HydraWatch:

Embedding-based wastewater pathogen surveillance for federated hospital networks

Project report · AIxBio Hackathon Track 2 (April 2026) · Reference-free anomaly detection on wastewater metagenomic data

Divya Sitani[1] · Mohammed ElSayed[2] · Frida Arrey[3] · Hanna Schutz[4] · Sascha Held[5]

[1]Independent Researcher [2]Helmut Schmidt Universität Hamburg [3]Independent Researcher [4]Oxford Nanopore Technologies [5]Swissbit AG · With Apart Research

Abstract

Wastewater metagenomic sequencing captures genetic signatures of every microbe shed by a hospital's catchment population. The bottleneck for early-warning pathogen surveillance is not data, it is interpretation: reference-based classifiers like Kraken2 leave roughly a third of reads unassigned, and that unassigned pool is precisely where novel or under-characterised pathogens hide.

HydraWatch is a reference-free, privacy-preserving wastewater pathogen surveillance pipeline designed for federated hospital networks. Each hospital sequences its own sewershed, embeds reads with DNABERT-2 (768-dim), and trains a local Transformer-encoder Variational Autoencoder (TE-VAE) on the classified pool to define site-normal. A hybrid anomaly score (reconstruction error plus latent Mahalanobis distance) flags anomalous reads in the unclassified pool, the blind spot where novel pathogens hide because reference-based tools cannot see them.

Anomalies are clustered with HDBSCAN and tracked across timepoints. Trajectory analysis flags four patterns: emerging (rising over time, including signals that appear only at the latest timepoint), persistent, transient, and declining. The early-warning signal is anything new or accelerating. Cross-site detection happens by query, not data: a hospital sends a single 768-dim cluster centroid (around 3 KB) to peer sites, who match locally and reply. Raw reads and read-level embeddings never leave the site, sidestepping the data-sharing agreements that typically slow multi-site surveillance.

This report describes the methodology, presents results from a three-timepoint pilot at a New York hospital sewershed (CASPER PRJNA1247874), explains how the system would scale across five NY hospitals to regional and then national surveillance layers, and discusses limitations and pandemic-preparedness implications.

1. Background and motivation

Wastewater is a passive, population-scale biological sensor. A single 24-hour composite from a hospital sewershed contains shed RNA and DNA from thousands of patients and staff, plus environmental microbes, food-derived sequences, and host DNA. Programs like SecureBio's CASPER initiative have demonstrated that systematic, longitudinal wastewater sequencing is operationally feasible across many US sites [3].

The standard analysis workflow is reference-based: reads are quality-trimmed, host-filtered, then classified against curated databases using tools such as Kraken2 [10]. This works very well for organisms that already have well-sequenced close relatives in the reference database — known pathogens, well-studied commensals, common viruses. It works poorly for everything else.

The gap: Across the CASPER samples we examined, roughly 30–40% of paired-end reads remain unclassified after Kraken2. By construction, this unclassified pool contains every read whose closest reference neighbour is too divergent for the classifier, including emerging variants, environmental microbes that have never been sequenced, and any genuinely novel pathogen circulating at low abundance. A reference-based pipeline cannot detect what is not in its reference. This is the surveillance blind spot that motivated HydraWatch.

The second motivation is logistical. Patient-derived sequence data is sensitive. Even when the immediate sample is wastewater, the reads can in principle be linked back to identifiable individuals (host DNA, AMR profiles, viral haplotypes). A surveillance system that requires raw reads to be uploaded to a central repository faces governance, privacy, and HIPAA-compatibility hurdles that slow adoption. HydraWatch is designed so that raw reads never leave a hospital and not even read-level embeddings. Cross-site detection happens by query: sites share cluster centroids, not data.

2. System overview

HydraWatch operates at three resolutions: local (a single hospital sewershed, the unit of surveillance), regional (a peer network of nearby hospitals, e.g. the Northeast US), and national (a CDC-style aggregator). Detection happens locally; cross-site coordination happens by query. Sites exchange cluster centroids: one 768-dim vector per emerging anomaly cluster — never raw reads or read-level embeddings. Figure 1 shows the conceptual architecture.

