BioCompliance: An AML-Inspired Compliance Engine for Benchtop DNA Synthesizers

Sebastián Soto Independent Researcher — AI Safety & Regulatory Compliance

With Apart Research

Abstract

Current U.S. biosecurity legislation (S.3741) mandates synthesis screening but places obligations on providers, not on benchtop synthesizer hardware post-sale. Edison, Toner & Esvelt (2026) demonstrated that unregulated DNA fragments sufficient to assemble 1918 influenza can be purchased for approximately $3,000 from dozens of providers, none of which verified identity or reported the attempt. Meanwhile, generative models like Evo 2 can now design functional viral genomes with novel proteins absent from any screening database. We present BioCompliance, an on-device compliance engine that transposes Anti-Money Laundering (AML) and Know-Your-Customer (KYC) frameworks to DNA synthesis. The system implements three deterministic enforcement modules: (1) tiered researcher credentialing, (2) evasion-zone sequence flagging, and (3) an Anti-Structuring Engine that detects suffix-prefix overlaps (15–40bp) in temporal order histories, blocking split-order assembly attacks before physical synthesis. A Biosecurity Officer dashboard with SAR export and built-in red-teaming enables conformity assessment per S.3741 §4(a)(5)(A). Designed as a behavioral complement to sequence screening (SecureDNA, IBBIS commec), BioCompliance targets an invariant that persists regardless of sequence novelty: the physical overlap required for fragment assembly.

---

1. Introduction

The Biosecurity Modernization and Innovation Act of 2026 (S.3741) represents the first U.S. legislation with enforcement teeth for nucleic acid synthesis screening. However, a critical gap remains: **S.3741 places obligations on "covered providers" — not on the end-users of benchtop synthesizers once the hardware is delivered.** Under §2(1), once a benchtop synthesizer is in the hands of a DIY/Independent researcher, there is no mandatory on-

device screening, no customer re-verification, and no split-order detection at the hardware level.

Edison, Toner & Esvelt (2026) demonstrated the severity of this gap: unregulated DNA fragments sufficient to generate infectious 1918 influenza virus were acquired from 24 out of 25 non-IGSC providers for approximately $3,000. *"Not one requested authorization, verified our identity, or reported the suspicious attempt to authorities."* A graduate student assembled the fragments on the first attempt.

Simultaneously, generative AI models are accelerating the threat. Evo 2 (Merchant et al., 2025) generated 16 functional phage genomes from scratch, including proteins *"evolutionarily distant"* from anything in nature — biology that cannot be caught by sequence-homology screening against known databases.

Our main contributions are:

1. **A novel Anti-Structuring Engine** that adapts AML "smurfing" detection to DNA synthesis, mathematically detecting suffix-prefix overlaps (15–40bp) to block split-order assembly attacks — a vulnerability identified by Edison et al. but not addressed by existing screening tools.
2. **A tiered compliance architecture** that transposes FATF/FIU regulatory frameworks (KYC, SAR, BSO) to benchtop synthesizers, filling the post-sale enforcement gap in S.3741.
3. **A three-layer governance framework** positioning synthesis compliance as the physical choke point that remains effective when model licensing (open-weight models) and sequence screening (AI-designed novel biology) are evadable.

---

2. Related Work

Sequence-based screening. SecureDNA (Sherman et al.) uses cryptographic DOPRF hashing to detect hazardous sequences with negligible false alarms. IBBIS Common Mechanism (commec) provides open-source screening for sequences ≥150bp. Both operate by matching queries against databases of known dangerous sequences — and both are limited when sequences are novel or AI-designed.

Split-order detection. Recent work on defending synthetic DNA orders against splitting-based obfuscation (Diggans & Leproust, 2025) proposes server-side detection of split orders across commercial providers. BioCompliance extends this approach to **on-device,**

real-time detection at the synthesis endpoint — the physical choke point before DNA becomes reality — rather than relying on inter-provider data sharing.

