tags are a powerful alternative to amplicon sequencing to explore diversity and structure of microbial communities. *Environ. Microbiol.* 16, 2659–2671. doi: 10.1111/1462-2920.12250
- Lou, J., Liang, W., Cao, L., Hu, I., Zhao, S., Chen, Z., et al. (2024). Predictive evolutionary modelling for influenza virus by site-based dynamics of mutations. *Nat. Commun.* 15:2546. doi: 10.1038/s41467-024-46918-0
- Louten, J. (2016). Detection and diagnosis of viral infections. *Essential Human Virol.*, 111–132. doi: 10.1016/B978-0-12-800947-5.00007-7
- Lu, H., Giordano, F., and Ning, Z. (2016). Oxford nanopore MinION sequencing and genome assembly. *Genomics Proteomics Bioinformatics* 14, 265–279. doi: 10.1016/j.gpb.2016.05.004
- Lundberg, S. M., and Lee, S. I. (2017). A Unified Approach to Interpreting Model Predictions. Proceedings of the 31st International Conference on Neural Information Processing Systems, Long Beach, 4–9 December 2017, 4768–4777. Available online at: <https://www.scirp.org/reference/referencespapers?referenceid=3862505> (Accessed September 22, 2025).
- Luvira, V., Leangwutiwong, P., Thippornchai, N., Thawornkuno, C., Chatchen, S., Chanchaonthana, W., et al. (2022). False positivity of anti-SARS-CoV-2 antibodies in patients with acute tropical diseases in Thailand. *Trop Med Infect Dis* 7:132. doi: 10.3390/tropicalmed7070132
- Marić, J., Krizanović, K., Riondet, S., Nagarajan, N., and Šikić, M. (2024). Comparative analysis of metagenomic classifiers for long-read sequencing datasets. *BMC Bioinformatics* 25:15. doi: 10.1186/s12859-024-05634-8
- Martínez, M. A., del Soto-Río, M. d. I. D., Gutiérrez, R. M., Chiu, C. Y., Greninger, A. L., Contreras, J. F., et al. (2014). DNA microarray for detection of gastrointestinal viruses. *J. Clin. Microbiol.* 53, 136–145. doi: 10.1128/jcm.01317-14
- Martinez-Martin, N., and Magnus, D. (2019). Privacy and ethical challenges in next-generation sequencing. *Expert Rev. Precis. Med. Drug Dev.* 4, 95–104. doi: 10.1080/23808993.2019.1599685
- Matsunaga, R., and Tsumoto, K. (2025). Accelerating antibody discovery and optimization with high-throughput experimentation and machine learning. *J. Biomed. Sci.* 32:46. doi: 10.1186/s12929-025-01141-x
- Miller, R. R., Montoya, V., Gardy, J. L., Patrick, D. M., and Tang, P. (2013). Metagenomics for pathogen detection in public health. *Genome Med.* 5:81. doi: 10.1186/gm485
- Ming, Z., Chen, X., Wang, S., Liu, H., Yuan, Z., Wu, M., et al. (2023). HostNet: improved sequence representation in deep neural networks for virus-host prediction. *BMC Bioinformatics* 24:455. doi: 10.1186/s12859-023-05582-9
- Mokili, J. L., Rohwer, F., and Dutilh, B. E. (2012). Metagenomics and future perspectives in virus discovery. *Curr. Opin. Virol.* 2, 63–77. doi: 10.1016/j.coviro.2011.12.004
- Morgan, J. L., Darling, A. E., and Eisen, J. A. (2010). Metagenomic sequencing of an in vitro-simulated microbial community. *PLoS One* 5:e10209. doi: 10.1371/journal.pone.0010209
- Moustakli, E., Christopoulos, P., Potiris, A., Zikopoulos, A., Mavrogianni, D., Karampas, G., et al. (2025). Long-read sequencing and structural variant detection: unlocking the hidden genome in rare genetic disorders. *Diagnostics (Basel)* 15:1803. doi: 10.3390/diagnostics15141803
- Msoni, N. S., Levy, J. I., Matteson, N. L., Ndlovu, N., Ntuli, P., Baer, A., et al. (2025). Wastewater-integrated pathogen surveillance dashboards enable real-time, transparent, and interpretable public health risk assessment and dissemination. *PLOS Glob Public Health* 5:e0004443. doi: 10.1371/journal.pgph.0004443
- Mswahili, M. E., and Jeong, Y.-S. (2024). Transformer-based models for chemical SMILES representation: a comprehensive literature review. *Heliyon* 10:e39038. doi: 10.1016/j.heliyon.2024.e39038