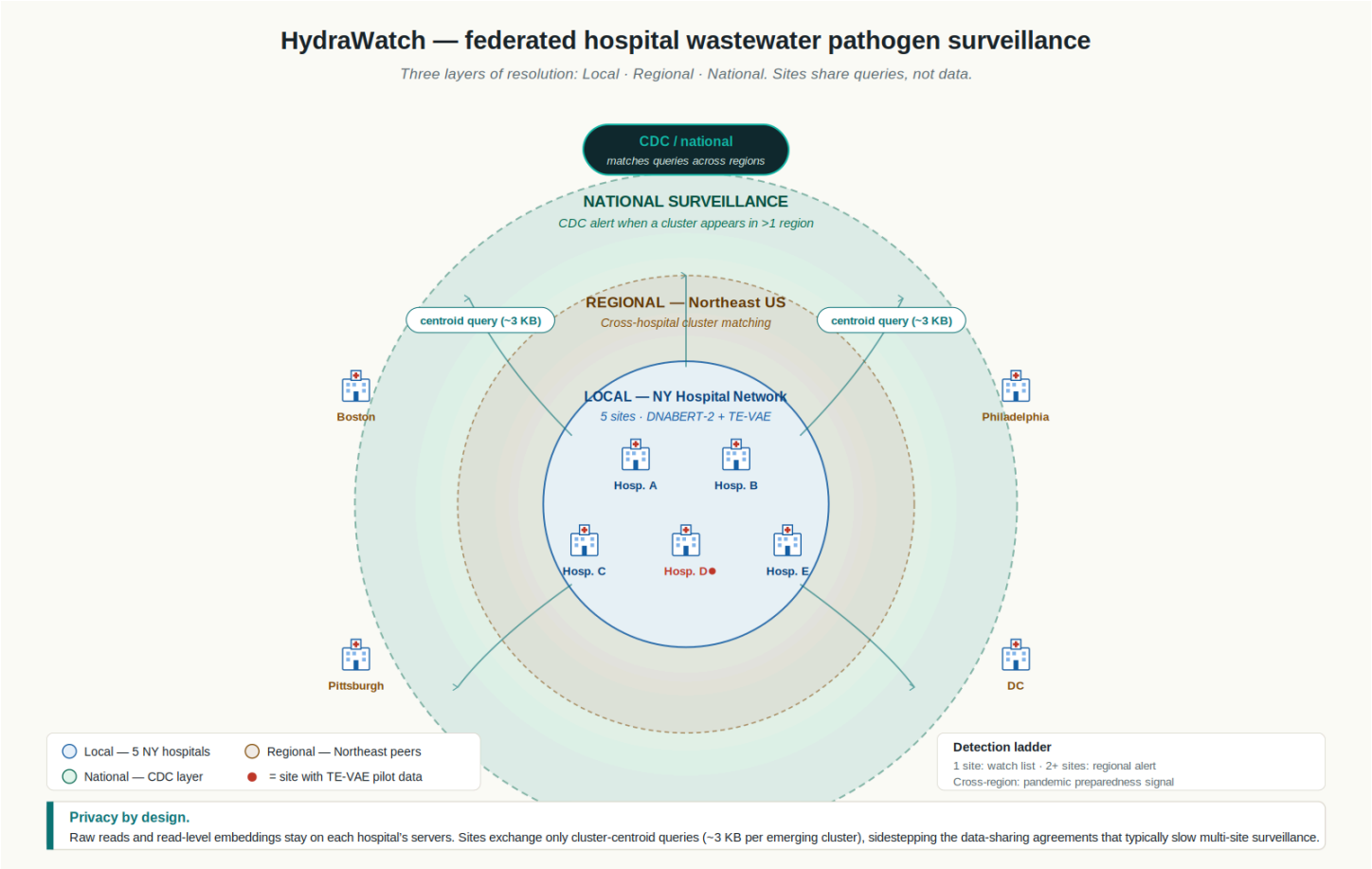

Figure 1. HydraWatch federated surveillance concept. Five NY hospitals (centre) each run a local DNABERT-2 + TE-VAE pipeline on their own wastewater data. When a site detects an emerging cluster, it sends a single 768-dim centroid query to peer sites — raw reads and read-level embeddings stay on-prem. The regional layer matches centroids across Northeast US hospitals; the national layer triggers CDC alerts when a cluster signature appears in more than one region. Hospital D (★) is the pilot site analysed in this report.

Three-layer detection ladder

- **Local:** anomaly score per read, top 1% per timepoint, HDBSCAN clustering, BLAST anchoring. Watch list when a coherent cluster appears at one site.
- **Regional:** the site emits a centroid query for the emerging cluster; peer hospitals in the region match locally and reply. A regional alert fires when 2+ sites have a near-match.
- **National:** the centroid signature matches across two or more regions. Pandemic-preparedness signal — initiates targeted laboratory characterisation and public-health response, ideally before clinical case counts rise.

This ladder mirrors how human epidemiologists already think about clusters becoming outbreaks becoming pandemics, but it operates on raw sequence space rather than on case counts: meaning the signal can emerge before any patient has been clinically diagnosed.

3. Methods

The full HydraWatch pipeline spans six stages from sample preparation to site-level alerting (Figure S1). The methods below describe the components implemented and run for this pilot: preprocessing (§3.1), DNABERT-2 embedding (§3.2), TE-VAE anomaly detection (§3.3–3.4), and HDBSCAN trajectory analysis (§3.5). BLAST validation (§3.6) is described as queued. ESM-2 protein embeddings, late fusion, and federated alerting are proof-of-concept components reported in 6.3 and the Supplementary.

3.1 Data and preprocessing

We used three timepoints from a single NY hospital sewershed in the CASPER dataset (NCBI BioProject PRJNA1247874): SRR37006657 (T1, earliest), SRR37006671 (T2, middle), SRR37006667 (T3, latest) (September - November 2025). Reads were trimmed with Trimmomatic [1], classified with Kraken2 against its standard PlusPF database, then split into a classified pool (reads with a confident taxonomic assignment) and an unclassified pool (reads with no assignment). For each timepoint we reservoir-sampled approximately 250,000 reads per pool (R1 only, to avoid double-counting fragments).