Regulatory landscape. S.3741 mandates customer screening (§4(a)(2)), split-order detection (§4(a)(1)(A)), and red-teaming procedures (§4(a)(5)(A)) for covered providers. The NTI "Three-Legged Stool" framework identifies customer screening, sequence screening, and provider verification as essential pillars. IFP's "Securing Benchtop DNA Synthesizers" report proposes phone-home screening infrastructure for on-device compliance.

Researcher credentialing. Jonas Sandbrink (Sentinel Bio, Track 4 sponsor) advocates for identity verification and tiered access controls at the point of use for bio-capable AI systems. Feldman, Feldman & Anton (2026) formalize this as "Know Your Scientist" (KYS) — a three-tier KYC framework for biological AI tools that shifts governance from content inspection to user verification. Their Tier I uses institutions as trust anchors, Tier II screens outputs via homology, and Tier III monitors behavioral patterns over time. Critically, KYS targets **API-level access control for AI design tools** (AlphaFold, ESM3, RFdiffusion). BioCompliance addresses the complementary downstream gap: **on-device compliance at the physical synthesis endpoint** — the choke point where AI-designed sequences become physical DNA.

How BioCompliance differs. Existing tools screen *what* is being synthesized. BioCompliance screens *who* is synthesizing, *how* they order, and whether their order patterns exhibit assembly signatures. This behavioral layer is orthogonal to sequence screening and remains effective against AI-designed novel sequences that have no database match. The AML→biosecurity transposition is, to our knowledge, novel in this field.

---

3. Methods

3.1 The Compliance Operating System

BioCompliance operates as a **deterministic, transactional rule engine** — no black-box models, no opaque scoring. Every decision is traceable to a specific rule, a core requirement for regulatory compliance and auditability. The system comprises three enforcement modules:

Module 1 — KYC & Device Telemetry: Maps the user's institutional identity to a Risk Tier (1–4), from DIY/Independent (highest risk) to BSL-3/4 Government Labs (trusted). Logs the

hardware MAC address for every synthesis attempt. In production, tier assignment would be verified through ORCID API integration and institutional email domain validation, following the Feldman et al. (2026) model where institutions vouch for affiliated researchers and assume accountability for vetting. The MVP uses self-declaration as a functional prototype. Enforces S.3741 §4(a)(2).

Module 2 — Evasion-Zone Flagging: Flags short sequences (<50bp) that bypass standard homology filters. Detects engineering markers: BsaI recognition sites (Golden Gate), Gibson Assembly proxy motifs. This module is critical because it catches Golden Gate Assembly — an alternative to Gibson that uses 4bp cohesive ends below the Anti-Structuring Engine's 15bp detection floor. Calibrated to known evasion strategies documented by Edison et al.

Module 3 — Anti-Structuring Engine: Analyzes the temporal order history of a specific user/device and mathematically detects suffix-prefix overlaps (15–40bp) between current and historical orders. Blocks "Split Orders" before physical synthesis. Directly implements S.3741 §4(a)(1)(A).

3.2 Anti-Structuring Detection

The core algorithm is inspired by AML "structuring" (smurfing) detection. For any previously ordered fragment *A* and incoming fragment *B*, the system evaluates:

```
suffix(A, k) == prefix(B, k)  OR  suffix(B, k) == prefix(A, k)
for k ∈ [15, 40]
```

The 15–40bp interval corresponds to the practical overlap range for Gibson Assembly — the dominant assembly method in molecular biology. The Anti-Structuring Engine targets this as a high-coverage detection heuristic, not as an absolute invariant.