- Nasir, A., Aamir, U. B., Kanji, A., Bukhari, A. R., Ansar, Z., Ghanchi, N. K., et al. (2023). Tracking SARS-CoV-2 variants through pandemic waves using RT-PCR testing in low-resource settings. *PLOS Glob Public Health* 3:e0001896. doi: 10.1371/journal.pgph.0001896
- Nayfach, S., Páez-Espino, D., Call, L., Low, S. J., Sberro, H., Ivanova, N. N., et al. (2021). Metagenomic compendium of 189,680 DNA viruses from the human gut microbiome. *Nat. Microbiol.* 6, 960–970. doi: 10.1038/s41564-021-00928-6
- Nazer, L. H., Zatarah, R., Waldrip, S., Ke, J. X. C., Moukheiber, M., Khanna, A. K., et al. (2023). Bias in artificial intelligence algorithms and recommendations for mitigation. *PLoS Digit Health* 2:e0000278. doi: 10.1371/journal.pdig.0000278
- Ni, P., Huang, N., Zhang, Z., Wang, D.-P., Liang, F., Miao, Y., et al. (2019). DeepSignal: detecting DNA methylation state from nanopore sequencing reads using deep-learning. *Bioinformatics* 35, 4586–4595. doi: 10.1093/bioinformatics/btz276
- Oehler, J. B., Wright, H., Stark, Z., Mallett, A. J., and Schmitz, U. (2023). The application of long-read sequencing in clinical settings. *Hum. Genomics* 17:73. doi: 10.1186/s40246-023-00522-3
- Pasolli, E., Asnicar, F., Manara, S., Zolfo, M., Karcher, N., Armanini, F., et al. (2019). Extensive unexplored human microbiome diversity revealed by over 150,000 genomes from metagenomes spanning age, geography, and lifestyle. *Cell* 176, 649–662.e20. doi: 10.1016/j.cell.2019.01.001
- Pavia, C. S., and Plummer, M. M. (2021). The evolution of rapid antigen detection systems and their application for COVID-19 and other serious respiratory infectious diseases. *J. Microbiol. Immunol. Infect.* 54, 776–786. doi: 10.1016/j.jmii.2021.06.003
- Peters, A., Vetter, P., Guitart, C., Lotfinejad, N., and Pittet, D. (2020). Understanding the emerging coronavirus: what it means for health security and infection prevention. *J. Hosp. Infect.* 104, 440–448. doi: 10.1016/j.jhin.2020.02.023
- Pigott, D. M., Golding, N., Mylne, A., Huang, Z., Henry, A. J., Weiss, D. J., et al. (2014). Mapping the zoonotic niche of Ebola virus disease in Africa. *eLife* 3:e04395. doi: 10.7554/eLife.04395
- Pita-Galeana, M. A., Ruhle, M., López-Vázquez, L., de Anda-Jáuregui, G., and Hernández-Lemus, E. (2025). Computational metagenomics: state of the art. *Int. J. Mol. Sci.* 26:9206. doi: 10.3390/ijms26189206
- Rampelli, S., Soverini, M., D'Amico, F., Barone, M., Tavella, T., Monti, D., et al. (2020). Shotgun metagenomics of gut microbiota in humans with up to extreme longevity and the increasing role of xenobiotic degradation. *mSystems* 5, e00124–e00120. doi: 10.1128/mSystems.00124-20
- Reich, N. G., McGowan, C. J., Yamana, T. K., Tushar, A., Ray, E. L., Osthus, D., et al. (2019). Accuracy of real-time multi-model ensemble forecasts for seasonal influenza in the US. *PLoS Comput. Biol.* 15:e1007486. doi: 10.1371/journal.pcbi.1007486
- Reintjes, R., and Zanuzdana, A. (2009). Outbreak investigations. *Mod. Infect. Dis. Epidemiol.*, 159–176. doi: 10.1007/978-0-387-93835-6_9
- Ren, J., Song, K., Deng, C., Ahlgren, N. A., Fuhrman, J. A., Li, Y., et al. (2020). Identifying viruses from metagenomic data using deep learning. *Quant Biol* 8, 64–77. doi: 10.1007/s40484-019-0187-4
- Richert-Pöggeler, K. R., Franzke, K., Hipp, K., and Kleespies, R. G. (2019). Electron microscopy methods for virus diagnosis and high resolution analysis of viruses. *Front. Microbiol.* 9. doi: 10.3389/fmicb.2018.03255
- Rives, A., Meier, J., Sercu, T., Goyal, S., Lin, Z., Liu, J., et al. (2021). Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proc. Natl. Acad. Sci. USA* 118:e2016239118. doi: 10.1073/pnas.2016239118