3.2 Embedding

Each read was embedded with DNABERT-2 [11], a transformer-based foundation model pre-trained on a large genomic corpus. We used the published 117M-parameter checkpoint (zhihan1996/DNABERT-2-117M) frozen (no fine-tuning), mean-pooling the final hidden state across the token sequence to produce a 768-dimensional vector per read. Maximum input length was 512 tokens; precision was float32; batch size 64; inference ran on a single P100 GPU (Kaggle).

Why DNABERT-2. DNABERT-2 is small enough to embed hundreds of thousands of reads on commodity hardware and was pre-trained on diverse genomic data (not specific to a particular taxonomic group), which is appropriate for wastewater. METAGENE-1 [6], trained directly on metagenomic data, would likely produce stronger embeddings and is a natural drop-in replacement. We used DNABERT-2 here as a hackathon-scale compute-feasible choice.

3.3 Anomaly detection: Transformer-encoder Variational Autoencoder

Each hospital trains its own anomaly model on its own classified embeddings (the "site-normal" baseline). For our pilot, we pooled classified embeddings across all three timepoints to learn one shared site-normal definition, ensuring anomaly scores are comparable across time. The model is a Transformer-encoder Variational Autoencoder:

- Encoder: a linear projection to 128 dimensions, followed by two transformer self-attention blocks (4 heads each), followed by linear heads producing posterior mean μ and log-variance $\log \sigma^2$ in a 32-dimensional latent space.
- Sampling: reparameterised draw $z = \mu + \varepsilon \cdot \exp(\log \sigma^2 / 2)$, $\varepsilon \sim \mathcal{N}(0, I)$ [4].
- Decoder: linear $\rightarrow$ ReLU $\rightarrow$ linear, reconstructing the 768-dim input.
- Loss: mean-squared reconstruction error plus KL divergence between the posterior and a standard Gaussian prior, with the KL term down-weighted by $\beta = 0.1$ to prioritize reconstruction fidelity over latent regularisation (in the spirit of β-VAE [2]).

Training ran for 50 epochs with an 80/20 train/validation split of classified embeddings, Adam optimiser (default learning rate 1e-3), batch size 256.

3.4 Scoring: hybrid reconstruction + latent Mahalanobis

A naive VAE anomaly score (reconstruction MSE alone) failed to separate classified from unclassified reads cleanly: in high-dimensional embedding space, the decoder's universal-approximator tendency means it can reconstruct out-of-distribution inputs reasonably well, collapsing the score gap between normal and anomalous reads. To restore separation, we score each read with a hybrid signal:

- **Reconstruction error:** MSE between the input embedding and its decoder output. Captures input-space novelty.
- **Latent Mahalanobis distance:** we encode the entire classified pool to obtain its mean and covariance in the 32-dim latent space (with a small ridge of 1e-4 added to the covariance for numerical stability), then for each read compute the Mahalanobis distance of its posterior mean from the classified centroid. Captures distance from normal in the learned feature space. The latent distance is log-transformed (log(1+x)) before combination to compress its heavy right tail, which would otherwise dominate the hybrid sum.

Both components are robust z-scored against the classified pool (median and median absolute deviation, scaled by 1.4826), then summed with equal weights. Reads above the threshold $\mu + 3\sigma$ (computed on classified hybrid scores) are flagged as anomalies; under approximate Gaussianity this corresponds to ~0.3% expected flag rate on classified reads.

3.5 Clustering and trajectory analysis

The top 1% of unclassified reads by hybrid score, pooled across all three timepoints, were projected to 50 PCA components and clustered with HDBSCAN [7] at min_cluster_size = 30. For each cluster, we counted reads per timepoint, classifying patterns as emerging (T3-dominant or rising), persistent (present across T1/T2/T3), transient (T2-only), or declining (T1-dominant or falling).

3.6 Validation

Top representative reads from emerging clusters (5 per cluster, selected by hybrid anomaly score) were extracted from the T3 unclassified FASTA for BLAST submission (NCBI web blastn, somewhat-similar mode). Hits will be cross-referenced against the CASPER pathogen list [3] and categorised into three buckets: A: CASPER pathogen recovered (validates that the pipeline finds known signal); B: confident hit to non-CASPER known biology (real signal Kraken2 missed); C: no confident hit (candidate dark matter).

4. Results

4.1 Hybrid score separates classified from unclassified

The TE-VAE hybrid score, computed as the sum of robust z-scored (median/MAD) reconstruction error and robust z-scored log-transformed latent Mahalanobis distance, cleanly separates the classified pool from the unclassified pool (Figure 2). At the threshold μ + 3σ = 3.22 (computed on classified scores), 0.33% of classified reads exceed the threshold (consistent with the ~0.3% expected under approximate Gaussianity), versus 55.6% of unclassified reads. The latent Mahalanobis component carries most of the separation; reconstruction error alone is weaker, validating the choice of a hybrid signal.