Evasion analysis. We acknowledge three attack vectors that circumvent suffix-prefix overlap detection: (1) **Golden Gate Assembly** uses BsaI recognition sites generating 4bp cohesive ends below the 15bp floor — but Module 2 explicitly detects BsaI sites, providing defense-in-depth. (2) **Ligation-based assembly** with synthetic adapters avoids sequence overlaps entirely — but has significantly lower efficiency for fragments >500bp and requires specialized reagents, raising the operational cost of evasion. (3) **Internal overlaps** embedded within fragments rather than at termini would evade suffix-prefix comparison — extending detection to all k-mer matches within fragments is a priority for post-hackathon development. No single detection method covers all assembly strategies; the system's

strength lies in combining overlap detection (Module 3), engineering marker detection (Module 2), and behavioral monitoring (Module 1) as layered defenses.

Scalability path. The current MVP uses brute-force comparison (O(n × k) per order against history). For institutional-scale deployment, proven sublinear alternatives exist: Suffix Automata / BNDM achieve O(n) search independent of history size; Parallel Failure-less Aho-Corasick (PFAC) provides linear-time multi-pattern matching compatible with FPGA hardware on synthesizers; and Locality-Sensitive Hashing (MinHash/LROD) enables O(1) lookups with encrypted, compressed historical storage. These are implementation-ready optimizations for post-hackathon scaling.

3.3 Tiered Response Model

Legitimate researchers routinely use 15–40bp overlaps for standard cloning. BioCompliance avoids paralyzing research through a **tiered response model**:

User Tier	Overlap Detected	System Response
Tier 1 (DIY/Independent)	Any overlap 15–40bp	**AUTO-BLOCK** + SAR to BSO
Tier 2 (Private Company)	Overlap + evasion-zone (<50bp)	**FLAG** + hold for BSO review
Tier 3 (University/Academia)	Overlap + ≥2 risk flags	**FLAG** + synthesis proceeds pending review
Tier 4 (BSL-3/4 Government)	Any pattern	**LOG** only — SAR available for audit

Table 1: Tiered response model. Banks do not block every large transaction — they file a SAR and escalate. The same principle applies here.

Additionally, standard cloning vectors (pUC19, pET series, pBR322) have well-characterized overlap signatures that can be **whitelisted** — analogous to how banks whitelist recurring transfers between known accounts. This further reduces false positives for routine molecular biology without weakening detection of anomalous fragmentation patterns.

3.4 Phone-Home Architecture

Each benchtop synthesizer runs the rule engine locally. Blocked orders and SARs transmit to a central BSO server, enabling cross-device split-order detection (S.3741 §4(a)(1)(A)), fleet-wide anomaly monitoring, and offline resilience with queued uploads.

4. Results

4.1 Interactive Proof of Concept

The Streamlit-based PoC demonstrates three workflows:

Figure 1 — KYC Onboarding: Researchers classified by institutional tier before accessing synthesis.

