iScience

CellPress
OPEN ACCESS

Review

Agnostic capture of pathogens for the detection and diagnostics of emerging threats

Anastasiya Sakkos,[1,4] Brandon Saint-John,[2,4] Tomas Tyml,[1] Eva Myskova,[1] Lorenzo Aureli,[1] Jamie L. Inman,[2] Antoine M. Snijders,[2] Nigel J. Mouncey,[1] Harshini Mukundan,[3,*] and Frederik Schulz[1,*]

[1]DOE Joint Genome Institute, Lawrence Berkeley National Laboratory, Berkeley, CA, USA
[2]Biological Systems and Engineering Division, Lawrence Berkeley National Laboratory, Berkeley, CA, USA
[3]Office of National and Homeland Security, Lawrence Berkeley National Laboratory, Berkeley, CA, USA
[4]These authors contributed equally
*Correspondence: hmukundan@lbl.gov (H.M.), fschulz@lbl.gov (F.S.)
https://doi.org/10.1016/j.isci.2026.114684

SUMMARY

The continued emergence of pathogens, whether novel, re-emerging, or engineered, poses a persistent global biosecurity and public health challenge. Recent outbreaks, including COVID-19, Lassa fever, Marburg virus, mpox, and avian influenza, underscore the urgent need for robust systems that enable rapid surveillance, early diagnosis, and timely countermeasures before widespread human transmission occurs. In this article, we focus on early detection technologies and systematically evaluate current diagnostic and sensing modalities. We highlight sequencing and spectroscopy as two complementary approaches capable of providing broad, agnostic detection and rich biological insight. Our analysis emphasizes that scientific innovation alone is insufficient: effective preparedness also requires improved data curation, integration, and sharing to build AI-ready resources that accelerate future responses. We argue for coordinated advances in both technological capabilities and supporting infrastructure to enable the rapid identification and characterization of emerging pathogens and to fully leverage modern science against evolving infectious threats.

INTRODUCTION

When a new pathogen emerges, people and resources are mobilized to assess its health risks and develop appropriate response measures. Indeed, different elements of this response were exercised during the COVID-19 pandemic, revealing both successes and challenges.[1] For instance, within just a few months after the genomic sequence of SARS-CoV-2 was identified, the Cryo-EM structure of its spike protein was determined,[2] followed by additional structures and functional insights.[3] The rapid development of PCR-based targeted diagnostics, sequence-based characterization of variants, and deployable immunoassays contributed to improved situational awareness and decision-making. Additional advances were made through the development of tools for epidemiological modeling and wastewater surveillance of SARS-CoV-2. In parallel to these efforts, the swift development and validation of an mRNA vaccine beginning in March of 2020, followed by its global manufacturing and distribution, demonstrated remarkable scientific and logistical achievements. Similar success was seen in the development of therapeutics.[4] Taken together, the global COVID-19 response showcased remarkable human ingenuity and scientific innovation, yet earlier availability of prediction, surveillance, and detection tools could have significantly reduced the peak of the pandemic. The lessons learned from these challenges underscore the importance of developing more agile response frameworks for future threats. They also underscore the need to systematize our response architecture in order to leverage the lessons learned and capabilities developed in addressing future outbreaks. Indeed, subsequently, recent experiences with mpox, avian influenza, Lassa, and Marburg virus outbreaks have already tested our improved capabilities and further substantiated the need for such robust response frameworks. However, in each of these cases, the identification of an emerging pathogen - in the context of zoonoses and risk to the human population - can be further accelerated to mitigate human impact. In this article, we have tried to identify rate limiting steps in our response and recommend strategic investment and development to establish a more resilient framework for the future.

In assessing our response to emerging threats in the aforementioned situations, two critical issues emerge. First, most methods required for the detection and diagnosis of threats in complex environmental or clinical samples are "targeted," meaning they are specifically designed and developed for a particular pathogen. Whereas this ensures specificity of identification, such diagnostics require redesign and redevelopment for every new emerging threat, significantly delaying deployment. For instance, with COVID-19, methods for the amplification of specific nucleic acid sequences with polymerase chain reaction (PCR) for the detection/diagnosis of SARS-CoV-2,[5] or antibody-based assays to detect capsid features of the virus were

developed, and had to be re-designed for many variants.[6] While these specific tools proved critical for improving situational awareness, guiding therapeutic intervention, informing quarantine decisions, and facilitating outbreak containment, such targeted approaches do not prepare us for the next emerging threat. Furthermore, these methods require extensive pathogen characterization and development of specific reagents, all of which is time-consuming, expensive, and technically challenging. The second issue that we observed is the lack of an integrated and standardized response framework to address emerging biological threats. This is significant for various reasons. For one, an effective response can greatly benefit from the availability of workflows that integrate scientific innovation, technical platforms, manufacturability, deployment, and workforce development. This can speed up the response time, allowing us to meet critical decision points during the course of outbreak progression. It is also important to capture and collate the data for posterity, especially to ensure AI-readiness for future use. An integrated and standardized architecture would thus not only enable timely and effective response today, but also increase the agility and speed of decision making in future events.

To address these dual challenges, we will explore agnostic technologies, which we define on a spectrum. Targeted methods, such as conventional PCR, are designed to detect a single, known signature. Agnostic methods, such as metagenomic sequencing or reagent-free spectroscopy, can detect any biological signature, including novel pathogens. Between these lie semi-agnostic approaches, such as pan-viral PCR or probe-capture panels, which detect any member of a broad class. In this review, we focus on two powerful modalities: sequencing and spectroscopy (Figure 1; Table 1). While other methods exist (e.g., microscopy and hybridization arrays), we chose these two because sequencing is the most mature agnostic method, while spectroscopy is a rapidly emerging reagent-free platform. Examining these complementary approaches allows us to cover a broad range of challenges, from sample collection to data analysis.

TOWARD AN AGNOSTIC DETECTION FRAMEWORK

The selection of environmental samples for surveillance and detection significantly impacts pathogen identification and future preparedness. Currently, outbreak-specific intelligence, including the nature of the pathogen, transmission routes, geographical prevalence, and animal hosts, guides sample collection and biomarker assessment choices. Pathogens and their signatures can be detected in diverse sources, including municipal wastewater, drinking water supplies, and food.[7–9] Further, both aerosolized pathogens (especially respiratory organisms) and volatile organic compounds (VOCs) exhaled by an infected individual can provide identifying information, adding air as another sample source.[10] Often, the transmission of an infection to human populations occurs from animal reservoirs (zoonosis); early assessment of which is also required to address and mitigate impact on human health, as is being evidenced with the current H5N1 outbreak.[11] For diagnostics, based on the pathogen and route of transmission and symptomatology, the choice of the clinical sample can vary. Once a pathogen has been shown to infect humans, clinical diagnosis becomes critical to address the presence, transmission, and impact. Clinical samples span a wide range, including bodily fluids such as saliva, sweat, urine, nasal mucus, and blood, all potential sources of pathogen-specific biomarkers.

The answer to the question of what to sample is currently driven by (early) insights into the pathogen in question, route of entry, and other factors (e.g., symptomatology, accessibility, and others). This decision is made for every emerging pathogen, a process that delays responsiveness or limits surveillance to categories/sub-types of pathogens based on the nature of the sampling method. Strategizing agnostic sampling strategies and interlinking information derived can truly expedite responsiveness to future outbreaks.

In addition to intact pathogens, biochemical signatures (biomarkers) such as nucleic acids, proteins, carbohydrates, metabolites, and fragments of bacterial, eukaryotic, and viral cells can be detected at the population level in samples, both environmental and clinical. The growing repertoire of peptide signatures enables mass spectrometry to identify bacteria in clinical samples on various taxonomic levels, expanding the use of proteinaceous samples. Thus, the range of biochemical signatures that can provide information on an emerging pathogen is significantly diverse.

Based on the sample chosen and the biochemical nature of the target being measured, processing to extricate and concentrate key signatures/biochemical markers is the next significant step in detection/diagnostics. Finally, most conventional sample processing methods are focused on a single biochemical signature type, and at the cost of other (potentially relevant) signatures. For instance, the extraction of nucleic acids involves the use of solvents and methods that destroy proteins and other hydrophilic components in the sample. Some investigators are working on more universal sampling methods that can capture and compartmentalize different biochemical signatures from a single sample, which together can provide more information about the threat than any single biomarker alone.[12] Further development of these exploratory methods can assist with better preparedness against uncharacterized and unanticipated threats in the future.

NUCLEIC ACID-BASED AGNOSTIC DETECTION

With advances in next-generation sequencing (NGS), agnostic diagnostics for identifying a range of illnesses have emerged as a viable strategy.[13] NGS has subsequently been adapted for agnostic surveillance in complex environmental samples such as wastewater.[14] Other methods, such as spectroscopic sensing, are also gaining validity for agnostic detection. Decoding a pathogen's nucleic acid sequence reveals its unique signature, enabling both targeted and untargeted/agnostic identification. This section covers targeted (e.g., PCR) and untargeted (metagenomic sequencing) methods for early pathogen detection and diagnosis.