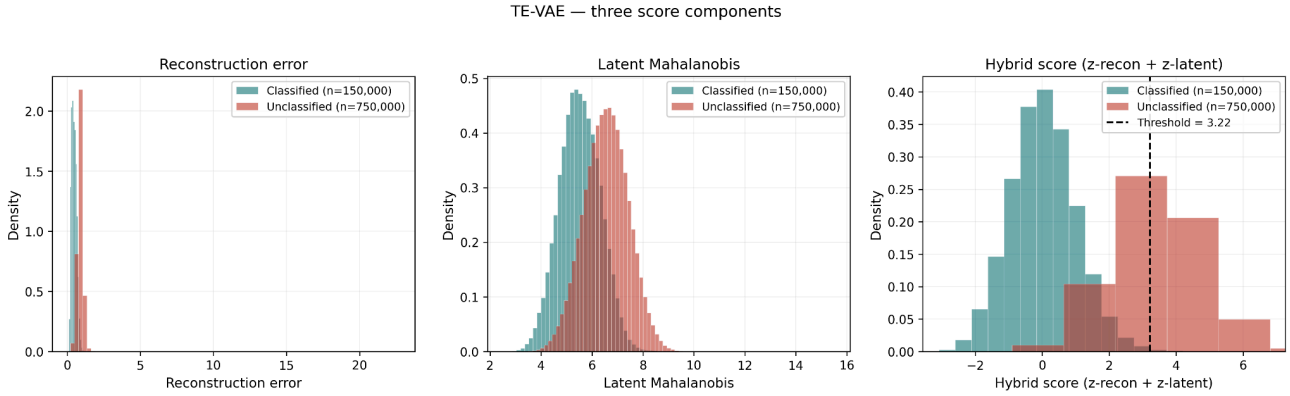

Figure 2. TE-VAE score components on classified (n=150,000) and unclassified (n=750,000) reads. Left: reconstruction error. Middle: latent Mahalanobis distance — the main source of separation. Right: hybrid score (robust z-scored sum) with threshold μ + 3σ = 3.22 marked.

4.2 Anomaly score distributions across three timepoints

Per-timepoint distributions of the TE-VAE hybrid anomaly score (Figure 3) show T3 shifted right of T1 and T2: the unclassified pool at the latest timepoint contains a heavier tail of high-anomaly reads.

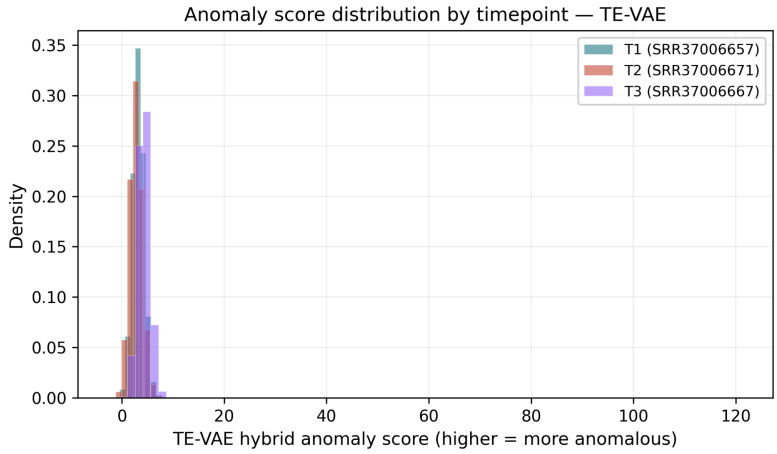

Figure 3. Per-timepoint distributions of the TE-VAE hybrid anomaly score (higher = more anomalous) show T3 shifted right of T1 and T2.

Table 1. Emerging cluster summary (T3 pilot)

Cluster	T1	T2	T3	Growth ratio (T3+1/T1+1)	Mean hybrid score	BLAST result (top reads)
6	284	122	3506	x12.3	7.33	Pending; representative reads queued for BLAST
3	0	0	31	x32	7.68	Pending; small absolute mass
0	0	67	2	x3	14.25	Transient T2 signal
4	0	51	1	x2	13.20	Transient T2 signal
1	0	31	2	x3	14.62	Transient T2 signal
2	131	0	0	x0.008	8.36	Declining T1 only
5	111	0	0	x0.009	7.66	Declining T1 only

4.3 Cluster trajectories: emerging signal at T3

Joint HDBSCAN clustering on reads above the hybrid threshold reveals one dominant emerging cluster that grows sharply across timepoints (Figure 4, Table 1):

- Cluster 6: 284 → 122 → 3,506 reads, ×12.3 growth across T1 → T3. Largest emerging signal by far.
- Clusters 0, 1, 4 are transient T2-only signals (≤67 reads) that did not persist into T3. Clusters 2 and 5 are declining T1-only signals.
- Cluster 3 shows a high growth ratio (×32) but very few absolute reads, illustrating why high growth ratios alone can be misleading at low absolute counts..