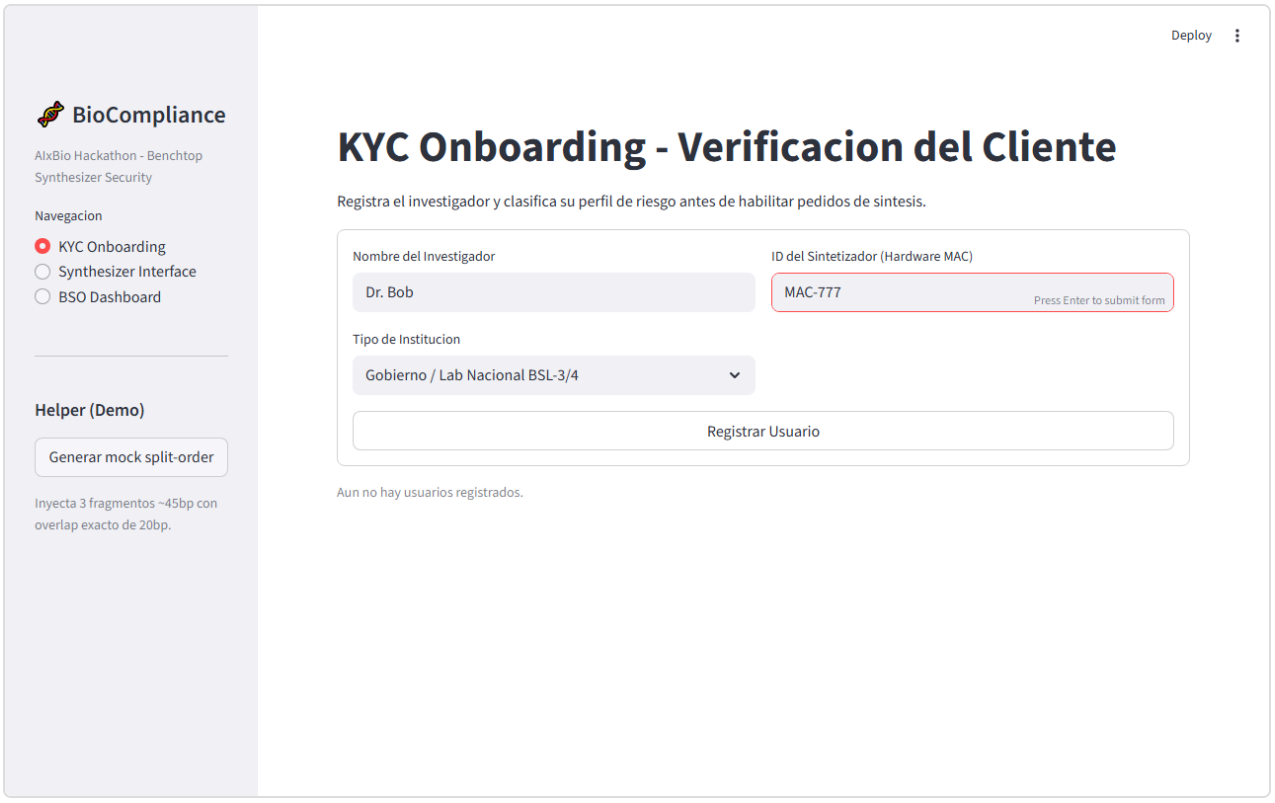


Figure 2 — Split-Order Attack Blocked: Three overlapping sub-50bp fragments submitted. Anti-Structuring Engine detected 20bp suffix-prefix overlap and blocked the order.