Enrichment for nucleic acids

For the extraction of nucleic acid signatures, commercially available RNA and DNA isolation kits and automation platforms are deployed and can reduce the complexity of such processes. Successful nucleic acid detection often requires enrichment, particularly when target genetic material is present in low abundance. The need for enrichment varies significantly between

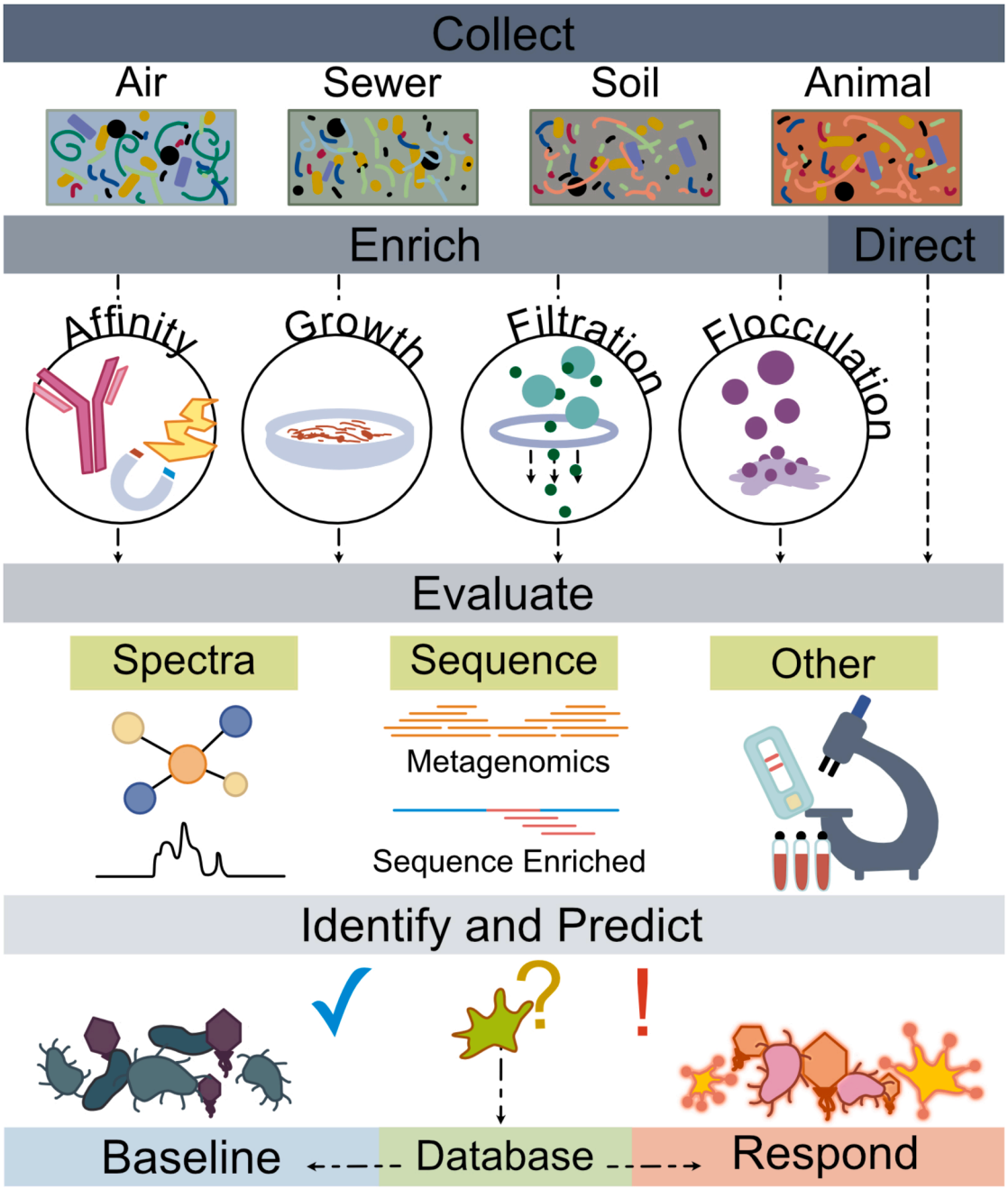

Figure 1. An integrated workflow for pathogen detection

Agnostic detection or diagnosis of emerging/unknown threats requires the interrogation of diverse environmental samples, such as wastewater and aerosols (detection), as well as clinical samples derived from human or animal hosts (One-Health, diagnostics). Such samples may need to be processed via methods such as enrichment, or may be used directly. Whereas many methods (labeled "Other") may be used or are being evaluated for agnostic detection and diagnostics, we specifically focus on spectral signatures and metagenomic and metatranscriptomic sequencing in order to identify and predict known and new pathogens. An integrated workflow that assimilates all of these components can not only establish a baseline for responsiveness but also facilitate rapid response. Indeed, the integrated database of such signatures, when combined with emerging technologies such as artificial intelligence and machine learning, can greatly expedite future preparedness.

clinical diagnostics, where pathogen loads are typically higher, and environmental detection, where pathogens may be present at trace levels in complex matrices, such as wastewater.[15,16]

Enrichment strategies range from generic approaches (separating viruses from bacterial and eukaryotic cells) to partially targeted methods (distinguishing ribosomal from messenger RNA),

Table 1. Advantages and limitations of diagnostic techniques of nucleic acid-based methods, spectroscopy, and data analytics

Technology/method	Advantages	Disadvantages/limitations	Agnostic potential
Nucleic acid-based methods			
PCR/RT-PCR	High sensitivity and specificity; rapid; well-established workflows.	Primer redesign needed for new variants.	Low (targeted). Broad-spectrum probes offer semi-agnostic capability for known targets.
Metagenomic amplicon sequencing	High sensitivity; good for microbial community profiling	Primer binding bias and PCR bias, assembly of short sequences (amplicons)	Mid (semi-agnostic). Can capture major groups of pathogens by using primers that bind in conserved regions of target genes.
Metagenomic sequencing (short-read)	High accuracy; good for variant calling	Lower sensitivity (only abundant taxa assemble well); Short reads are suboptimal for complex genome assembly with repeat regions; expensive	High (Agnostic). Can identify the full spectrum of organisms and viruses in a sample without prior knowledge.
Metagenomic sequencing (long-read)	Can produce complete genomes; real-time analysis; portable platforms (e.g., ONT).	Lower sensitivity (only abundant taxa assemble well); Historically higher error rates; requires high-quality input DNA; expensive	High (Agnostic). Excellent for novel pathogen discovery and genome characterization.
Probe capture enrichment	Increases sensitivity for low-abundance targets; reduces sequencing cost/depth required.	Relies on probes for known sequences; potential for bias; not truly agnostic.	Mid (Semi-Agnostic). Enriches for a broad but pre-defined set of pathogens, enhancing agnostic discovery within those groups.
Spectroscopy-based methods			
Multi-wavelength/Hyperspectral Spectroscopy	Reagent-free; non-destructive; rapid (seconds per sample); portable.	Complex signatures require advanced analytics (AI/ML); Lack of standardized spectral libraries; instrument variability.	High (Agnostic). Can detect biochemical shifts caused by any pathogen, known or unknown, without specific reagents.
MALDI-TOF Mass Spectrometry	Rapid bacterial/fungal ID; clinically established.	Requires culture/enrichment; relies heavily on curated databases of known organisms.	Mid (Semi-Agnostic). Identifies organisms and viruses present in a sample, but limited to what exists in the reference database.
Data analytics			
Bioinformatics for sequence analysis	Identifies pathogens, AMR and virulence factors; evolutionary analysis; pathogenicity prediction	Dependent on database completeness; differentiating pathogens from commensals is challenging.	High (Agnostic). Enabling analysis of unknown sequences through genome composition and gene content.
Machine Learning for spectroscopy	Classifies complex spectral data; can identify subtle patterns indicative of infection.	Requires large, high-quality training datasets; risk of overfitting; interpretability can be a challenge.	High (Agnostic). Enabling the classification of samples without needing to identify specific molecular interactions.

to highly specific techniques using targeted probes. The latter forms the foundation for specific detection methods such as PCR, discussed later in discussion.

The attempt to monitor SARS-CoV-2 levels in wastewater represented a renaissance for multiple traditional enrichment methods, including the use of additives such as milk protein, polyethylene glycol, and silica to flocculate the particles from the bulk solution.[17–19] Other approaches, such as filtration and ultrafiltration, have also been successfully used for isolating viral particles from the environment for metagenomic analysis.[20] An affinity-based method by Thermo Fisher Scientific uses paramagnetic beads that are positively charged to attract a range of viral species. Magnetized polymer cages from Ceres were used even prior to the SARS-CoV-2 pandemic for the purpose of trapping viral particles, bacteria, and certain metabolites using affinity dyes.[21,22] Magnetized affinity beads provide a rapid manner of pulling particles of interest out of solution without the use of centrifugation or large laboratory equipment. These approaches are also highly amenable to automation. The cost of reagents, however, is a limiting factor for large-scale use of such methods, and still needs to be addressed.

Collected nucleic acids can be further enriched through the removal of high abundance low-complexity sequences, which can mask low-abundance targets. A common target for depletion is ribosomal RNA (rRNA), for which several commercially available methods exist.[23] Aside from the degradation-based depletion of rRNAs, another approach is to leverage the presence of polyadenylation signals on eukaryotic messenger

RNAs (mRNA) and on many RNA viruses that infect animals and humans to increase the abundance of those sequences and thereby reduce the overall amount of total RNA needed.[24,25] This increase in signal for low-abundance transcripts does come at a cost in terms of additional processing steps, and potential selection of 3′ ends alone.

In addition to concentrating nucleic acids, entire organisms can be enriched through culture-based methods (Figure 1). Culturing pathogens from clinical or environmental samples may amplify a viable target to detectable levels, which is crucial for increasing the sensitivity of downstream whole-genome sequencing. However, this approach has limitations for agnostic detection, as many pathogens are slow-growing, require a host, and are unculturable under standard laboratory conditions, or may be outcompeted by other microbes in the sample. This can lead to a biased view of the microbial community and the potential loss of the true causative agent.

