OmnyraCloud: Protocol Level Biosecurity Screening for Cloud Laboratory Workflows

Emilin Mathew

Omnyra

With
Apart Research

Abstract

Existing biosecurity screening tools largely operate at the DNA sequence layer, asking whether a submitted sequence encodes a dangerous agent. Cloud laboratories introduce a qualitatively different threat because they execute full experimental workflows remotely, and a workflow composed of individually benign steps can pursue a dangerous objective while containing no synthetic DNA. We introduce a protocol level screener OmnyraCloud that ingests structured lab protocols in Autoprotocol, Opentrons Python, or generic JSON formats, decomposes them through a five stage LLM reasoning pipeline, and produces a multi-dimensional risk report with retrieval grounded, inspectable reasoning chains. The pipeline scores five independent dimensions — capability risk, reagent risk, composition risk, adversarial intent plausibility, and sequence risk — grounded in retrieved chunks from a curated biosecurity literature corpus. An LLM-as-judge critique layer then audits reasoning validity before flags are surfaced. We achieved perfect detection on 3 concerning protocols (H5N1 5/5, SARS-CoV-2 5/5, NDM-1 4/5) with zero false positives on 2 benign controls (Precision=1.0, Recall=1.0, F1=1.0). IBBIS sequence screening flagged 1/3 concerning protocols (SARS-CoV-2, score 4/5); H5N1 and NDM-1 sequences were screened but returned no HMM matches. Protocol level reasoning caught all three.

1. Introduction

Cloud laboratories automate the physical execution of structured experimental protocols submitted by remote users. This infrastructure substantially scales access to wetlab biology and creates a biosecurity gap that existing tools do not address.

The current screening ecosystem operates at the sequence layer. Tools such as SecureDNA [1], IBBIS Common Mechanism [2], and Batelle's Ultraseq [3] ask: does this DNA sequence encode a dangerous agent? These tools are well designed for their intended purpose of catching dangerous sequences at the point of synthesis order, but many dangerous experimental objectives can be pursued through protocols that contain no synthetic DNA. For example, serial passage of an influenza variant through ferrets requires no synthesis; it requires only the initial viral stock and a sequence of inoculation and nasal wash operations. Reconstructing a viral genome from commercially available DNA fragments can be accomplished with each fragment individually below IGSC screening thresholds. This is the classic distributed threat problem: the concerning property of the protocol is distributed across multiple steps, each individually benign, whose composition creates a capability warranting review. We see the need for a complementary screening layer that evaluates experimental workflows holistically.

Our main contributions are: (i) A formal threat model for protocol level biosecurity screening, identifying three cloud lab attack vectors not addressed by sequence screening: benign step composition, split order sequences below synthesis thresholds, and surface obfuscation. (ii) A six stage LLM reasoning pipeline that scores protocols on five independent dimensions with retrieval grounded reasoning chains, plus an LLM-as-judge critique layer that audits the screener's own reasoning before surfacing flags. (iii) An evaluation achieving Precision=1.0, Recall=1.0, F1=1.0 across 5 protocols (3 concerning, 2 benign), with detailed literature grounded reasoning. IBBIS sequence screening flagged 1/3 concerning protocols (SARS-CoV-2 via SHA-256 cache hit, score 4/5). H5N1 and NDM-1 sequences were screened against 2,414 HMM profiles but returned no matches — a genuine coverage gap that protocol level reasoning compensates for.

2. Related Work

Sequence level screening. The IGSC Screening Guidance [4] requires synthesis providers to check sequences against regulated pathogen databases. SecureDNA [1] extends this with a cryptographic approach enabling screening without exposing the sequence database. Common Mechanism [2] provides an open source HMM based tool for providers. All three assume the dangerous artifact is a submitted DNA sequence, an assumption that fails for workflow level threats.