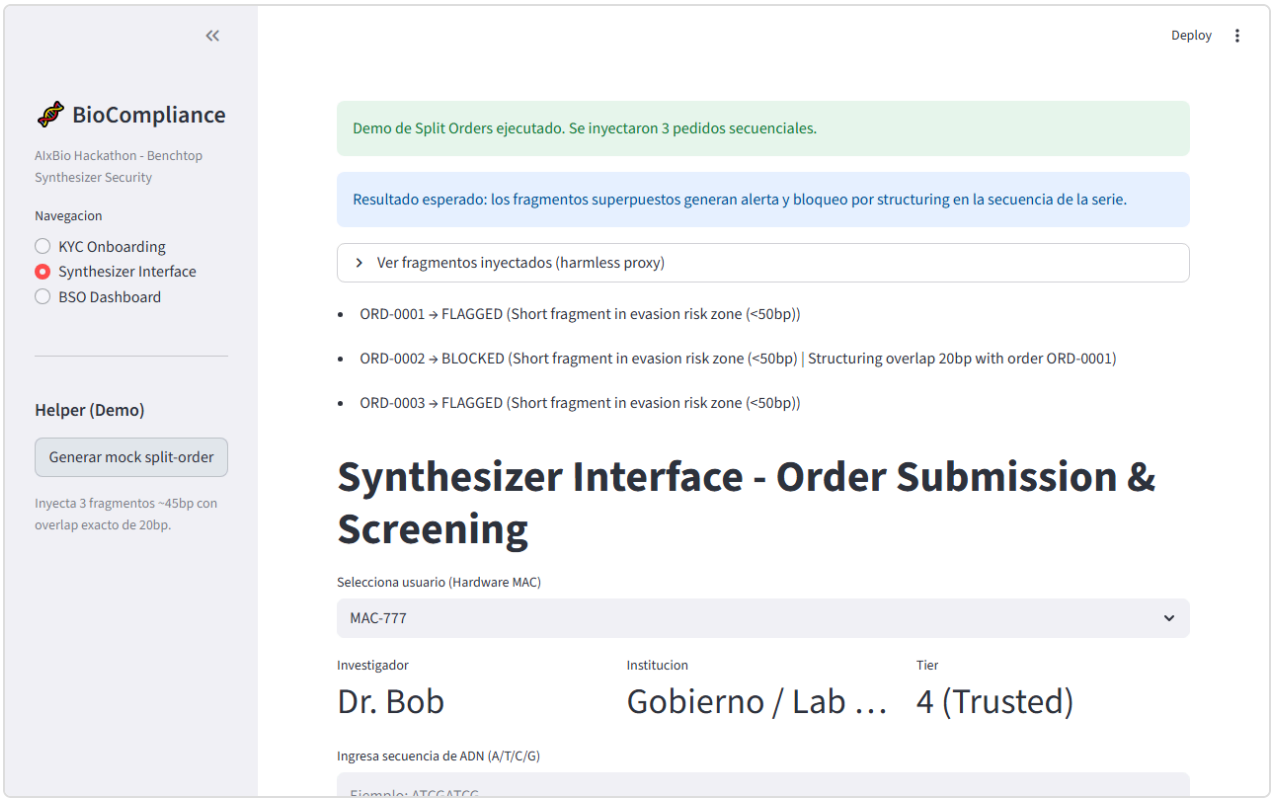
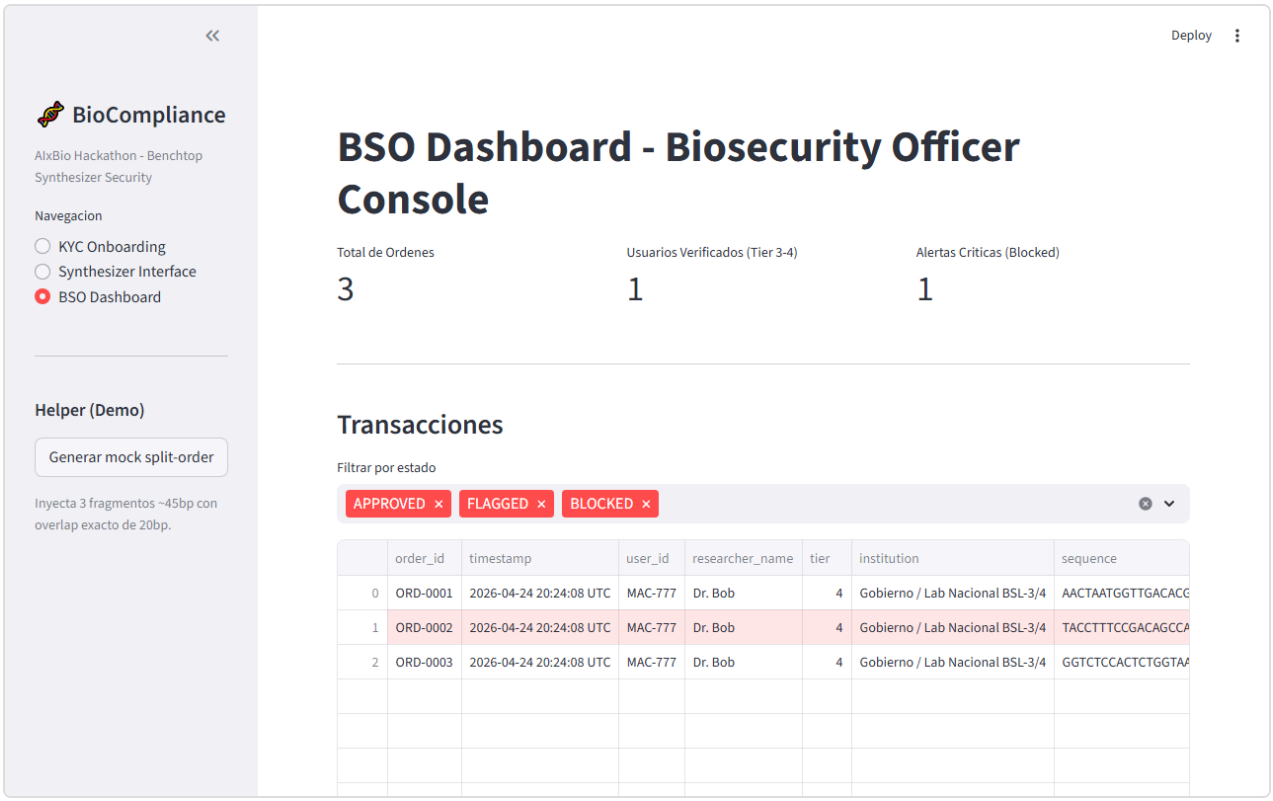