Measurement: From targeted to broad-spectrum polymerase chain reaction to metagenomics

Measurement of the signature can aid in both detection and diagnostic applications. Detection is defined as the identification of a pathogen or associated signatures in diverse environmental samples. These include matrices such as wastewater (which proved a key tool for community-level pathogen monitoring, notably for SARS-CoV-2 during the COVID-19 pandemic), air, water, soil, and others. Diagnosis is the measurement of a pathogen or relevant biomarker signatures in a clinical sample from an infected individual or animal.

Polymerase chain reaction and RT-polymerase chain reaction assays

While conventional polymerase chain reaction (PCR) is a highly targeted method, its principles are foundational, and broad-spectrum assays represent a step toward more agnostic detection. PCR and quantitative reverse transcription PCR (qRT-PCR) assays target genomic regions or genes distinct to known pathogens. Numerous PCR-based detection and diagnostic platforms exist and have been extensively reviewed in the literature,[26] with widespread application in the environmental detection of pathogens and microbes. PCR served as the gold-standard technology for SARS-CoV-2 identification during the pandemic, enabling improved situational awareness and decision-making.[27] Broad-spectrum probes can capture variants more effectively, potentially enabling more robust surveillance, albeit with lower accuracy. For instance, emerging bioinformatic methods such as the FEVER platform[6] enable broad-spectrum identification of viral variants while maintaining pathogen specificity, benefiting population surveillance. This technology was crucial for BioWatch, the DHS platform for biothreat detection in aerosols,[28,29] enabling the identification of known biological threat agents. For environmental surveillance, loop-mediated isothermal amplification (LAMP) has demonstrated sensitivity comparable to traditional RT-qPCR, with reduced inhibition from wastewater contaminants and colorimetric quantification capabilities.[30] Further, LAMP has been successfully combined with CRISPR-mediated detection using SHERLOCK (specific high-sensitivity enzymatic reporter unlocking).[31]

Metagenomics and metatranscriptomics

Metagenomics encompasses the study of complete nucleotide sequences (genomes) from organisms in bulk samples (e.g., wastewater, soil, and feces). In contrast, metatranscriptomics focuses on studying active genes (gene expression) within such communities. Both metagenomic and metatranscriptomic sequencing of environmental samples represent powerful approaches, capable of identifying the full spectrum of organisms present in environmental or clinical samples, including novel or unexpected threats, making them essential tools for comprehensive biosurveillance and clinical diagnostics. However, selecting appropriate sequencing technologies is crucial, particularly in complex environments.

Complex samples contain diverse microbial and biological materials, including single-stranded and double-stranded RNA and DNA viruses, bacteria, and eukaryotic cells. Analysis of such samples requires either pre-sequencing separation, sequence-based enrichment during library preparation, or specific data acquisition strategies. In clinical diagnosis, metagenomic analyses enable the identification of new and emerging pathogens and their variants, particularly when conventional targeted diagnostics fail to identify disease-causing agents in patients with complex symptoms.[32] However, both clinical and environmental samples often contain pathogens at low concentrations, masked by predominant background nucleic acids from hosts or non-pathogens. While deeper sequencing offers one solution, it demands more resources. Alternative approaches include enriching target sequences[33] or depleting unwanted ones.[34]

Metatranscriptomics has shown particular promise for the agnostic detection of RNA viruses in complex samples such as wastewater.[35] However, this evolving method faces ongoing challenges with non-targeted sequencing data. Studies reveal that even after viral sequence enrichment, bacterial RNAs predominate in wastewater samples, with most viral candidates being bacteriophages.[35] Researchers are exploring innovative solutions, such as leveraging polyadenylation or uridylation in eukaryotic viruses,[36,37] for sequence-level enrichment. Since not all viral pathogens possess poly(A) tails, this approach provides only partial enrichment. While enzymatic RNA depletion offers another partial solution, additional methodological developments are clearly needed.

An emerging strategy that utilizes metagenomic and metatranscriptomic approaches is the detection and analysis of extracellular nucleic acids (eNAs). Detection of eRNA indicates actively transcribing pathogens, while eDNA presence can suggest pathogen degradation. While parsing signals from eNAs versus intact microorganisms remains challenging in complex environments such as wastewater, successful applications exist, such as using propidium monoazide to selectively detect viable Campylobacter spp. in food.[38]

An important factor in metagenomics and metatranscriptomics that significantly influences the information obtainable from sequence data is the selection of the sequencing platform. Short-read sequencing platforms (e.g., Illumina NovaSeq) provide low error rates and robust performance for *de novo* and read mapping-based assemblies, making them suitable for detecting new variants of known pathogens.[39] While assembled sequences or unassembled reads can identify known pathogens

and inform functional profiling, short-read data rarely yield complete genomes, especially for repeat-rich pathogen genomes in complex samples. Long-read sequencing technologies, such as Oxford Nanopore Technologies (ONT) and PacBio IsoSeq, address these limitations by generating extended reads that bridge genome assembly breakpoints. Recent improvements in protocols and chemistries have significantly reduced their initially high error rates.[40,41] Long-read assemblies are more likely to yield complete or nearly complete genomes, enabling better strain/species delineation and pathogenicity assessment. Recent innovations include direct RNA sequencing, which streamlines library preparation.[42] While ONT read accuracy has significantly improved, it has historically lagged behind short-read platforms. However, several challenges still remain. These include the requirement for sufficient high-molecular-weight input nucleic acid, potential biases during library preparation that may affect the detection of low-abundance species, and the high computational demands for real-time analysis, which can be a barrier in resource-limited settings. The unique advantages of portability, real-time data access, and long reads that resolve complex genomic regions make ONT platforms highly promising for field-deployable surveillance and rapid characterization,[43] reducing infrastructure requirements for remote biosurveillance and providing a cost-effective alternative to traditional short-read sequencing.

Sequence-based enrichment and background depletion

Substantial variation in pathogen target abundances may result in low-abundance targets being masked by high-abundance ones.[44] A method that can potentially enhance sensitivity and expand genome coverage of selected low-abundance pathogens in a complex sample matrix is selective Whole Genome Amplification (sWGA). The technique uses phi29 DNA polymerase with specific primer sets in a multiple displacement amplification process, generating fragments up to 70-kb from template genomes.[45] Critical to success is the careful design and evaluation of sWGA primer sets to maximize target genome amplification while minimizing non-target DNA amplification. This process is relatively straightforward in clinical samples dominated by a single non-target (human) DNA, where existing primer design pipelines can be applied.[46,47] However, applying broadly targeted (e.g., lineage-specific) sWGA to complex environmental samples such as wastewater requires reconfigured primer design pipelines to avoid significant off-target effects. Combining sWGA with degenerate primers targeting broader pathogen groups could reduce sequencing effort while maintaining accurate genome reconstruction of both known and emerging pathogens.

Probe capture enrichment represents another semi-agnostic but powerful method for virus sequencing in complex sample matrices. This sequence-specific approach utilizes extensive panels of probes, which can be single-stranded DNA/RNA or, as in some commercial offerings, double-stranded DNA, designed to target conserved genomic regions. The process involves hybridizing biotinylated probes with the sample during library preparation, followed by capture of probe-target complexes using streptavidin-coated magnetic beads. After washing to remove nonspecific or unbound sequences, targets are eluted by disrupting the biotin-streptavidin interaction and, if necessary, amplified using random primers. This method has demonstrated effectiveness across various applications, including viral genome analysis,[48,49] pathogen detection in wastewater samples,[50,51] and functional targeted metagenomics.[52,53] Numerous protocols and commercial kits are available, such as the Twist Comprehensive Viral Research Panel (Twist Bioscience, San Francisco, CA), which features over 1 million unique double-stranded DNA probes derived from more than 3,100 human and animal viral genomes.

Metagenomic and transcriptomic sensitivity can be further enhanced by reducing non-target sequences through various approaches. Selective lysis and host DNA depletion have proven effective in clinical samples dominated by human DNA,[54,55] though this method only reduces cells susceptible to weak osmotic pressure (e.g., mammalian cells) and may affect viral DNA.[56] For sWGA, off-target sequence amplification can be mitigated using blocking agents specifically designed to bind problematic non-target sequences, with 3′ modifications (e.g., C3 spacer) suppressing prevalent off-target DNA amplification.[57] Similar to probe capture enrichment, biotinylated probes can be designed to capture and remove abundant off-target sequences from samples. An alternative approach leverages Cas9 combined with specific guide RNAs, to deplete unwanted, known high-abundance sequences (e.g., host DNA and ribosomal RNA genes), thereby increasing the relative proportion and detection sensitivity of remaining low-abundance pathogen sequences.

It is noteworthy to acknowledge that while untargeted and enrichment/depletion strategies aim for comprehensive or agnostic detection, biases can be introduced at nearly every step from sample collection through data analysis, meaning truly unbiased representation remains a challenging but achievable goal for current technologies.

Bioinformatics for sequence-based detection

Data analytics plays a crucial role in modern pathogen identification and pathogenicity prediction. Nucleic acid-based methods enable the discriminative identification of cellular and viral pathogens through sequence similarity comparisons with known pathogens in existing databases, such as those maintained by NCBI (e.g., GenBank, non-redundant, Pathogen Detection portal) and the Bacterial and Viral Bioinformatics Resource Center (BV-BRC). Analytical approaches range from sequence homology searches of reads or contigs against pathogen genome databases to metrics such as average nucleotide identity (ANI) or average amino acid identity (AAI) applied to assembled genomes or large fragments for taxonomic classification. Despite inherent limitations, these bioinformatic analyses provide critical insights into pathogen identity, community composition within a sample, functional potential (including the presence of antimicrobial resistance (AMR) and virulence factors), and pathogen evolution, which are essential for surveillance and outbreak response.