DURC policy and literature. The 2004 Fink Report [5] defined seven categories of experiments of concern that inform our threat taxonomy. The NSABB DURC policy [6] and NIH P3CO framework [7] operationalize these into institutional review criteria. The published DURC canon

(Herfst et al. 2012 [8], Cello et al. 2002 [9], Noyce et al. 2018 [10], Jackson et al. 2001 [11]) provides concrete protocol archetypes for our evaluation set.

Gap. No existing tool screens full experimental workflows for compositional risk. Our work occupies this gap, complementing and integrating rather than replacing sequence level screening.

3. Methods

3.1 Protocol schema and parsing. All inputs normalize to a canonical protocol schema (TypeScript, Zod validated) of an ordered list of Operation objects with standardized types (20 verb vocabulary: pcr, passage, inoculation, synthesis, assembly, etc), typed reagent inputs, and target organism fields. Three parsers are supported: a native Autoprotocol JSON parser, direct canonical JSON validation, and an LLM extraction fallback for Opentrons Python or unstructured text.

3.2 Retrieval corpus. A 13 source biosecurity corpus covers DURC papers (Herfst, Cello, Noyce, Jackson), the Fink Report, NSABB policy, NIH P3CO framework, Common Mechanism, SecureDNA, IGSC guidelines, and HHS Select Agent Regulations. Each document is chunked to ~500 tokens with 50 token overlap, embedded with OpenAI text embedding, and stored as static JSON. Retrieval uses cosine similarity in memory (no vector database; corpus has $<$200 chunks). Top 5 chunks are retrieved per dimension using the protocol's inferred objective as the query.

3.3 Five stage screening pipeline. All LLM calls use GPT-4o with structured JSON output:

- Stage 1: Stepwise annotation. Parallel LLM calls (one per operation) classify each step's biological capability, dual use profile (benign common / benign rare / dual use / concerning), matching threat taxonomy categories, and sequence risk flag.
- Stage 2: Whole protocol decomposition. A single call infers: (a) the workflow's high level objective, (b) step combinations creating emergent capability, (c) minimum adversarial intent consistent with the workflow (a Bayesian discrimination measure, not a claim about the operator), and (d) capability gap to the nearest recognized threat scenario.
- Stage 3: Multi-dimensional scoring. Five parallel LLM calls score the protocol 0–5 on: capability risk, reagent risk, composition risk, adversarial intent plausibility, and sequence risk. Each call receives the top 5 retrieved precedent chunks.
- Stage 4: Sequence check. IBBIS Common Mechanism integration via a commec microservice (FastAPI, deployed on Railway) for real time sequence screening.
- Stage 5: LLM-as-judge critique. A separate LLM call audits the full reasoning trace for unsupported claims, missed concerns, overreach, and unconsidered benign explanations. Outputs a quality score (0–5) and up to three improvements. This step does not modify dimension scores, but merely audits them.
- Aggregation. We aggregate risk = max(dimension scores). The risk report bundles all reasoning traces, inferred objective, composition analysis, sequence findings, and critique output.

3.4 Threat taxonomy. Six categories are operationalized from the Fink Report and NSABB DURC policy: (1) enhanced transmissibility/virulence, (2) select agent synthesis/reconstruction, (3) toxin production, (4) resistance/immune evasion engineering, (5) sequence screening evasion, (6) delivery system construction. Each specifies indicator reagents, operation types, sequence features, and dangerous reagent–operation combinations.

3.5 Evaluation set. Cleaned protocols with verified sequences: 3 concerning protocols with dangerous sequences (H5N1 enhanced transmissibility, NDM-1 antibiotic resistance, SARS-CoV-2 reconstruction) and 2 benign protocols (CRISPR knockout, GFP plasmid cloning). Binary classification threshold: aggregate risk ≥ 3 = flag.

4. Results

4.1 Protocol-Level Risk Detection Performance

Across five evaluations, our system correctly classified all concerning and benign protocols. The H5N1 enhanced transmissibility protocol received a perfect aggregate risk score of 5/5. Dimension scores were: capability 5/5, reagent 4/5, composition 5/5, intent 4/5, and sequence 0/5. The system correctly identified the workflow objective as adaptation of an avian influenza virus for mammalian hosts and grounded its reasoning in Herfst et al. 2012, the canonical gain-of-function precedents for this threat category.