- Rooney, A. M., Raphenya, A. R., Melano, R. G., Seah, C., Yee, N. R., MacFadden, D. R., et al. (2022). Performance characteristics of next-generation sequencing for the detection of antimicrobial resistance determinants in *Escherichia coli* genomes and metagenomes. *mSystems* 7:e00022–22. doi: 10.1128/msystems.00022-22
- Rosenboom, I., Scheithauer, T., Friedrich, F. C., Pörtner, S., Hollstein, L., Pust, M.-M., et al. (2022). Wochenende — modular and flexible alignment-based shotgun metagenome analysis. *BMC Genomics* 23:748. doi: 10.1186/s12864-022-08985-9
- Roux, S., Adriaenssens, E. M., Dutilh, B. E., Koonin, E. V., Kropinski, A. M., Krupovic, M., et al. (2019). Minimum information about an uncultivated virus genome (MIUViG). *Nat. Biotechnol.* 37, 29–37. doi: 10.1038/nbt.4306
- Roux, S., Matthijnssens, J., and Dutilh, B. E. (2021). Metagenomics in virology. *Encyclopedia Virol.*, 133–140. doi: 10.1016/B978-0-12-809633-8.20957-6
- Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat. Mach. Intell.* 1, 206–215. doi: 10.1038/s42256-019-0048-x
- Rumelhart, D. E., Hinton, G. E., and Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature* 323, 533–536.
- Samek, W., Wiegand, T., and Müller, K.-R. (2017). Explainable artificial intelligence: understanding, visualizing and interpreting deep learning models. doi: 10.48550/arXiv.1708.08296

- Santiago-Rodriguez, T. M., and Hollister, E. B. (2022). Unraveling the viral dark matter through viral metagenomics. *Front. Immunol.* 13. doi: 10.3389/fimmu.2022.1005107
- Saravanan, K. A., Panigrahi, M., Kumar, H., Rajawat, D., Nayak, S. S., Bhushan, B., et al. (2022). Role of genomics in combating COVID-19 pandemic. *Gene* 823:146387. doi: 10.1016/j.gene.2022.146387
- Schirmer, M., D'Amore, R., Ijaz, U. Z., Hall, N., and Quince, C. (2016). Illumina error profiles: resolving fine-scale variation in metagenomic sequencing data. *BMC Bioinformatics* 17:125. doi: 10.1186/s12859-016-0976-y
- Schuele, L., Masirika, L. M., Cassidy, H., Clausen, P. T. L. C., Zaeck, L. M., Boter, M., et al. (2025). Metagenomic sequencing of mpox virus clade Ib lesions identifies possible bacterial and viral co-infections in hospitalized patients in eastern DRC. *Microbiol Spect.* 13:e00512-25. doi: 10.1128/spectrum.00512-25
- Schulz, F., Roux, S., Paez-Espino, D., Jungbluth, S., Walsh, D. A., Deneff, V. J., et al. (2020). Giant virus diversity and host interactions through global metagenomics. *Nature* 578, 432–436. doi: 10.1038/s41586-020-1957-x
- Senior, A. W., Evans, R., Jumper, J., Kirkpatrick, J., Sifre, L., Green, T., et al. (2020). Improved protein structure prediction using potentials from deep learning. *Nature* 577, 706–710. doi: 10.1038/s41586-019-1923-7
- Sereika, M., Kirkegaard, R. H., Karst, S. M., Michaelsen, T. Y., Sørensen, E. A., Wollenberg, R. D., et al. (2022). Oxford nanopore R10.4 long-read sequencing enables the generation of near-finished bacterial genomes from pure cultures and metagenomes without short-read or reference polishing. *Nat. Methods* 19, 823–826. doi: 10.1038/s41592-022-01539-7
- Sevim, V., Lee, J., Egan, R., Clum, A., Hundley, H., Lee, J., et al. (2019). Shotgun metagenome data of a defined mock community using Oxford nanopore, PacBio and Illumina technologies. *Sci Data* 6:285. doi: 10.1038/s41597-019-0287-z
- Shafin, K., Pesout, T., Lorig-Roach, R., Haukness, M., Olsen, H. E., Bosworth, C., et al. (2020). Nanopore sequencing and the Shasta toolkit enable efficient de novo assembly of eleven human genomes. *Nat. Biotechnol.* 38, 1044–1053. doi: 10.1038/s41587-020-0503-6
- Shkoporov, A. N., Stockdale, S. R., Lavelle, A., Kondova, I., Heuston, C., Upadrasa, A., et al. (2022). Viral biogeography of the mammalian gut and parenchymal organs. *Nat. Microbiol.* 7, 1301–1311. doi: 10.1038/s41564-022-01178-w