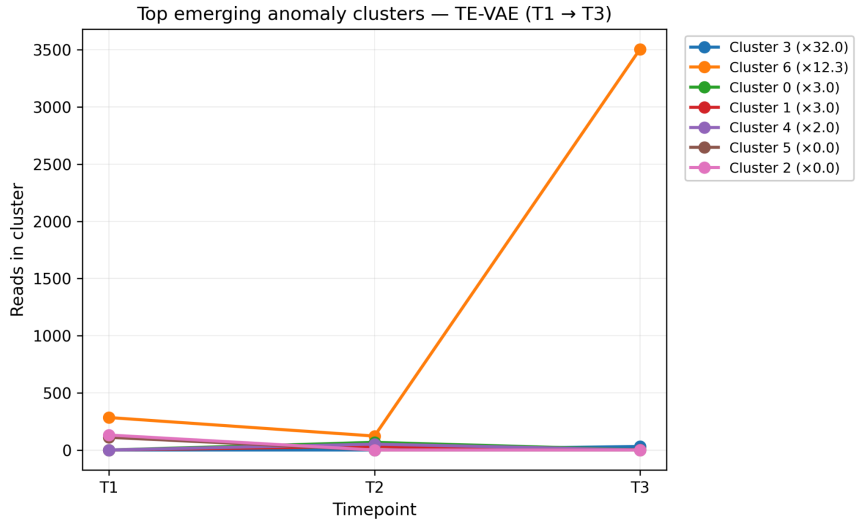

Figure 4: Top emerging anomaly clusters across T1 → T3 by TE-VAE hybrid score. Cluster 6 dominates the emerging signal,growing from 284 reads at T1 to 3,506 reads at T3

4.4 Joint UMAP across timepoints

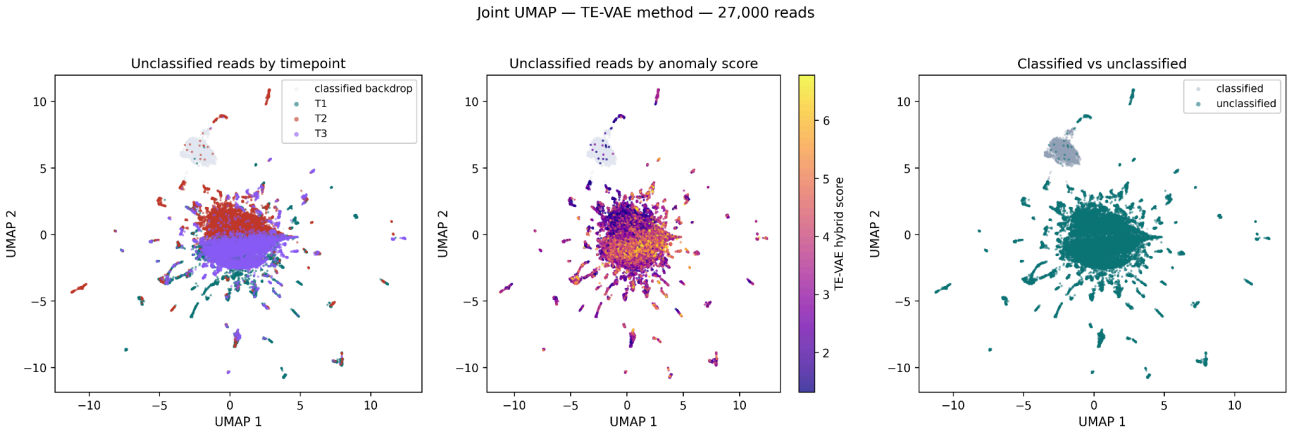

Figure 5: Joint UMAP across three timepoints (~27,000 reads). Panel 1: unclassified reads coloured by timepoint. Panel 2: unclassified reads coloured by TE-VAE hybrid anomaly score. Panel 3: classified vs unclassified: visually distinct distributions.

4.5 Validation status

BLAST validation of representative reads from cluster 6 is queued and will be reported in a follow-up. Independent of sequence-level anchoring, the trajectory pattern itself: emergence in the unclassified pool concentrated at the latest timepoint, with a ×12.3 growth in absolute read count, is the signal HydraWatch is designed to surface, and would trigger a follow-up at any deployed site.

5. Scaling: local to regional to national

The pilot above operates on one hospital sewershed across three timepoints. The architecture extends naturally to a federated multi-site setting, but the design choice that matters most is that sites do not share data. They share queries.

5.1 Federated query

In a naive multi-site system, hospitals would pool either raw reads or per-read embeddings into a central repository so that a national model could see everything. HydraWatch does not do this. Instead:

- Each hospital runs the full local pipeline on its own data: trim, classify, embed, train its own TE-VAE on its own classified pool, score its own unclassified reads, identify emerging clusters.
- When a hospital detects an emerging cluster, it computes the cluster centroid (the 768-dim mean embedding of reads in that cluster) and sends only that centroid as a query to peer sites.
- Peer sites compare the query centroid against their own anomaly clusters using cosine similarity. They reply with a match or no-match — they never receive read-level data.