Key limitations of these approaches include the general dependence on database completeness for all taxa, the challenge of differentiating true pathogens from background commensal microbiota in complex samples, and accurately interpreting low-abundance hits which may represent early-stage infections or contaminants. Current databases typically contain genomes and/or conserved marker genes such as 16S rRNA,

18S rRNA, or viral RNA-dependent RNA polymerase genes.[58] However, major pathogen databases are primarily limited to experimentally verified (disease-associated) pathogens and may not be well suited to identify novel phylogenetically unrelated pathogens. This limitation highlights the need for expanded databases and sequence repositories capable of identifying patterns that predict novel and unprecedented pathogenicity signatures. Ideal databases would incorporate comprehensive, curated metadata on pathogenicity mechanisms, zoonoses, and host susceptibility.

The increasing availability of genomic and phenotypic information in public databases has spurred development of various supervised machine learning models for pathogenic potential prediction.[59] These approaches, trained on protein content or nucleic acid sequences, have demonstrated promising results in pathogenicity determination, though accuracy varies when analyzing incomplete metagenomic data or unknown genomes.[60–63] However, disease association cannot always be inferred from genetic sequences alone, as pathogenicity often results from complex interactions beyond genetic makeup. Comprehensive genome-to-phenome studies are essential for accurate pathogenicity prediction. Multiple factors influence pathogenicity, including environmental conditions such as altered climate, critical biomass requirements, life cycle modifications, host adaptations, vector ecology and other environmental factors.

SPECTROSCOPY-BASED AGNOSTIC DETECTION

Genomic techniques largely target nucleic acids. However, living systems are not only composed of nucleic acids, but also contain proteins, carbohydrates, metabolites, lipids, and other biomolecules. Interrogating signatures associated with these biochemically disparate molecules can provide valuable insights for early detection and diagnosis of infection/threat. Thus, use of alternative investigative approaches that target such biomolecules can provide additional, or even unique clarity on emerging threats.

Spectroscopy is defined as the investigation and measurement of signatures associated with the interaction of matter with electromagnetic radiation.[64] Such methods can be based on absorption, emission, or emanation of fluorescence upon interaction with the radiation. All biomolecules interact with electromagnetic radiation, albeit at different wavelengths across the spectrum and in varied ways. Measurement or characterization of these interactions can provide unique insights into the nature of the biological entity, as well as facilitate matrix (background) characterization. Based on the choice of the method, spectroscopy can also allow for reagent-free, label-free, non-destructive characterization of biomolecules, which can be extremely valuable for early diagnostics and surveillance of emerging threats.[65–67] The choice of the spectroscopic method also dictates the nature of the information generated. For instance, in vibrational spectroscopy, transmitted light is absorbed by the molecular bonds in the sample constituents at different wavelengths, resulting in an absorption pattern that is measurable by a detector.[68] The composition of the sample results in a unique spectral signature. While each of the aforementioned methods provides distinctive spectral information, they are indicative of only limited biomolecules. Integrating information across the electromagnetic spectrum can provide more holistic information on biosignatures in complex samples, and consequently, multi-wavelength spectroscopy holds much promise in this regard.

Spectral signatures are extremely sensitive and specific - however, small permutations in background or biochemistry can change the spectral profiles. Before the advent of machine learning, characterization of such signatures and associated changes was extremely intensive and, therefore, time-consuming. Thus, the use of spectral sensing for pathogen-agnostic detection and diagnostics has been limited.[67] However, with the emergence of machine learning and the advent of advanced remote sensing capabilities, we are uniquely positioned to realize the full potential of this reagent-free signature regime, especially when we do not limit ourselves to specific wavelengths. So far, largely, spectral sensing for detection and diagnostic applications - i.e., for biosignature identification - has operated either in specific wavelength regimes (e.g., infrared, optical, UV, and other), or used ligands to address the specificity of measured targets (fluorescence spectroscopy). For instance, SARS-CoV-2 viral particles could be detected from nasal swab suspensions and saliva using near-infrared spectroscopy and validated with RT-PCR.[69–71] Mukundan and colleagues have applied optical spectroscopy, with fluorescence-based ligands for specificity, for the detection and diagnosis of pathogens in a variety of diseases using ultra-sensitive waveguide-based platforms.[72,73] However, these methods are more targeted - akin to the PCR-based methods for nucleic acids- and need to be redesigned and redeveloped for every new pathogen. It is important to note that MALDI-TOF mass spectrometry, which analyzes protein profiles, has achieved widespread clinical success for bacterial/fungal identification. However, we classify it as semi-agnostic because its strength lies in rapidly matching a sample's unique protein fingerprint to a library of known organismal fingerprints. While it can agnostically screen a sample for any organism in its database, it is less suited for identifying a novel pathogen that is not in the database, compared to metagenomics, which can assemble a complete, unknown genome that can then be characterized by its constituent genes and prediction functions. Agnostic application of spectral sensing is a more daunting goal in the biological realm, although use cases in space-science (ChemCam and SuperCam instruments on Mars Rovers) and in remote sensing exist, largely for non-biological targets. More recently, there has been a surge in multi-wavelength and hyperspectral sensing in the detection and diagnostics realm as well, allowing for integrated signatures across the electromagnetic spectrum. Needless to say, the advent of expedited data analytic methods, computational tools, and machine learning has catapulted the real-time usability of these data-heavy methods. Indeed, recently, multi-wavelength spectroscopy was demonstrated to provide near real-time identification of SARS-CoV-2 metabolites in saliva samples (3 s/sample), without reagents or sample processing, in a blinded clinical study.[67] Previous work has shown that after developing a baseline in 5 min, the speed of spectral measurements can take as little as 0.18 s[74] However, more work is required to determine the specificity of such modalities

and the reproducibility and inter- and intra-assay variability associated with the use of machine learning in biological systems.[75] Overall, the development of portable, fast spectrometers will be a key factor that allows for the widespread deployment of multi-wavelength spectroscopy closer to the wastewater location. Some challenges with spectroscopy include the complexity of spectral signatures from heterogeneous matrices, the lack of comprehensive and standardized spectral reference libraries (especially for microbes in complex backgrounds), and variability between instruments hindering standardization, the need for large, high-quality datasets for training machine learning models, and historical limitations in speed and specificity/sensitivity, particularly with portable devices.

Beyond biosignatures, spectroscopy can also provide further information on the matrix/background, which can be extremely relevant for source identification and characterization of threat. For instance, spectroscopy has been shown in different use cases for characterizing complex environmental samples, e.g., wastewater and clinical biofluids, thereby demonstrating relevance to both detection and diagnostic applications. Hand-held spectrometers at distinctive wavelengths are used to measure characteristics such as chemical oxygen demand, total suspended solids, and biochemical oxygen demand[76] in wastewater. Using analytical methods such as partial least-squares regression, investigators were able to accurately recapitulate validated measurements from the spectra, demonstrating the promise of such methods. There have been limited studies using spectroscopy to detect viruses in complex backgrounds in environmental samples such as wastewater.[77]

Sample processing and enrichment for spectroscopy

Several well-characterized methods, both laboratory and field-based, exist for processing samples for specific biomarkers such as proteins, carbohydrates, lipids, and others. It must be noted that several methods, such as immunoassays based on urine or saliva, do not require any pre-processing of the samples, primarily because of the use of ligands that impart outcome specificity. Enriching proteins is crucial for accessing functional information and unique biomarkers, as proteins indicate active biological processes, unlike nucleic acids, which primarily show potential. This is highly relevant to this review, as protein analysis complements nucleic acid data for a more comprehensive, agnostic threat assessment. Both generic and custom methods are employed for protein concentration and enrichment. Generic methods described for the concentration or desalting of peptides vary based on the target application, purity requirements, and scale. These methods include protein concentration from heterogeneous samples through sedimentation,[78] precipitation using salts such as ammonium sulfate,[78] and precipitation with acetone.[79] Filtration provides a more versatile approach for concentration and solute removal. Ultrafiltration using selective membranes enables protein concentration through low-speed centrifugation or vacuum-based methods. Commercial systems such as EDM Millipore Amicons, use porous regenerated cellulose membranes to retain biomolecules based on molecular size, accommodating various input/output volumes and concentrating diverse solutes (proteins, nucleic acids, viral particles). Outside of small-scale applications, ultrafiltration is an industrially relevant process offering precision separation for food components such as milk proteins and other proteinaceous components that are recovered from waste.[80,81] The adaptability of membrane technology to vacuum-based filtration enhances flexibility and portability for field applications and deployable biosurveillance.

For more specific enrichment, affinity tags offer targeted approaches for various applications. They are used for the purification of recombinant protein fusions; however, this requires a validated design for a heterologous expression platform.[82] The development of targeted antibodies that recognize and bind label-free proteins in complex samples has enabled more selective enrichment. Newer alternatives such as ProImmune's Ankyrons provide animal-free, *in vitro* screened options that bind targets similarly to immunoglobulins, offering both standardized and custom design.[83] These affinity and antibody-based methods are particularly valuable when high purity is required.