Figure 3 — BSO Dashboard: Biosecurity Officer console with red-flagged orders and SAR export.



4.2 Threat Scenario Walkthrough

Scenario: A Tier 1 individual attempts to reconstruct a 1,500bp gene by ordering three ~500bp fragments with 20bp Gibson Assembly overlaps across 48 hours.

Step	Adversary Action	BioCompliance Response
1	Registers with DIY affiliation	KYC assigns **Tier 1**. MAC logged.
2	Submits Fragment A (500bp)	**Approved** — no prior history. SAR logged.
3	Submits Fragment B (500bp) 24h later	**20bp overlap detected. AUTO-BLOCKED.** SAR filed.
4	Modifies Fragment B to remove exact overlap	**18bp partial overlap detected. BLOCKED.**
5	BSO receives cumulative SARs	Pattern: Tier 1, sequential assembly overlaps, 24h velocity. **Escalated.**

Table 2: Even if evasion succeeds on a single pair, the cumulative SAR trail creates an audit record the BSO can act on — mirroring AML investigation principles.

4.3 Built-In Red-Teaming

S.3741 §4(a)(5)(A) mandates *"adversarial testing at random intervals to ensure compliance."* The "Generar mock split-order" button stress-tests the Anti-Structuring Engine on demand, enabling conformity assessment without external penetration testing.

---

5. Discussion and Limitations

The Three-Layer Governance Stack

BioCompliance occupies a specific and currently unoccupied position in the biosecurity governance pipeline:

Governance Layer	Provider	Controls	Evasion Vector
Model Licensing	Governments / Sandbrink	Who can **DESIGN** sequences	Open-weight models (Evo 2 is open)
Sequence Screening	SecureDNA / commec	**WHAT** is dangerous	Novel biology (no database match)
Synthesis Compliance	**BioCompliance**	**WHO/HOW/PATTERN**	High-coverage heuristic: dominant assembly method targeted

Table 3: The synthesis endpoint is the only physical choke point where digital threats become biological reality. Without the third layer, the first two are evadable.

AI-Designed Biology and the Ceiling of Homology Screening

Evo 2 generated 16 functional phage genomes including proteins absent from any known organism (Merchant et al., 2025). Mutations in 13 of these genomes *"could not be recapitulated from any known natural sequences."* This demonstrates a fundamental ceiling for sequence-homology screening: it cannot flag biology that has never existed. BioCompliance's behavioral approach remains effective regardless of whether the ordered sequence is known, novel, or AI-designed.

Dual-Use Considerations

Proxy sequences only. Engineering markers used in the prototype (BsaI, Gibson proxy motifs) are benign demonstration proxies. No hazardous biological material or sequences were used.