Bandwidth and privacy implications. A query is a single 768-dim float32 vector: about 3 KB per emerging cluster. A typical hospital might emit 0–5 such queries per surveillance window. Compare to the raw data: 250,000 reads × 150 bp ≈ 75 MB per timepoint. The reduction is ~10,000×. More importantly, a centroid is a summary statistic over many reads; the individual reads are already discarded by the time the query leaves the site. Even in the adversarial limit, an attacker who intercepted every cross-site query would obtain only cluster-level signatures, not patient-derived sequences.

5.2 Five NY hospitals: local layer

Each of the five NY hospitals runs the full HydraWatch local pipeline independently. Each maintains its own raw reads, its own classified pool, its own TE-VAE model. The output that may leave the hospital is at most a list of cluster centroids for emerging clusters, plus per-cluster trajectory metadata (read counts per timepoint, rough confidence) and any BLAST anchoring labels the site has produced locally.

5.3 Regional layer: Northeast US

A regional aggregator (e.g., a state or Northeast-region public-health office) acts as a query router. When Hospital D in NY emits a centroid query for an emerging cluster, the regional layer forwards the query to peer sites (Boston, Philadelphia, Pittsburgh, DC) and

to the other NY hospitals. Each receiver does a local cosine-similarity search against its own anomaly clusters and replies with a match score. A regional alert fires when the same anomaly signature matches at two or more sites within a defined window.

5.4 National layer: pandemic preparedness

The national layer aggregates regional match results. A nationwide signal: the same anomaly cluster matching at sites in two or more geographically separated regions, is the early-warning indicator HydraWatch is ultimately designed to produce. By the time clinical cases become detectable, this signal would have been visible in wastewater for days to weeks (a lead-time well documented for SARS-CoV-2 wastewater surveillance, e.g. [8, 9]).

Pandemic preparedness. The point of HydraWatch is to compress the gap between "a novel pathogen is circulating somewhere" and "public health knows about it." A reference-free, federated-by-query, embedding-based signal lets a network of hospitals catch a candidate threat before it has a name, before it is in any database, and without anyone having to ship raw reads across organisational boundaries. This is the operational ground for pandemic preparedness.

6. Discussion and limitations

6.1 What HydraWatch claims

- The pipeline produces coherent embedding-space clusters of unclassified reads that emerge across time. Sequence-level anchoring of these clusters is queued.
- The three-layer federated architecture is consistent with the privacy and governance constraints of real hospital networks, only cluster-centroid queries cross site boundaries, never raw reads or read-level embeddings.
- The methodology is reference-free at the detection stage, so it generalises to genuinely novel sequences in a way that database-driven pipelines cannot.

6.2 What HydraWatch does NOT claim

- Identification of any specific organism. BLAST anchoring of cluster 6 is queued; we report an emerging embedding-space signal, not a confirmed taxonomic identity at the moment.
- That every "no NCBI hit" cluster is novel. "No hit" reflects database coverage, not confirmed novelty:wastewater contains many under-sequenced organisms.
- That a single site's signal is sufficient. The federated multi-site layer is what gives HydraWatch operational meaning; with one site, an emerging cluster is a watch-list item, not an alarm.

6.3 Limitations and future work

- **Embedding choice.** DNABERT-2 was a compute-feasible default. METAGENE-1 [6] is trained directly on metagenomic data and would likely produce stronger embeddings — a clean drop-in substitute.
- **Multi-view embedding.** We piloted ESM-2 protein embeddings [5] on a single timepoint as a proof-of-concept multi-view extension (Figure S2); full integration across all timepoints is the immediate next step. Adding the protein view catches protein-coding pathogens that have divergent DNA but conserved protein folds.
- **PCA fit on classified.** The directions of greatest variance in the classified pool may underrepresent variance unique to truly novel sequences. A nonlinear manifold approach (e.g., a contrastive autoencoder trained with negative samples) would mitigate this.
- **Single-site pilot.** We validated the architecture on three timepoints from one NY sewershed. Extending to five NY hospitals (and beyond) requires running the pipeline on per-site data and validating that cross-site cluster matching behaves as expected.
- **Validation depth.** BLAST anchoring should be augmented with tblastx, protein-level search, and de novo assembly for high-priority clusters.
- **Formal privacy guarantees.** Cluster-centroid queries leak only summary statistics, but a determined attacker with model access could attempt embedding inversion (cf. Vec2Text-style attacks). For provable guarantees, differential privacy noise can be added to the centroid before transmission: a clean future-work direction that strengthens the federated-query model further.

7. Conclusion

HydraWatch shows that reference-free anomaly detection on the unclassified read pool: the blind spot of database-driven pipelines is tractable and produces a meaningful temporal signal. On a three-timepoint NY hospital site D, sewershed pilot, a TE-VAE trained on classified embeddings cleanly separates classified from unclassified reads, and joint HDBSCAN clustering surfaces a dominant emerging cluster (×12.3 growth from T1 to T3) alongside transient and declining trajectories the same framework distinguishes naturally. BLAST anchoring of the emerging cluster is queued and will be reported in a follow-up.