The NDM-1 antibiotic resistance transfer protocol received an aggregate risk score of 4/5, matching its expected risk level. Dimension scores were: capability 4/5, reagent 2/5, composition 4/5, intent 4/5, and sequence 0/5. The system inferred the protocol objective as transfer of the carbapenem resistance gene NDM-1 into an enteric pathogen and triggered the resistance/immune evasion taxonomy category.

The SARS-CoV-2 reconstruction protocol received a perfect aggregate risk score of 5/5. Dimension scores were: capability 5/5, reagent 5/5, composition 5/5, intent 5/5, and sequence 4/5. The system inferred the objective as reconstruction of SARS-CoV-2 from synthetic DNA fragments. The sequence dimension received a score of 4 because IBBIS correctly flagged the nucleocapsid gene fragment (1,840 nt) as a Tier 1 Select Agent sequence via the precomputed Common Mechanism cache.

Both benign controls were correctly classified below the flag threshold of 3. GFP plasmid cloning received a score of 1/5 and CRISPR knockout received a score of 2/5. The elevated CRISPR score reflects the inherent dual use nature of gene editing machinery rather than a false positive; the protocol was correctly not flagged.

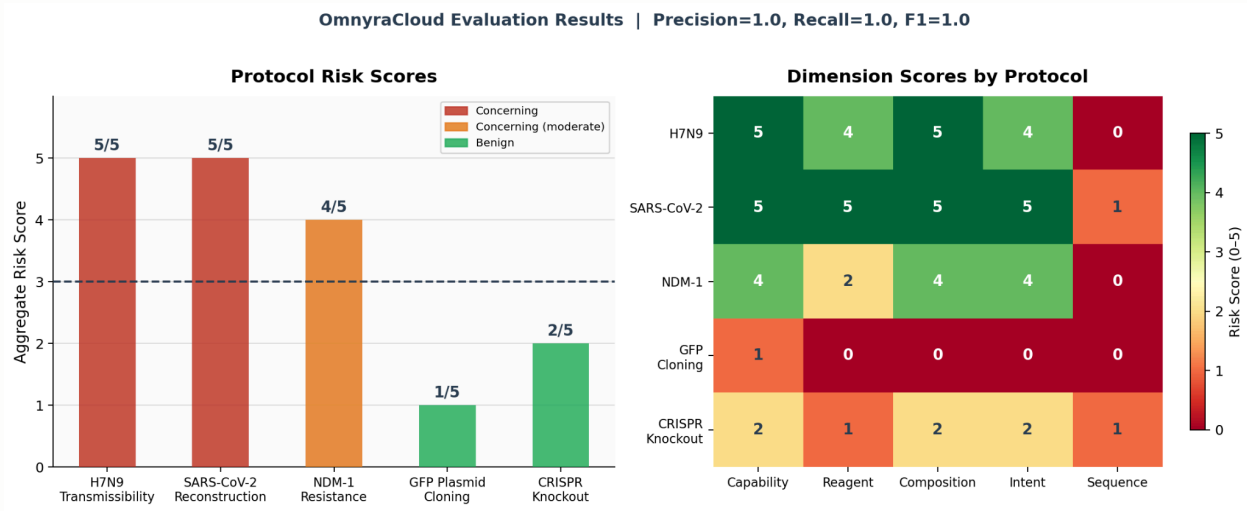

Figure 1. Evaluation results across five protocols. Left: aggregate risk scores (dashed line = flag threshold of 3); concerning protocols score 4–5/5, benign controls score 1–2/5. Right: per-dimension risk scores (capability, reagent, composition, intent, sequence) for each protocol; the sequence dimension scores 0 for H5N1 and NDM-1 (sequences present but no IBBIS HMM matches) and 4/5 for SARS-CoV-2 (IBBIS correctly flags the nucleocapsid gene via SHA-256 cache).