- Singh, R., Lanchantin, J., Robins, G., and Qi, Y. (2016). DeepChrome: deep-learning for predicting gene expression from histone modifications. *Bioinformatics* 32, i639–i648. doi: 10.1093/bioinformatics/btw427
- Slavov, S. N. (2025). Routine detection of viruses through metagenomics: where do we stand? *Am. J. Trop. Med. Hyg.* 112:652. doi: 10.4269/ajtmh.24-0652
- Soenksen, L. R., Ma, Y., Zeng, C., Boussieux, L., Villalobos Carballo, K., Na, L., et al. (2022). Integrated multimodal artificial intelligence framework for healthcare applications. *NPJ Digit Med* 5:149. doi: 10.1038/s41746-022-00689-4
- Song, S., Ma, L., Xu, X., Shi, H., Li, X., Liu, Y., et al. (2021). Rapid screening and identification of viral pathogens in metagenomic data. *BMC Med. Genet.* 14:289. doi: 10.1186/s12920-021-01138-z
- Srivastava, V., Kumar, R., Wani, M. Y., Robinson, K., and Ahmad, A. (2025). Role of artificial intelligence in early diagnosis and treatment of infectious diseases. *Infect. Dis.* 57, 1–26. doi: 10.1080/23744235.2024.2425712
- Takemae, N., Kuba, Y., Oba, K., and Kageyama, T. (2024). Direct genome sequencing of respiratory viruses from low viral load clinical specimens using the target capture sequencing technology. *Microbiol. Spect.* 12, e0098624–e0098624. doi: 10.1128/spectrum.00986-24
- Tampuu, A., Bzhilava, Z., Dillner, J., and Vicente, R. (2019). ViraMiner: deep learning on raw DNA sequences for identifying viral genomes in human samples. *PLoS One* 14:e0222271. doi: 10.1371/journal.pone.0222271
- Thomas, T., Gilbert, J., and Meyer, F. (2012). Metagenomics – a guide from sampling to data analysis. *Microb Inform Exp* 2:3. doi: 10.1186/2042-5783-2-3
- Tisza, M. J., and Buck, C. B. (2021). A catalog of tens of thousands of viruses from human metagenomes reveals hidden associations with chronic diseases. *Proc. Natl. Acad. Sci. USA* 118:e2023202118. doi: 10.1073/pnas.2023202118
- Travers, K. J., Chin, C.-S., Rank, D. R., Eid, J. S., and Turner, S. W. (2010). A flexible and efficient template format for circular consensus sequencing and SNP detection. *Nucleic Acids Res.* 38:e159. doi: 10.1093/nar/gkq543
- Unsal, S., Atas, H., Albayrak, M., Turhan, K., Acar, A. C., and Doğan, T. (2022). Learning functional properties of proteins with language models. *Nat. Mach. Intell.* 4, 227–245. doi: 10.1038/s42256-022-00457-9
- Vashisht, V., Vashisht, A., Mondal, A. K., Farmaha, J., Alptekin, A., Singh, H., et al. (2023). Genomics for emerging pathogen identification and monitoring: prospects and obstacles. *Biomed Informatics* 3, 1145–1177. doi: 10.3390/biomedinformatics3040069
- Villanueva-Miranda, I., Xiao, G., and Xie, Y. (2025). Artificial intelligence in early warning systems for infectious disease surveillance: a systematic review. *Front. Public Health* 13:1609615. doi: 10.3389/fpubh.2025.1609615
- Wagner, M. M., Gresham, L. S., and Dato, V. (2006). Case detection, outbreak detection, and outbreak characterization. *Handb. Biosurveillance*, 27–50. doi: 10.1016/B978-012369378-5/50005-3
- Wang, D., Chen, Y., Xiang, S., Hu, H., Zhan, Y., Yu, Y., et al. (2023). Recent advances in immunoassay technologies for the detection of human coronavirus infections. *Front. Cell. Infect. Microbiol.* 12:1040248. doi: 10.3389/fcimb.2022.1040248
- Wang, D., Coscoy, L., Zylberberg, M., Avila, P. C., Boushey, H. A., Ganem, D., et al. (2002). Microarray-based detection and genotyping of viral pathogens. *Proc. Natl. Acad. Sci. USA* 99, 15687–15692. doi: 10.1073/pnas.242579699
- Wang, D., Urisman, A., Liu, Y.-T., Springer, M., Ksiazek, T. G., Erdman, D. D., et al. (2003). Viral discovery and sequence recovery using DNA microarrays. *PLoS Biol.* 1:E2. doi: 10.1371/journal.pbio.0000002