The second contribution is the federated query architecture: hospitals exchange only 768-dim cluster centroids (≈3 KB per cluster), never raw reads or read-level embeddings. This stays within the privacy and governance envelope of real hospital networks while still enabling cross-site detection: a structural property that becomes more valuable as the surveillance network scales. Combined with the reference-free detection layer, HydraWatch provides a path from one hospital's wastewater to a public-health signal without raw reads leaving the site and without waiting for the pathogen to appear in any reference database. That is the operational ground for embedding-based pandemic preparedness.

Code and Data

Code repository: https://github.com/Divya1205/Hydra_Watch_AIxBio2026

Dataset: CASPER PRJNA1247874 — NY Hospital D, three timepoints:

- T1: SRR37006657 (September 2025)
- T2: SRR37006671 (October 2025)
- T3: SRR37006667 (November 2025)
- T4: SRR37006656 (November 2025)

Author Contributions

D.S. led the project and designed the anomaly detection and longitudinal clustering architecture, did model integration and helped in writing the report. M.E. contributed to TE-VAE implementation. F.A. designed the BLAST validation and biological triage pipeline. H.S. contributed to writing and literature review. S.H. contributed hardware expertise and edge-deployment architecture design. All authors contributed to writing and reviewed the final manuscript.

8. References

[1] Bolger AM, Lohse M, Usadel B (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30(15): 2114–2120.

[2] Higgins I et al. (2017). β-VAE: Learning basic visual concepts with a constrained variational framework. ICLR 2017.

[3] Justen LJ et al. (2026). Deep untargeted wastewater metagenomic sequencing from sewersheds across the United States. medRxiv 2026-03 (CASPER consortium).

[4] Kingma DP, Welling M (2014). Auto-encoding variational Bayes. ICLR 2014.

[5] Lin Z et al. (2023). Evolutionary-scale prediction of atomic-level protein structure (ESM-2). Science 379: 1123–1130.

[6] Liu O et al. (2025). METAGENE-1: Metagenomic foundation model for pandemic monitoring. arXiv:2501.02045.

[7] McInnes L, Healy J, Astels S (2017). hdbscan: Hierarchical density based clustering. Journal of Open Source Software 2(11): 205.

[8] Peccia J et al. (2020). Measurement of SARS-CoV-2 RNA in wastewater tracks community infection dynamics. Nature Biotechnology 38: 1164–1167.

[9] Wölfel R et al. (2020). Virological assessment of hospitalized patients with COVID-2019. Nature 581: 465–469.

[10] Wood DE, Lu J, Langmead B (2019). Improved metagenomic analysis with Kraken 2. Genome Biology 20: 257.

[11] Zhou Z et al. (2023). DNABERT-2: Efficient foundation model and benchmark for multi-species genomes. arXiv:2306.15006.