Data analytics for spectral signatures

As with nucleic acid methods, data analytics are a critical element for the success and applicability of these tools for detection and diagnostics. Converting spectral measurements into the detection of viruses, various machine learning and artificial intelligence approaches can be used to build models. Prior to training models, the spectra generated are typically preprocessed to remove artifacts and increase the signal-to-noise ratio. These methods include normalization, baseline correction, smoothing, and calculating the first or second derivative of the data.[84] Following potential preprocessing, several models, such as partial least squares,[85] random forest,[86] or deep learning approaches have been used to train models.[87] In addition to building the model, understanding the important wavelengths for the models is useful for interpreting what the models are learning. There are several resources that correlate wavelengths with bonds, lipids, proteins, or other molecular components. With interpretable models, the types of molecular structures can be found, even in complex samples, to try to evaluate what the model is learning. With this approach, previous work has shown that identifying viruses, in various backgrounds, with spectroscopy is possible, and we believe applying a spectroscopic device with sufficient resolution for the detection of viruses is feasible.[88]

Databases of spectra exist, but focus primarily on singular compounds (SpectraBase.com). However, these spectra only represent compounds that are small and highly purified. Currently, no database of spectra exists of large structures in complex backgrounds such as viruses or bacteria in wastewater. One challenge is that data are not standardized. Each spectrometer has its own range of wavelengths where spectra are measured, and at different resolutions. Previous work has attempted to address this by developing standardized data formats, but these formats are not widespread across all spectroscopic applications.[89] Another key limitation is the speed of spectroscopy. Of the aforementioned spectroscopy papers, the fastest platforms take more than a minute for a single sample. While several technological advances in spectrometers have resulted in greater spectral resolution, this also means that the number of features easily outpaces the number of

samples that can be generated. More advanced machine learning and artificial intelligence approaches may help to overcome the first problem. Yet, the lack of data standardization and inability to develop datasets much larger than the number of features can make it difficult to develop robust machine learning models.[90] While essential for analyzing complex spectra, artificial intelligence and machine learning models cannot reliably predict novel signatures absent from training data, underscoring the need for foundational spectral data generation. Indeed, previous work from our team and others has called out the need for standardization, especially when applying machine learning classifiers for the identification of biological patterns in complex samples.[91] Thus, it is important to ensure systematic, standardized, reproducible machine learning methods to derive physiologically relevant information for decision-making.

The critical bottleneck remains the development of extensive, standardized spectral libraries representing pathogen diversity within relevant complex matrices, which are currently scarce. The development of spectral libraries containing measurements of known concentrations of biomarkers and signatures (viruses, background matrices) in various complex backgrounds (e.g., wastewater) is relevant to using these methods broadly. While we understand that this database will be ever-changing, given the transience of living communities, and flow-induced dynamics - collection of large datasets will allow for the systematic capture of uncertainties within such systems, which is a critical element to such analytic pipelines. More broadly, there is a need to develop robust models and datasets based on the large amount of spectroscopic data that is already available. Previous work has shown AI/ML methods such as transfer learning, can be used for inference from models trained on different datasets and backgrounds.[92] With the latest advancements in AI/ML, we believe that broadly trained foundational models for spectroscopy may be able to help provide tools that can fast-track the adoption of spectroscopy for pathogen-agnostic detection.

CONCLUSION

Timely development and deployment of detection and diagnostic tools is crucial for reducing outbreak-associated mortality and morbidity.[93] Even in routine healthcare settings, effective and timely diagnostics are essential for tailored responses and addressing persistent threats such as antimicrobial resistance.[94] While targeted diagnostics contribute to this goal, they have significant limitations: 1) inability to detect unanticipated and emerging threats, and 2) longer development time during outbreaks.[94] This underscores the critical need for pathogen-agnostic platform technologies that enhance preparedness for both expected and unanticipated biological events.[95] In this article, we systematically explored the different elements and gaps in the use of two agnostic detection/diagnostic modalities - sequencing and spectroscopy.

Sequencing-based and spectroscopic methods show particular promise in this context.[67,96]

The two methods are in highly different stages of maturity as far as agnostic biodetection is concerned, offering a broad spectrum for the discussion of factors contributing to the translation of these technologies from design to deployment. Sequencing-based methods have demonstrated effectiveness in monitoring known threats in complex environmental samples, particularly wastewater. The current challenge lies in transitioning from detecting known/anticipated pathogens to predictively assessing emerging threats. Spectroscopy, on the other hand, has only been limitedly used in this context; but recent evolution in capabilities and machine learning offer great promise for more widespread use in the near future. Indeed, the advancement of remote sensing methods has enabled reagent-free and rapid situational awareness and signature discovery in complex environments. Thus, there is significant potential to enable agnostic biodetection and diagnostics, which can significantly advance responsiveness during unanticipated outbreaks - and as such, should be a focus of future research investment. Such technologies can also enable warfighter support in remote areas of the world via improved situational awareness and decision making, and impact routine care in clinics and hospitals.[95]

Whether it is sequencing, spectroscopy, or a completely different method, the rapid use of innovative technologies in a real-world situation requires standardization and streamlining - that is a pre-designed resilient architecture for detection/diagnosis. Standardized protocols, data collection, sharing, pipelines, manufacturing, deployment, and finally, decision-making are all elements of such an architecture. Ultimately, there is also a need for an authoritative agency that can enable deployment and associated decision making, which can greatly mitigate mortality and morbidity associated with outbreaks. For example, pathogens can be transmitted through air, water, soil, animals (zoonoses), or other sources. Thus, the choice of the sample can be a significant variable, and optimization of protocols and practices across multiple sample types is critical for resilience. Indeed, the impact of wastewater surveillance was realized during the COVID-19 pandemic, as was the significance of rapid in-home testing via lateral flow immunoassays. If the usability of such methods can be further shifted to the left of the outbreak curve via the use of agnostic modalities that do not depend on specific ligands or reagents, then the impact of such technologies will definitely be further amplified. Filtered air from modern buildings (airports, commercial centers) could provide particles for sequencing and detection, covering both static locations and high-flux zones with international traffic. The DHS's BioWatch program demonstrated biothreat agent detection in air samples, and with advanced agnostic sequencing, spectral methods, and AI/ML capabilities, can now encompass an expanded pathogen range. This is also true for diagnostics, where currently we rely heavily on targeted methods such as lateral flow immunoassays and PCR for rapid testing. Further optimization of untargeted methods for early information can significantly allow for healthcare decisions that can save lives. This is also significant for biodefense applications and in normal routine health care practice.

A key element of our response architecture is data - we strongly advocate for collecting and curating data from annual and occasional outbreaks and pandemics in a streamlined and systematic manner to enable predictive analytics, pattern-based assessments, and future AI-readiness. All in all, we believe that an integrated decision tree connecting pathogen-agnostic detection from environmental and clinical matrices with

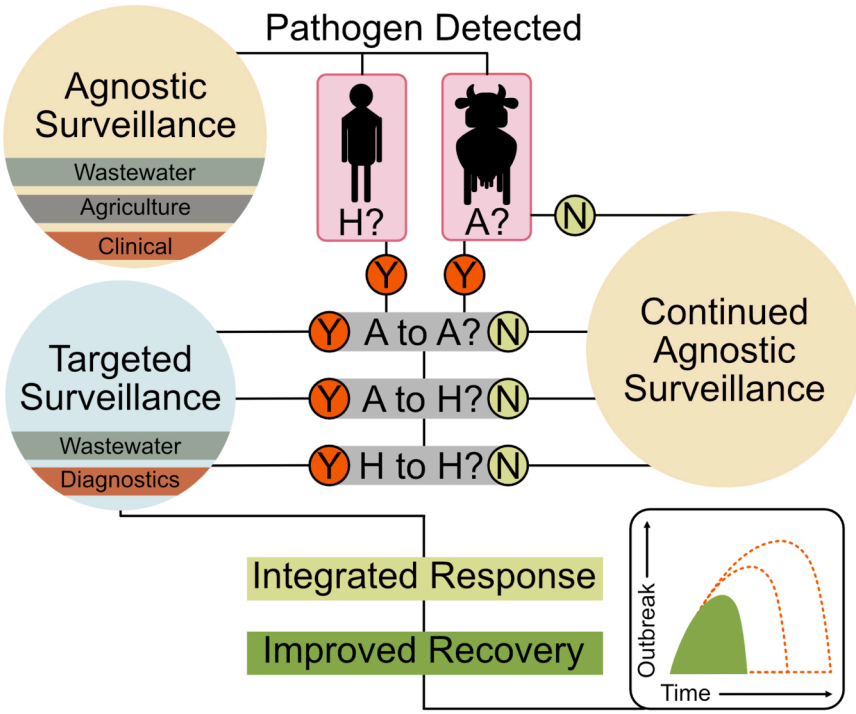

Figure 2. Integration of pathogen-agnostic detection into strategies for outbreak response

Agnostic surveillance of environmental samples can provide early warning of pathogens circulating in animal or human populations within monitored geographic areas. When a pathogen is detected in both environmental samples and animal or human populations, targeted assays should be rapidly developed for effective, frequent monitoring. This becomes particularly crucial when evidence emerges of animal-to-human or human-to-human transmission. The ultimate goal of early warning systems is to enable an integrated response that accelerates recovery and limits outbreak duration, thereby preventing progression to pandemic scale.

appropriate response measures is essential, impacting various stages from prediction to social regulation implementation (Figure 2). As depicted in the figure, this implementation strategy begins with broad agnostic surveillance to provide early warning, which then triggers the development of targeted assays for more focused monitoring once a threat is identified and characterized, thereby enabling a swift and efficient public health response.