Actor differentiation. Nation-states can bypass commercial compliance through indigenous BSL-4 facilities. Non-state actors rely on exploiting regulatory seams in commercial providers and benchtop synthesizers — making them the primary threat population for transactional pattern detection (Yassif, NTI).

LLM-assisted evasion. Edison et al. warn that *"large language models can suggest fragmentation and evasion strategies."* Our layered defense combines overlap detection (Module 3), engineering marker detection (Module 2), and behavioral monitoring (Module 1) to raise the cost of evasion across multiple assembly strategies simultaneously.

Responsible disclosure. This project demonstrates detection methods, not evasion methods. The split-order attack vector was publicly documented by Edison et al. (2026) and S.3741's legislative text.

Limitations

Proof of concept scope. The prototype runs as a single Streamlit instance. Phone-home cross-device correlation is designed but not yet implemented.

No sequence homology. BioCompliance screens behavioral metadata, not sequence content. Integration with SecureDNA's API is a post-hackathon priority for a unified screening stack.

Single-researcher testing. Built and tested by one researcher in 48 hours. Broader usability testing and institutional feedback are needed.

Scalability untested. The Anti-Structuring algorithm is O(n × k) per order against order history. Performance with large institutional histories (10,000+ orders) requires suffix-indexing optimization.

Future Work

1. Deploy on-device compliance on commercial benchtop synthesizer firmware
2. Integrate SecureDNA API for combined behavioral + sequence screening
3. Cross-device split-order correlation via centralized BSO server
4. Policy engagement: propose biosecurity FIU node to receive SARs from providers and on-device systems

---

6. Conclusion

We presented BioCompliance, a deterministic compliance engine that transposes Anti-Money Laundering frameworks to DNA synthesis security. By targeting the physical invariant of assembly overlaps (15–40bp) rather than sequence content, the system addresses a fundamental limitation of homology-based screening that is increasingly acute as generative AI models design functional biology absent from any known database. The tiered KYC + Anti-Structuring + BSO architecture fills the post-sale enforcement gap in S.3741 and provides a regulatory framework that policymakers already understand — directly translating from FATF/FIU infrastructure that has been operational in the financial sector for decades.

Code and Data

Code repository: GitHub — BioCompliance **Live demo:** `python -m streamlit run app.py --server.port=8501` **No hazardous sequences** were used in any stage of this project.

LLM Usage Statement

We used Claude (Anthropic) to assist with code implementation, report drafting, and regulatory analysis. All claims, results, and architectural decisions were independently verified and designed by the author. The core AML→biosecurity transposition concept, Anti-Structuring algorithm design, and policy recommendations are original contributions.

References

1. Edison, R., Toner, S. & Esvelt, K. M. "Assembling unregulated DNA segments bypasses synthesis screening: regulate fragments as select agents." *Nature Communications* **17**, 3189 (2026). https://doi.org/10.1038/s41467-025-67955-3
2. S.3741 — Biosecurity Modernization and Innovation Act of 2026. 119th Congress, 2nd Session.
3. Sherman, B. et al. "A system capable of verifiably and privately screening global DNA synthesis." SecureDNA. https://securedna.org
4. IFP. "Securing Benchtop DNA Synthesizers." Institute for Progress.

5. NTI | bio. "Benchtop DNA Synthesis Devices: Capabilities, Biosecurity Implications, and Governance" (2023).
6. Merchant, A. et al. "Generative design of bacteriophages with genome language models." *bioRxiv* (2025). Arc Institute & Stanford. https://doi.org/10.1101/2025.09.12.675911
7. Feldman, J., Feldman, T. & Anton, A. I. "Know Your Scientist: KYC as Biosecurity Infrastructure." *arXiv:2602.06172* [cs.CR] (2026). Georgia Institute of Technology & Yale Law School.
8. Diggans, J. & Leproust, E. "Defending Synthetic DNA Orders Against Splitting-Based Obfuscation." *bioRxiv* (2025). https://doi.org/10.1101/2025.03.12.642526
9. Puzis, R. et al. "Increased Cyber-Biosecurity for DNA Synthesis." *Nature Biotechnology* (2020). https://doi.org/10.1038/s41587-020-00761-y