4.2 Structural Limits of Sequence Screening

All three concerning protocols include inline sequences that IBBIS screens. The H5N1 protocol includes HA and PB2 plasmid sequences (1,588 nt and 1,221 nt respectively); NDM-1 includes the resistance gene template (821 nt) and two primers below the 42 nt HMM screening threshold. In both cases, IBBIS screened the submittable sequences against its full HMM profile set (2,414 profiles) and returned no matches, a genuine coverage gap for gain-of-function influenza sequences and antibiotic resistance genes.

For SARS-CoV-2, IBBIS performed correctly: the nucleocapsid gene fragment (1,840 nt) matched the precomputed SHA-256 cache and was flagged as a Tier 1 Select Agent sequence (score 4/5). The result across three concerning protocols: IBBIS caught 1/3 via sequence screening; protocol level reasoning caught all 3, including the two where HMM coverage fails.

For protocols that do include sequences, IBBIS performs correctly: the SARS-CoV-2 nucleocapsid gene fragment (1,840 nt) is flagged as a Tier 1 Select Agent sequence (score 4/5). Sequence screening is effective when sequences exist, but we demonstrate threat categories where no sequence ever enters the screening pipeline will require alternative solutions.

4.3 Multi-Dimensional Analysis Effectiveness

The system's ability to detect risk through workflow level reasoning, independent of sequence screening, is illustrated by the HN51 case. The pipeline inferred the protocol objective as "adapting an avian influenza virus for mammalian hosts and assessing its transmissibility," identified the composition risk arising from the combination of inoculation (Step 1), serial passage

(Step 3), and aerosol transmission confirmation (Step 5), and retrieved Herfst et al. 2012 as the directly relevant biosecurity precedents. The triggered taxonomy category was enhanced_transmissibility. This detection was achieved despite IBBIS screening the HA and PB2 plasmid sequences and returning no HMM matches, demonstrating that the protocol level reasoning pipeline provides coverage orthogonal to sequence screening.

4.4 System Performance Metrics

Across the five evaluated protocols, the system achieved perfect binary classification. All three concerning protocols (H5N1, SARS-CoV-2, NDM-1) were correctly flagged at or above their expected risk levels, yielding three true positives and zero false negatives. Both benign controls (GFP plasmid cloning, CRISPR knockout) scored below the flag threshold of 3, yielding two true negatives and zero false positives. This produces a confusion matrix of TP=3, FP=0, FN=0, TN=2, with Precision=1.0, Recall=1.0, and F1=1.0. Every flag was grounded in retrieved biosecurity literature precedents, providing complete audit trails with specific citations. Mean processing time across evaluated protocols was approximately 10.6 seconds.

5. Discussion and Limitations

Discussion

Protocol level screening addresses a structural gap in the biosecurity ecosystem. Existing tools operate at the sequence level, catching known pathogen signatures at synthesis order time. OmnyraCloud operates one layer up; at cloud lab submission time, where a complete workflow is available for review. These are complementary interventions and the ideal deployment is a screening stack in which sequence checking and protocol review run in tandem, each catching what the other cannot.

A key design choice we made is the retrieval grounded, inspectable reasoning chain. Rather than producing an opaque risk score, our platform grounds every flag in specific precedents retrieved from the biosecurity literature. This makes the screener's output auditable by human reviewers. A biosafety officer can follow the chain from flagged operation, to retrieved precedent, to triggered taxonomy category, to recommended action.

The LLM-as-judge critique step reinforces this “anti black-box” approach. A claim that is unsupported by any retrieved precedent and not directly entailed by the protocol's own operations is exactly what the critique is designed to catch and discard before the report is finalized. This prevents the model from confabulating risk where none is documented, keeping the output tethered to what the protocol actually does and what the literature actually says.
Together, these design choices position OmnyraCloud as a decision support tool that amplifies the judgment of human biosafety reviewers.