- Watkins, R. E., Eagleson, S., Hall, R. G., Dailey, L., and Plant, A. J. (2006). Approaches to the evaluation of outbreak detection methods. *BMC Public Health* 6:263. doi: 10.1186/1471-2458-6-263
- Wen, L., and Tang, F. (2025). Single-cell omics sequencing technologies: the long-read generation. *Trends Genet.* doi: 10.1016/j.tig.2025.07.012
- Wenger, A. M., Peluso, P., Rowell, W. J., Chang, P.-C., Hall, R. J., Concepcion, G. T., et al. (2019). Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nat. Biotechnol.* 37, 1155–1162. doi: 10.1038/s41587-019-0217-9
- Wilder-Smith, A. (2021). COVID-19 in comparison with other emerging viral diseases: risk of geographic spread via travel. *Trop. Dis. Travel Med. Vaccines* 7:3. doi: 10.1186/s40794-020-00129-9
- Willmington, C., Belardi, P., Murante, A. M., and Vainieri, M. (2022). The contribution of benchmarking to quality improvement in healthcare. A systematic literature review. *BMC Health Serv. Res.* 22:139. doi: 10.1186/s12913-022-07467-8
- Wu, S., Fang, Z., Tan, J., Li, M., Wang, C., Guo, Q., et al. (2021). DeePhage: distinguishing virulent and temperate phage-derived sequences in metavirome data with a deep learning approach. *Gigascience* 10:giab056. doi: 10.1093/gigascience/giab056
- Wu, L.-Y., Wijesekara, Y., Piedade, G. J., Pappas, N., Brussaard, C. P. D., and Dutilh, B. E. (2024). Benchmarking bioinformatic virus identification tools using real-world metagenomic data across biomes. *Genome Biol.* 25:97. doi: 10.1186/s13059-024-03236-4
- Xia, Y., Li, X., Wu, Z., Nie, C., Cheng, Z., Sun, Y., et al. (2023). Strategies and tools in illumina and nanopore-integrated metagenomic analysis of microbiome data. *iMeta* 2:e72. doi: 10.1002/imt2.72
- Yagin, F. H., Cicek, İ. B., Alkhateeb, A., Yagin, B., Colak, C., Azzeh, M., et al. (2023). Explainable artificial intelligence model for identifying COVID-19 gene biomarkers. *Comput. Biol. Med.* 154:106619. doi: 10.1016/j.combiomed.2023.106619
- Yang, Z., Shan, Y., Liu, X., Chen, G., Pan, Y., Gou, Q., et al. (2024). VirID: beyond virus discovery—an integrated platform for comprehensive RNA virus characterization. *Mol. Biol. Evol.* 41:msae202. doi: 10.1093/molbev/msae202
- Ye, Y., Pandey, A., Bawden, C., Sumsuzzman, D. M., Rajput, R., Shoukat, A., et al. (2025). Integrating artificial intelligence with mechanistic epidemiological modeling: a scoping review of opportunities and challenges. *Nat. Commun.* 16:581. doi: 10.1038/s41467-024-55461-x
- Yimer, S. A., Booij, B. B., Tobert, G., Hebbeler, A., Oloo, P., Brangel, P., et al. (2024). Rapid diagnostic test: a critical need for outbreak preparedness and response for high priority pathogens. *BMJ Glob. Health* 9:e014386. doi: 10.1136/bmjgh-2023-014386
- Yurdem, B., Kuzlu, M., Gullu, M. K., Catak, F. O., and Tabassum, M. (2024). Federated learning: overview, strategies, applications, tools and future directions. *Heliyon* 10:e38137. doi: 10.1016/j.heliyon.2024.e38137
- Zhang, C., Liu, P., Li, J., Han, M., Liu, Y., Xing, W., et al. (2025). Adaptation of single molecule real time (SMRT) sequence technology for hepatitis C virus genome sequencing and identification of resistance-associated substitutions. *Virology* 605:110481. doi: 10.1016/j.virol.2025.110481
- Zhao, C., Li, F., Peng, Z., Zhou, X., and Zhuge, Y. (2023). A structured multi-head attention prediction method based on heterogeneous financial data. *PeerJ Comput Sci* 9:e1653. doi: 10.7717/peerj-cs.1653
- Zwiers, L. C., Grobbee, D. E., Uijl, A., and Ong, D. S. Y. (2024). Federated learning as a smart tool for research on infectious diseases. *BMC Infect. Dis.* 24:1327. doi: 10.1186/s12879-024-10230-5