While key elements for surveillance and response infrastructure exist, technological gaps require strategic investment. Critical priorities include framework development, integration, methodology standardization, and creating machine learning-compatible data repositories, as well as the integration of global efforts to enable predictive assessments. Established domestic workflows could serve as blueprints for field-deployable biomonitoring, adaptable to remote locations, including military conflict zones and natural disaster areas, enabling defense missions as well. Domestic and global biosurveillance implementation will provide insights into unforeseen challenges and opportunities for improving sensitivity, accuracy, and cost-effectiveness over time. Indeed, the high cost on the economy with every lost productive day, and the mental and physical duress of humankind can be mitigated with the scientific innovation and standardization of processes.

ACKNOWLEDGMENTS

This work was supported by the Defense Threat Reduction Agency "Pathogen-agnostic Detection of Evolving and Newly Emerging Threats" (Principal Investigator: Schulz, Project Manager: Dr. Francesconi) at the U.S. Department of Energy Joint Genome Institute (JGI) (https://ror.org/04xm1d337). The work conducted by the JGI, a DOE Office of Science User Facility, is supported by the Office of Science of the U.S. Department of Energy operated under Contract No. DE-AC02-05CH11231.

DECLARATION OF INTERESTS

Frederik Schulz is CEO of SampleX. SampleX and its activities are unrelated to the work presented in this article. All other authors declare no competing interests.

DECLARATION OF GENERATIVE AI AND AI-ASSISTED TECHNOLOGIES IN THE WRITING PROCESS

During the preparation of this work, the authors used Gemini v3.0 and ChatGPT 5.1 in order to perform the final proofreading. After using this tool/service, the authors reviewed and edited the content as needed and took full responsibility for the content of the publication.

REFERENCES

1. Wu, F., Zhao, S., Yu, B., Chen, Y.-M., Wang, W., Song, Z.-G., Hu, Y., Tao, Z.-W., Tian, J.-H., Pei, Y.-Y., et al. (2020). A new coronavirus associated with human respiratory disease in China. Nature *579*, 265–269.
2. Shi, W., Cai, Y., Zhu, H., Peng, H., Voyer, J., Rits-Volloch, S., Cao, H., Mayer, M.L., Song, K., Xu, C., et al. (2023). Cryo-EM structure of SARS-CoV-2 postfusion spike in membrane. Nature *619*, 403–409.
3. Huang, Y., Yang, C., Xu, X.-F., Xu, W., and Liu, S.-W. (2020). Structural and functional properties of SARS-CoV-2 spike protein: potential antivirus drug development for COVID-19. Acta Pharmacol. Sin. *41*, 1141–1149.
4. Vitiello, A., La Porta, R., Trama, U., Ferrara, F., Zovi, A., Auti, A.M., Di Domenico, M., and Boccellino, M. (2022). Pandemic COVID-19, an update of current status and new therapeutic strategies. Naunyn-Schmiedebergs Arch Pharmacol *395*, 1159–1165.

5. Okba, N.M.A., Müller, M.A., Li, W., Wang, C., GeurtsvanKessel, C.H., Corman, V.M., Lamers, M.M., Sikkema, R.S., de Bruin, E., Chandler, F.D., et al. (2020). SARS-CoV-2 specific antibody responses in COVID-19 patients. Preprint at bioRxiv. https://doi.org/10.1101/2020.03.18.20038059.
6. Stromberg, Z.R., Theiler, J., Foley, B.T., Myers Y Gutiérrez, A., Hollander, A., Courtney, S.J., Gans, J., Deshpande, A., Martinez-Finley, E.J., Mitchell, J., et al. (2022). Fast Evaluation of Viral Emerging Risks (FEVER): A computational tool for biosurveillance, diagnostics, and mutation typing of emerging viral pathogens. PLoS Glob. Public Health *2*, e0000207.
7. Xiao, K., and Zhang, L. (2023). Wastewater pathogen surveillance based on One Health approach. Lancet Microbe *4*, e297.
8. Cabral, J.P.S. (2010). Water Microbiology. Bacterial Pathogens and Water. Int. J. Environ. Res. Public Health *7*, 3657–3703.
9. Bintsis, T. (2017). Foodborne pathogens. AIMS Microbiol. *3*, 529–563.
10. Remy, R., Kemnitz, N., Trefz, P., Fuchs, P., Bartels, J., Klemenz, A.-C., Rührmund, L., Sukul, P., Miekisch, W., and Schubert, J.K. (2022). Profiling of exhaled volatile organics in the screening scenario of a COVID-19 test center. iScience *25*, 105195.
11. Sah, R., Srivastava, S., Kumar, S., Mehta, R., Donovan, S., Sierra-Carrero, L., Luna, C., Woc-Colburn, L., Cardona-Ospina, J.A., Hinestroza-Jordan, M., et al. (2024). Concerns on H5N1 avian influenza given the outbreak in U.S. dairy cattle. Lancet Reg. Health. Am. *35*, 100785.
12. Lenz, K.D., Jakhar, S., Chen, J.W., Anderson, A.S., Purcell, D.C., Ishak, M.O., Harris, J.F., Akhadov, L.E., Kubicek-Sutherland, J.Z., Nath, P., and Mukundan, H. (2021). A centrifugal microfluidic cross-flow filtration platform to separate serum from whole blood for the detection of amphiphilic biomarkers. Sci. Rep. *11*, 5287.
13. Qin, D. (2019). Next-generation sequencing and its clinical application. Cancer Biol. Med. *16*, 4–10.
14. Sun, C., Zhang, B., Ning, D., Zhang, Y., Dai, T., Wu, L., Li, T., Liu, W., Zhou, J., and Wen, X. (2021). Seasonal dynamics of the microbial community in two full-scale wastewater treatment plants: diversity, composition, phylogenetic group based assembly and co-occurrence pattern. Water Res. *200*, 117295.
15. Wölfel, R., Corman, V.M., Guggemos, W., Seilmaier, M., Zange, S., Müller, M.A., Niemeyer, D., Jones, T.C., Vollmar, P., Rothe, C., et al. (2020). Virological assessment of hospitalized patients with COVID-2019. Nature *581*, 465–469.
16. Robinson, C.A., Hsieh, H.-Y., Hsu, S.-Y., Wang, Y., Salcedo, B.T., Belenchia, A., Klutts, J., Zemmer, S., Reynolds, M., Semkiw, E., et al. (2022). Defining biological and biophysical properties of SARS-CoV-2 genetic material in wastewater. Sci. Total Environ. *807*, 150786.
17. Salvo, M., Moller, A., Alvareda, E., Gamazo, P., Colina, R., and Victoria, M. (2021). Evaluation of low-cost viral concentration methods in wastewaters: Implications for SARS-CoV-2 pandemic surveillances. J. Virol. Methods *297*, 114249.
18. Gonzales-Gustavson, E., Cárdenas-Youngs, Y., Calvo, M., da Silva, M.F.M., Hundesa, A., Amorós, I., Moreno, Y., Moreno-Mesonero, L., Rosell, R., Ganges, L., et al. (2017). Characterization of the efficiency and uncertainty of skimmed milk flocculation for the simultaneous concentration and quantification of water-borne viruses, bacteria and protozoa. J. Microbiol. Methods *134*, 46–53.
19. Yanaç, K., Adegoke, A., Wang, L., Uyaguari, M., and Yuan, Q. (2022). Detection of SARS-CoV-2 RNA throughout wastewater treatment plants and a modeling approach to understand COVID-19 infection dynamics in Winnipeg, Canada. Sci. Total Environ. *825*, 153906.
20. Roux, S., Matthijnssens, J., and Dutilh, B.E. (2021). Metagenomics in virology. Encycl. Virol., 133–140. https://doi.org/10.1016/B978-0-12-809633-8.20957-6.
21. Ii, A.N., Lin, S.-C., Lepene, B., Zhou, W., Kehn-Hall, K., and van Hoek, M.L. (2021). Use of magnetic nanotrap particles in capturing Yersinia pestis virulence factors, nucleic acids and bacteria. J. Nanobiotechnology *19*, 186.
22. Shafagati, N., Patanarut, A., Luchini, A., Lundberg, L., Bailey, C., Petricoin, E., 3rd, Liotta, L., Narayanan, A., Lepene, B., and Kehn-Hall, K. (2014). The use of Nanotrap particles for biodefense and emerging infectious disease diagnostics. Pathog. Dis. *71*, 164–176.
23. Adiconis, X., Borges-Rivera, D., Satija, R., DeLuca, D.S., Busby, M.A., Berlin, A.M., Sivachenko, A., Thompson, D.A., Wysoker, A., Fennell, T., et al. (2013). Comparative analysis of RNA sequencing methods for degraded or low-input samples. Nat. Methods *10*, 623–629.
24. Kuai, L., Fang, F., Butler, J.S., and Sherman, F. (2004). Polyadenylation of rRNA in Saccharomyces cerevisiae. Proc. Natl. Acad. Sci. USA *101*, 8581–8586.
25. Haile, S., Corbett, R.D., Bilobram, S., Mungall, K., Grande, B.M., Kirk, H., Pandoh, P., MacLeod, T., McDonald, H., Bala, M., et al. (2019). Evaluation of protocols for rRNA depletion-based RNA sequencing of nanogram inputs of mammalian total RNA. PLoS One *14*, e0224578.
26. Yang, S., and Rothman, R.E. (2004). PCR-based diagnostics for infectious diseases: uses, limitations, and future applications in acute-care settings. Lancet Infect. Dis. *4*, 337–348.
27. Dutta, D., Naiyer, S., Mansuri, S., Soni, N., Singh, V., Bhat, K.H., Singh, N., Arora, G., and Mansuri, M.S. (2022). COVID-19 diagnosis: A comprehensive review of the RT-qPCR method for detection of SARS-CoV-2. Diagnostics *12*, 1503.
28. National Research Council, Institute of Medicine, Board on Health Sciences Policy, Division on Earth and Life Studies, Board on Life Sciences, and Committee on PCR Standards for the BioWatch Program (2015). BioWatch PCR assays: Building confidence, ensuring reliability: Abbreviated version (National Academies Press).
29. National Academies of Sciences, Engineering, and Medicine (2017). Enhancing BioWatch Capabilities Through Technology and Collaboration. In Proceedings of a Workshop. Washington, DC (The National Academies Press). https://doi.org/10.17226/6557.
30. Spiteri, S., Marino, F., Girolamini, L., Pascale, M.R., Derelitto, C., Caligaris, L., Paghera, S., and Cristino, S. (2024). Loop-mediated isothermal amplification (LAMP): An innovative approach for the environmental monitoring of SARS-CoV-2. Pathogens *13*, 1022. https://doi.org/10.3390/pathogens13111022.
31. Joung, J., Ladha, A., Saito, M., Kim, N.-G., Woolley, A.E., Segel, M., Barretto, R.P.J., Ranu, A., Macrae, R.K., Faure, G., et al. (2020). Detection of SARS-CoV-2 with SHERLOCK one-pot testing. N. Engl. J. Med. *383*, 1492–1494.
32. Gu, W., Deng, X., Lee, M., Sucu, Y.D., Arevalo, S., Stryke, D., Federman, S., Gopez, A., Reyes, K., Zorn, K., et al. (2021). Rapid pathogen detection by metagenomic next-generation sequencing of infected body fluids. Nat. Med. *27*, 115–124.
33. Gaudin, M., and Desnues, C. (2018). Hybrid capture-based next generation sequencing and its application to human infectious diseases. Front. Microbiol. *9*, 2924.
34. Parris, D.J., Kariithi, H., and Suarez, D.L. (2022). Non-target RNA depletion strategy to improve sensitivity of next-generation sequencing for the detection of RNA viruses in poultry. J. Vet. Diagn. Invest. *34*, 638–645.
35. Spurbeck, R.R., Catlin, L.A., Mukherjee, C., Smith, A.K., and Minard-Smith, A. (2023). Analysis of metatranscriptomic methods to enable wastewater-based biosurveillance of all infectious diseases. Front. Public Health *11*, 1145275.
36. Barr, J.N., and Fearns, R. (2010). How RNA viruses maintain their genome integrity. J. Gen. Virol. *91*, 1373–1387.
37. Huo, Y., Shen, J., Wu, H., Zhang, C., Guo, L., Yang, J., and Li, W. (2016). Widespread 3'-end uridylation in eukaryotic RNA viruses. Sci. Rep. *6*, 25454.
38. Okada, A., Tsuchida, M., Rahman, M.M., and Inoshima, Y. (2022). Two-round treatment with propidium monoazide completely inhibits the detection of dead Campylobacter spp. Cells by quantitative PCR. Front. Microbiol. *13*, 801961.