---

Appendix: Limitations and Dual-Use Considerations

Prototype Scope and Honest Assessment

BioCompliance is a proof-of-concept built in 48 hours by a single researcher. The MVP demonstrates architectural design and detection logic, not production-grade security. Key gaps between the PoC and a deployable system:

- **KYC verification:** The prototype uses self-declared tier assignment (free text input). A production system requires institutional verification via ORCID API, institutional email domain validation, and the institutional vouching model described by Feldman et al. (2026). Without verified identity, the tier system provides no real security — only a workflow demonstration.
- **No quantitative evaluation:** The threat scenario walkthrough (Table 2) is a constructed demonstration, not a measured evaluation. A rigorous assessment would require a synthetic dataset of benign Gibson Assembly orders and adversarial split-orders, with measured false positive and false negative rates. This is a priority for post-hackathon development.
- **Detection coverage:** The Anti-Structuring Engine detects Gibson Assembly overlaps (15–40bp) but does not cover all assembly methods. Golden Gate Assembly is partially mitigated by Module 2's BsaI detection. Ligation-based assembly and internal (non-terminal) overlaps are not detected. No single detection layer is sufficient — the system's design relies on layered defense across all three modules.

Evasion Vectors and Adversarial Analysis

Attack Vector	Detection Status	Mitigation
Gibson Assembly (15–40bp overlaps)	✅ Detected by Module 3	Core detection capability
Golden Gate Assembly (BsaI, 4bp ends)	⚠️ Partially detected by Module 2	BsaI recognition site flagging
Ligation-based (synthetic adapters)	❌ Not detected	Low efficiency for >500bp fragments; high operational cost
Internal overlaps (non-terminal)	❌ Not detected	Requires k-mer scanning extension (future work)
Longer fragments avoiding overlap window	⚠️ Partially mitigated	Evasion-zone flagging + velocity monitoring

We do not claim that the Anti-Structuring Engine is an absolute invariant. It is a high-coverage heuristic targeting the dominant assembly method, reinforced by engineering marker detection and behavioral monitoring.

Dual-Use Risks

This project demonstrates a **detection system**, not an evasion toolkit. The split-order attack vector was publicly documented by Edison et al. (2026) in *Nature Communications* and is explicitly referenced in S.3741. No hazardous biological sequences or materials were used.

Ethical Considerations

The KYC tier system introduces differential friction by institutional affiliation. This mirrors how financial regulations apply enhanced due diligence to higher-risk customers, but requires careful calibration to avoid disproportionate burden on independent researchers. Future work should include community input on tier thresholds.

Scalability Constraints

The current overlap detection is brute-force (O(n × k) per order). For institutional-scale deployment (10,000+ orders per device), sublinear alternatives are implementation-ready: Suffix Automata (BNDM) for O(n) search, Aho-Corasick (PFAC) for FPGA-compatible linear matching, and MinHash/LSH for O(1) lookup with encrypted history storage. The phone-home architecture for cross-device correlation is designed but not implemented in the MVP.

Suggestions for Future Improvements

1. Integrate SecureDNA API for combined behavioral + sequence screening in a single pipeline
2. Implement cross-device split-order correlation via encrypted SAR aggregation at the BSO server
3. Deploy as firmware module on commercial benchtop synthesizers (Ansa Biotechnologies, DNA Script)
4. Propose establishment of a dedicated biosecurity FIU node for centralized SAR triage
5. Conduct usability testing with partner institutions to calibrate tier thresholds and minimize legitimate research friction
6. Defend against remote DNA injection attacks (Puzis et al., 2020) where malware modifies sequences in transit between design software and the synthesizer — the phone-home architecture provides a natural integration point for transit integrity verification