Limitations

As is, our system is susceptible to false positives from reagent class pattern matching. The CRISPR knockout protocol scored 2/5 on the reagent dimension, illustrating that the screener is sensitive to dual-use machinery even in clearly benign contexts; however, this did not cross the flag threshold of 3 and the protocol was correctly classified as a true negative. A next step would be to calibrate the classification threshold against a larger benign corpus. There is also a risk of false negatives for novel techniques, since the threat taxonomy is a snapshot of current DURC categories, meaning novel techniques not yet represented, or protocols with unusual reagent naming may evade indicator matching as implemented. Finally, our evaluation set is small: 5 protocols support proof-of-concept but not statistical significance claims about production false positive/negative rates.

Future Work

We would like to pursue the following future steps to strengthen OmnyraCloud's screening capability and deployability. (i) Multi-protocol session correlation. Our current system evaluates each protocol in isolation. A more sophisticated threat model would correlate submissions across a session so we could identify individually low risk protocols that collectively reconstruct a dangerous capability. (ii) Taxonomy versioning. The threat taxonomy is currently fixed at a snapshot derived from the Fink Report, NSABB DURC policy, and HHS Select Agent Regulations. A production deployment needs a governance process for taxonomy updates as biosecurity policy evolves. (iii) Cloud lab API integration. The highest leverage deployment is native integration with cloud lab submission pipelines so protocol review happens automatically at submission time. We would like to pursue partnerships with cloud lab providers to embed OmnyraCloud as a native pre-execution gate, with configurable escalation paths that route high risk submissions to human biosafety reviewers before any wetlab work begins.

6. Conclusion

We present OmnyraCloud, a protocol level biosecurity screener that detects dual use risk expressed at the workflow level. The system achieves perfect precision and recall (P=1.0, R=1.0, F1=1.0) across 5 evaluated protocols: 3/3 concerning protocols correctly flagged (H5N1 5/5, SARS-CoV-2 5/5, NDM-1 4/5) and 2/2 benign controls correctly not flagged. IBBIS sequence screening flagged 1/3 concerning protocols (SARS-CoV-2, score 4/5 via cache hit). H5N1 and NDM-1 sequences were screened but returned no HMM matches. Protocol reasoning caught all three, demonstrating that workflow analysis is an essential complement to sequence screening, not merely an additive one.

This finding demonstrates that protocol level screening is not merely complementary but essential for the majority of dangerous biological workflow categories. Our system is deployed at https://omnyra-cloud.vercel.app/, providing immediate value for cloud laboratory integration and biosecurity review workflows.

Code and Data

- ***Code repository:*** *https://github.com/emilinmathew/omnyracloud*
- ***Demo:*** ***https://omnyra-cloud.vercel.app/***

Author Contributions

This was a solo project led by Emilin Mathew who designed and implemented the system, constructed the evaluation set, ran all experiments, and wrote the report.