39. Van Poelvoorde, L.A.E., Delcourt, T., Coucke, W., Herman, P., De Keersmaecker, S.C.J., Saelens, X., Roosens, N.H.C., and Vanneste, K. (2021). Strategy and performance evaluation of low-frequency variant calling for SARS-CoV-2 using targeted deep Illumina sequencing. Front. Microbiol. *12*, 747458.

40. Zee, A., Deng, D.Z.Q., Adams, M., Schimke, K.D., Corbett-Detig, R., Russell, S.L., Zhang, X., Schmitz, R.J., and Vollmers, C. (2022). Sequencing Illumina libraries at high accuracy on the ONT MinION using R2C2. Genome Res. *32*, 2092–2106.

41. Karst, S.M., Ziels, R.M., Kirkegaard, R.H., Sørensen, E.A., McDonald, D., Zhu, Q., Knight, R., and Albertsen, M. (2021). High-accuracy long-read amplicon sequences using unique molecular identifiers with Nanopore or PacBio sequencing. Nat. Methods *18*, 165–169.

42. Viehweger, A., Krautwurst, S., Lamkiewicz, K., Madhugiri, R., Ziebuhr, J., Hölzer, M., and Marz, M. (2019). Direct RNA nanopore sequencing of full-length coronavirus genomes provides novel insights into structural variants and enables modification analysis. Genome Res. *29*, 1545–1554.

43. Ciuffreda, L., Rodríguez-Pérez, H., and Flores, C. (2021). Nanopore sequencing and its application to the study of microbial communities. Comput. Struct. Biotechnol. J. *19*, 1497–1511.

44. Kantor, R.S., and Jiang, M. (2024). Considerations and opportunities for probe capture enrichment sequencing of emerging viruses from wastewater. Environ. Sci. Technol. *58*, 8161–8168.

45. Leichty, A.R., and Brisson, D. (2014). Selective whole genome amplification for resequencing target microbial species from complex natural samples. Genetics *198*, 473–481.

46. Dwivedi-Yu, J.A., Oppler, Z.J., Mitchell, M.W., Song, Y.S., and Brisson, D. (2023). A fast machine-learning-guided primer design pipeline for selective whole genome amplification. PLoS Comput. Biol. *19*, e1010137.

47. Clarke, E.L., Sundararaman, S.A., Seifert, S.N., Bushman, F.D., Hahn, B.H., and Brisson, D. (2017). Swga: A primer design toolkit for selective whole genome amplification. Bioinformatics *33*, 2071–2077.

48. Briese, T., Kapoor, A., Mishra, N., Jain, K., Kumar, A., Jabado, O.J., and Lipkin, W.I. (2015). Virome capture sequencing enables sensitive viral diagnosis and comprehensive virome analysis. mBio *6*, e01491-15.

49. Wylie, T.N., Wylie, K.M., Herter, B.N., and Storch, G.A. (2015). Enhanced virome sequencing using targeted sequence capture. Genome Res. *25*, 1910–1920.

50. Jiang, M., Wang, A.L.W., Be, N.A., Mulakken, N., Nelson, K.L., and Kantor, R.S. (2024). Evaluation of the impact of concentration and extraction methods on the targeted sequencing of human viruses from wastewater. Environ. Sci. Technol. *58*, 8239–8250.

51. Williams, R.C., Farkas, K., Garcia-Delgado, A., Adwan, L., Kevill, J.L., Cross, G., Weightman, A.J., and Jones, D.L. (2024). Simultaneous detection and characterization of common respiratory pathogens in wastewater through genomic sequencing. Water Res. *256*, 121612.

52. Baba, H., Kuroda, M., Sekizuka, T., and Kanamori, H. (2023). Highly sensitive detection of antimicrobial resistance genes in hospital wastewater using the multiplex hybrid capture target enrichment. mSphere *8*, e0010023.

53. Kushwaha, S.K., Manoharan, L., Meerupati, T., Hedlund, K., and Ahrén, D. (2015). MetCap: a bioinformatics probe design pipeline for large-scale targeted metagenomics. BMC Bioinf. *16*, 65.

54. Gan, M., Wu, B., Yan, G., Li, G., Sun, L., Lu, G., and Zhou, W. (2021). Combined nanopore adaptive sequencing and enzyme-based host depletion efficiently enriched microbial sequences and identified missing respiratory pathogens. BMC Genom. *22*, 732.

55. Wu-Woods, N.J., Barlow, J.T., Trigodet, F., Shaw, D.G., Romano, A.E., Jabri, B., Eren, A.M., and Ismagilov, R.F. (2023). Microbial-enrichment method enables high-throughput metagenomic characterization from host-rich samples. Nat. Methods *20*, 1672–1682.

56. Oechslin, C.P., Lenz, N., Liechti, N., Ryter, S., Agyeman, P., Bruggmann, R., Leib, S.L., and Beuret, C.M. (2018). Limited correlation of shotgun metagenomics following host depletion and routine diagnostics for viruses and bacteria in low concentrated surrogate and clinical samples. Front. Cell. Infect. Microbiol. *8*, 375.

57. Vestheim, H., and Jarman, S.N. (2008). Blocking primers to enhance PCR amplification of rare sequences in mixed samples - a case study on prey DNA in Antarctic krill stomachs. Front. Zool. *5*, 12.

58. Galagoda, R., Chanto, M., Takemura, Y., Tomioka, N., Syutsubo, K., Honda, R., Yamamoto-Ikemoto, R., and Matsuura, N. (2023). Quantitative 16S rRNA gene amplicon sequencing for comprehensive pathogenic bacterial tracking in a municipal wastewater treatment plant. ACS ES&T Water *3*, 923–933.

59. Ferrer Florensa, A., Almagro Armenteros, J.J., Kaas, R.S., Conradsen Clausen, P.T.L., Nielsen, H., Rost, B., and Aarestrup, F.M. (2025). Whole-genome prediction of bacterial pathogenic capacity on novel bacteria using protein language models, with PathogenFinder2. Preprint at bioRxiv. https://doi.org/10.1101/2025.04.12.648497.

60. Cosentino, S., Voldby Larsen, M., Møller Aarestrup, F., and Lund, O. (2013). PathogenFinder–distinguishing friend from foe using bacterial whole genome sequence data. PLoS One *8*, e77302.

61. Deneke, C., Rentzsch, R., and Renard, B.Y. (2017). PaPrBaG: A machine learning approach for the detection of novel pathogens from NGS data. Sci. Rep. *7*, 39194.

62. Naor-Hoffmann, S., Svetlitsky, D., Sal-Man, N., Orenstein, Y., and Ziv-Ukelson, M. (2022). Predicting the pathogenicity of bacterial genomes using widely spread protein families. BMC Bioinf. *23*, 253.