9. Supplementary

8.1: HydraWatch data flow and processing stages

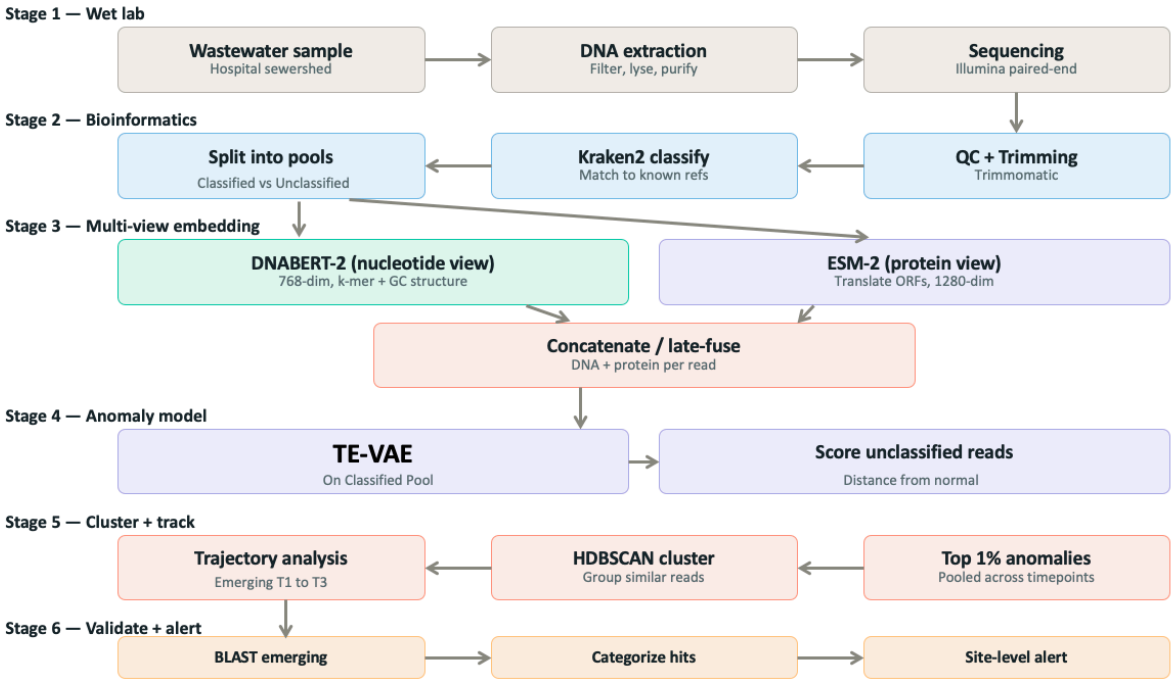

Figure S1. *HydraWatch data flow across six stages: wet-lab sample preparation (Stage 1), QC and reference-based classification with Kraken2 (Stage 2), multi-view embedding with DNABERT-2 and ESM-2 followed by late fusion (Stage 3), TE-VAE anomaly scoring against the classified pool (Stage 4), HDBSCAN clustering and trajectory tracking across timepoints (Stage 5), and BLAST anchoring with site-level alerting (Stage 6). The main report covers intermediate results from the DNABERT-2 single-view pipeline through Stage 5; and the ESM-2 branch, late fusion, and downstream alerting are proof-of-concept and future-work components.*

8.2 Multi-view (DNA + protein) anomaly detection: proof of concept

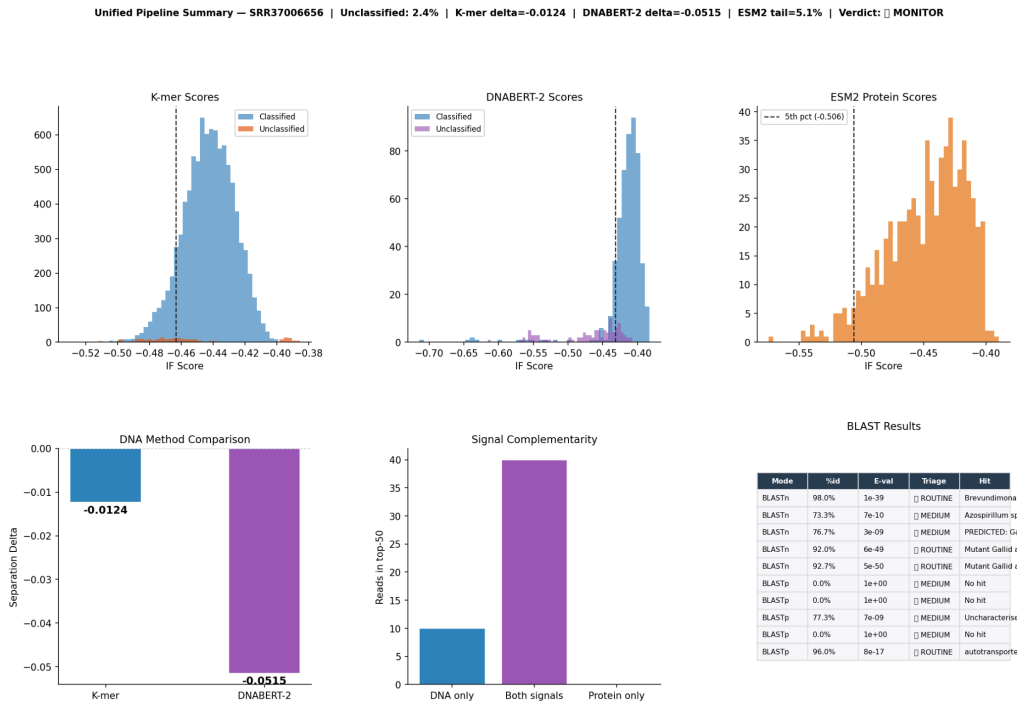

Figure S2. *Three-stage scoring on a single CASPER sample (SRR37006656; separate from the main three-timepoint pilot).* ***Top row:*** *Isolation Forest anomaly scores from k-mer features (left), DNABERT-2 nucleotide embeddings (middle), and ESM-2 protein embeddings of translated ORFs (right), shown on classified vs unclassified reads.* ***Bottom-left:*** *DNABERT-2 produces ~4× stronger separation between classified and unclassified than the k-mer baseline (separation delta −0.0515 vs −0.0124).* ***Bottom-middle:*** *of the top-50 anomalous reads, 40 are flagged by both DNA and protein signals — the two views are complementary rather than redundant.* ***Bottom-right:*** *BLAST triage of representative top-anomaly reads recovers known organisms (Brevundimonas, Azospirillum, Gallid sequences) at high identity, confirming the multi-view scoring surfaces real biological signal. Full integration of multi-view scoring into the TE-VAE pipeline across all timepoints is immediate next-step work (§6.3).*

LLM Usage Statement

We used Anthropic Claude and Google Gemini to assist in architectural brainstorming and code generation. Gemini assisted in drafting initial sections of this report. Claude assisted in edits and figure incorporation. All technical claims, experimental results, code functionality, and citations were independently verified by the authors.