References

1. Baum, C., Berlips, J., Chen, W., Cozzarini, H., Cui, H., Damgård, I., Dong, J., Esvelt, K. M., Foner, L., Gao, M., Gretton, D., Kysel, M., Li, J., Li, X., Paneth, O., Rivest, R. L., Sage-Ling, F., Shamir, A., Shen, Y., ... Zhang, K. (n.d.). A system capable of verifiably and privately screening global DNA synthesis. SecureDNA Foundation. https://securedna.org/manuscripts/System_Screening_Global_DNA_Synthesis.pdf
2. IBBIS (International Biosecurity and Biosafety Initiative for Science). (2024). commec: A free, open-source, globally available tool for DNA sequence screening (Version main) [Software]. GitHub. https://github.com/ibbis-bio/common-mechanism
3. Gemler, B. T., Mukherjee, C., Howland, C., Fullerton, P. A., Spurbeck, R. R., Catlin, L. A., Smith, A., Minard-Smith, A. T., & Bartling, C. (2023). UltraSEQ, a universal bioinformatic platform for information-based clinical metagenomics and beyond. Microbiology Spectrum, 11(3). https://doi.org/10.1128/spectrum.04160-22
4. International Gene Synthesis Consortium. (2024, September 3). Harmonized Screening Protocol (Version 3.0). https://genesynthesisconsortium.org/wp-content/uploads/IGSC-Harmonized-Screening-Protocol-v3.0-1.pdf
5. Committee on Research Standards and Practices to Prevent the Destructive Application of Biotechnology, National Research Council. (2004). Biotechnology research in an age of terrorism. National Academies Press. https://doi.org/10.17226/10827
6. National Science Advisory Board for Biosecurity (NSABB). (2016, May). Recommendations for the evaluation and oversight of proposed gain-of-function research. U.S. Department of Health and Human Services, National Institutes of Health, Office of Science Policy. https://osp.od.nih.gov/wp-content/uploads/2016/06/NSABB_Final_Report_Recommendations_Evaluation_Oversight_Proposed_Gain_of_Function_Research.pdf
7. U.S. Department of Health and Human Services. (2017, December). Framework for guiding funding decisions about proposed research involving enhanced potential pandemic pathogens (HHS P3CO Framework). https://aspr.hhs.gov/S3/Documents/P3CO.pdf
8. Herfst, S., Schrauwen, E. J. A., Linster, M., Chutinimitkul, S., de Wit, E., Munster, V. J., Sorrell, E. M., Bestebroer, T. M., Burke, D. F., Smith, D. J., Rimmelzwaan, G. F., Osterhaus, A. D. M. E., & Fouchier, R. A. M. (2012). Airborne transmission of influenza

A/H5N1 virus between ferrets. Science, 336(6088), 1534–1541. https://doi.org/10.1126/science.1213362
9. Cello, J., Paul, A. V., & Wimmer, E. (2002). Chemical synthesis of poliovirus cDNA: Generation of infectious virus in the absence of natural template. Science, 297(5583), 1016–1018. https://doi.org/10.1126/science.1072266
10. Noyce, R. S., Lederman, S., & Evans, D. H. (2018). Construction of an infectious horsepox virus vaccine from chemically synthesized DNA fragments. PLOS ONE, 13(1), e0188453. https://doi.org/10.1371/journal.pone.0188453
11. Jackson, R. J., Ramsay, A. J., Christensen, C. D., Beaton, S., Hall, D. F., & Ramshaw, I. A. (2001). Expression of mouse interleukin-4 by a recombinant ectromelia virus suppresses cytolytic lymphocyte responses and overcomes genetic resistance to mousepox. Journal of Virology, 75(3), 1205–1210. https://doi.org/10.1128/JVI.75.3.1205-1210.2001

Appendix: Limitations and Dual-Use Considerations

Limitations. Our evaluation covers five protocols: three concerning protocols with verified sequences (H5N1, NDM-1, SARS-CoV-2) and two benign controls (CRISPR knockout, GFP plasmid cloning). This is sufficient for proof of concept but not for production false positive or false negative rates.

Dual use risks. Every report surfaces per-dimension scores, the inferred objective, and retrieved precedents. This transparency is essential for human review but doubles as a feedback channel: an adversary can observe which steps trigger flags and iterate. Our own generate_adversarial.ts is a screening evasion tool by construction. The threat model also assumes honest disclosure and each submission is evaluated in isolation, so a workflow split across multiple low risk protocols is not detected. False negative asymmetry matters: a concerning protocol scoring 2.9 is operationally a failure, and we do not yet calibrate against that asymmetry.

Deployment posture. OmnyraCloud is a research prototype and a decision support tool for human biosafety reviewers, not an autonomous gate. Responsible deployment requires rate limiting and logging of submissions to detect adversarial probing, versioning of the taxonomy and corpus on a defined cadence, and a human-in-the-loop review path for any flagged workflow. We judge that publishing this work strengthens defenders more than attackers, since protocol level screening is the layer at which defenders are currently absent.

LLM Usage Statement

Claude Code was used to assist with scaffolding implementation code. The threat taxonomy, evaluation protocol design, pipeline architecture, and all results were designed and verified by the author. All evaluation runs were executed by the author and results were independently verified against ground truth labels.