63. Balaji, A., Kille, B., Kappell, A.D., Godbold, G.D., Diep, M., Elworth, R.A.L., Qian, Z., Albin, D., Nasko, D.J., Shah, N., et al. (2022). SeqScreen: accurate and sensitive functional screening of pathogenic sequences via ensemble learning. Genome Biol. *23*, 133.

64. Workman, J., Jr. (2024). A review of the latest spectroscopic research in pharmaceutical and biopharmaceutical applications. Spectroscopy (Springf.) *39*, 25–29. https://doi.org/10.56530/spectroscopy.at8171q52024.

65. Brauchle, E., and Schenke-Layland, K. (2013). Raman spectroscopy in biomedicine - non-invasive in vitro analysis of cells and extracellular matrix components in tissues. Biotechnol. J. *8*, 288–297.

66. Bataller, B.G., and Capareda, S.C. (2018). A rapid and non-destructive method for quantifying biomolecules in Spirulina platensis via Fourier transform infrared – Attenuated total reflectance spectroscopy. Algal Res. *32*, 341–352.

67. Saint-John, B., Wolf-Yadlin, A., Jacobsen, D.E., Inman, J.L., Gart, S., Keener, M., McMurray, C., Snijders, A.M., Mukundan, H., Kubicek-Sutherland, J.Z., and Brown, J.B. (2025). Reagent-free hyperspectral diagnosis of SARS-CoV-2 infection in saliva samples. ECS Sens. Plus *4*, 014601.

68. Locke, A., Fitzgerald, S., and Mahadevan-Jansen, A. (2020). Advances in optical detection of human-associated pathogenic bacteria. Molecules *25*, 5256.

69. Nogueira, M.S., Leal, L.B., Marcarini, W.D., Pimentel, R.L., Muller, M., Vassallo, P.F., Campos, L.C.G., Dos Santos, L., Luiz, W.B., Mill, J.G., et al. (2021). Rapid diagnosis of COVID-19 using FT-IR ATR spectroscopy and machine learning. Sci. Rep. *11*, 15409.

70. Kazmer, S.T., Hartel, G., Robinson, H., Richards, R.S., Yan, K., van Hal, S.J., Chan, R., Hind, A., Bradley, D., Zieschang, F., et al. (2022). Pathophysiological response to SARS-CoV-2 infection detected by infrared spectroscopy enables rapid and robust saliva screening for COVID-19. Biomedicines *10*, 351.

71. Martinez-Cuazitl, A., Vazquez-Zapien, G.J., Sanchez-Brito, M., Limon-Pacheco, J.H., Guerrero-Ruiz, M., Garibay-Gonzalez, F., Delgado-Macuil, R.J., de Jesus, M.G.G., Corona-Perezgrovas, M.A., Pereyra-Talamantes, A., and Mata-Miranda, M.M. (2021). ATR-FTIR spectrum analysis of saliva samples from COVID-19 positive patients. Sci. Rep. *11*, 19980.

72. Mukundan, H., Anderson, A.S., Grace, W.K., Grace, K.M., Hartman, N., Martinez, J.S., and Swanson, B.I. (2009). Waveguide-based biosensors for pathogen detection. Sensors (Basel) *9*, 5783–5809.

73. Jakhar, S., Sakamuri, R., Vu, D., Dighe, P., Stromberg, L.R., Lilley, L., Hengartner, N., Swanson, B.I., Moreau, E., Dorman, S.E., and Mukundan, H. (2021). Interaction of amphiphilic lipoarabinomannan with host carrier lipoproteins in tuberculosis patients: Implications for blood-based diagnostics. PLoS One *16*, e0243337.

74. Cohen, D.J.F., Li, N.C., Ioussoufovitch, S., and Diop, M. (2023). Fast estimation of adult cerebral blood content and oxygenation with hyperspectral time-resolved near-infrared spectroscopy. Front. Neurosci. *17*, 1020151.

75. Martinez, K.M., Wilding, K., Llewellyn, T.R., Jacobsen, D.E., Montoya, M.M., Kubicek-Sutherland, J.Z., Batni, S., Manore, C., and Mukundan, H. (2024). Factors influencing accuracy, interpretability and reproducibility in the use of machine learning in biology. Preprint at Res. Sq. https://doi.org/10.21203/rs.3.rs-4171489/v1.

76. Melendez-Pastor, I., Almendro-Candel, M., Navarro-Pedreño, J., Gómez, I., Lillo, M., and Hernández, E. (2013). Monitoring urban wastewaters' characteristics by visible and short wave near-infrared spectroscopy. Water (Basel) *5*, 2026–2036.

77. Corpuz, M.V.A., Buonerba, A., Zarra, T., Hasan, S.W., Korshin, G.V., Belgiorno, V., and Naddeo, V. (2022). Advances in virus detection methods for wastewater-based epidemiological applications. Case Stud. Chem. Environ. Eng. *6*, 100238.

78. Lebowitz, J., Lewis, M.S., and Schuck, P. (2002). Modern analytical ultracentrifugation in protein science: a tutorial review. Protein Sci. *11*, 2067–2079.

79. Simpson, D.M., and Beynon, R.J. (2010). Acetone precipitation of proteins and the modification of peptides. J. Proteome Res. *9*, 444–450.

80. Sánchez-Moya, T., Hidalgo, A.M., Ros-Berruezo, G., and López-Nicolás, R. (2020). Screening ultrafiltration membranes to separate lactose and protein from sheep whey: application of simplified model. J. Food Sci. Technol. *57*, 3193–3200.

81. Shahid, K., Srivastava, V., and Sillanpää, M. (2021). Protein recovery as a resource from waste specifically via membrane technology-from waste to wonder. Environ. Sci. Pollut. Res. Int. *28*, 10262–10282.

82. Guan, D., and Chen, Z. (2014). Challenges and recent advances in affinity purification of tag-free proteins. Biotechnol. Lett. *36*, 1391–1406.

83. Park, Y.-J., Jankowski, W., Hurst, N.C., Fry, J.W., Schwabe, N.F., Tan, L.C.C., and Sauna, Z.E. (2024). Functional activity and binding specificity of small ankyrin repeat proteins called ankyrons against SARS-CoV-2 variants. Preprint at bioRxiv. https://doi.org/10.1101/2024.10.11.617752.

84. Rinnan, Å., Berg, F.v.d., and Engelsen, S.B. (2009). Review of the most common pre-processing techniques for near-infrared spectra. Trends Analyt. Chem. *28*, 1201–1222.

85. Mehmood, T., Liland, K.H., Snipen, L., and Sæbø, S. (2012). A review of variable selection methods in Partial Least Squares Regression. Chemometr. Intell. Lab. Syst. *118*, 62–69.

86. de Santana, F.B., de Souza, A.M., and Poppi, R.J. (2018). Visible and near infrared spectroscopy coupled to random forest to quantify some soil quality parameters. Spectrochim. Acta Mol. Biomol. Spectrosc. *191*, 454–462.

87. Schuetzke, J., Szymanski, N.J., and Reischl, M. (2023). Validating neural networks for spectroscopic classification on a universal synthetic dataset. npj Comput. Mater. *9*, 100–112.

88. Sakudo, A., Suganuma, Y., Kobayashi, T., Onodera, T., and Ikuta, K. (2006). Near-infrared spectroscopy: promising diagnostic tool for viral infections. Biochem. Biophys. Res. Commun. *341*, 279–284.

89. Tucker, S., Dubb, J., Kura, S., von Lühmann, A., Franke, R., Horschig, J.M., Powell, S., Oostenveld, R., Lührs, M., Delaire, É., et al. (2023). Introduction to the shared near infrared spectroscopy format. Neurophotonics *10*, 013507.

90. Bishop, C.M. (2006). Pattern Recognition and Machine Learning (Information Science and Statistics) (Springer).

91. Martinez, K.M., Wilding, K., Llewellyn, T.R., Jacobsen, D.E., Montoya, M.M., Kubicek-Sutherland, J.Z., Batni, S., Manore, C., and Mukundan, H. (2025). Evaluating the factors influencing accuracy, interpretability, and reproducibility in the use of machine learning classifiers in biology to enable standardization. Sci. Rep. *15*, 16651.

92. Mishra, P., and Passos, D. (2021). Realizing transfer learning for updating deep learning models of spectral data to be used in new scenarios. Chemometr. Intell. Lab. Syst. *212*, 104283.

93. Cabrera, C., Pilobello, K., Dalvin, S., Bobrow, J., Shah, D., Garg, L.F., Chalise, S., Doyle, P., Miller, G.A., Walt, D.R., et al. (2022). Systematic approach to address early pandemic's diagnostic unmet needs. Front. Microbiol. *13*, 910156.

94. Manore, C., Graham, T., Carr, A., Feryn, A., Jakhar, S., Mukundan, H., and Highlander, H.C. (2019). Modeling and cost benefit analysis to guide deployment of POC diagnostics for non-typhoidal Salmonella infections with antimicrobial resistance. Sci. Rep. *9*, 11245.

95. Bartlow, A.W., Stromberg, Z.R., Gleasner, C.D., Hu, B., Davenport, K.W., Jakhar, S., Li, P.-E., Vosburg, M., Garimella, M., Chain, P.S.G., et al. (2022). Comparing variability in diagnosis of upper respiratory tract infections in patients using syndromic, next generation sequencing, and PCR-based methods. PLOS Glob. Public Health *2*, e0000811.

96. Gauthier, N.P.G., Chorlton, S.D., Krajden, M., and Manges, A.R. (2023). Agnostic sequencing for detection of viral pathogens. Clin. Microbiol. Rev. *36*, e0011922.