OliGraph: graph-based screening of large oligopools[1]

Christopher Kent
University of Birmingham

Samuel A.S. Richardson
University of Birmingham

Ryan Teo
University of Birmingham

With
Apart Research

Abstract

Existing synthesis screening tools cannot evaluate short oligonucleotide pools, whose overlapping fragments can be reassembled into regulated sequences via polymerase cycling assembly (PCA) yet fall below gene-length detection thresholds. We present OliGraph, an open-source tool that constructs a bi-directed overlap graph from an oligonucleotide pool and extracts contigs for downstream gene-length screening. An optional PCA mode retains only cross-strand overlaps consistent with PCA chemistry. We validated OliGraph in a blinded study across ten simulated pools (70–9,184 oligonucleotides, 30–300 bp) spanning four risk categories. BLAST screening of individual oligonucleotides failed to identify sequences of concern in most pools: three returned zero hits, and vector noise obscured true positives in the remainder. After OliGraph assembly, contig-level BLAST matched the longest assembled sequences (up to 1,905 bp) to sequences of concern at 97–100% identity. In one pool, assembly collapsed 1,634 individual BLAST results into 10 hits from a single contig, all assigned to the same source organism. PCA mode correctly distinguished assemblable from non-assemblable fragments within the same pool. Two pools with no assemblable structure yielded no contigs. OliGraph processed all pools in under 0.2 seconds, fast enough for real-time order screening and consistent with proposals to bring oligonucleotide orders within the scope of synthesis screening regulation.

[1] Research conducted at the AIxBio Hackathon, April 2026

1. Introduction

Screening synthetic DNA orders against databases of regulated sequences is one of the most effective biosecurity interventions available, because synthesis providers occupy a narrow bottleneck between digital sequence design and physical biological material. As AI-driven tools lower the expertise required to design functional biological constructs, this bottleneck grows more consequential: a broader range of actors, including non-state groups with limited laboratory infrastructure, can now translate a digital sequence into a viable agent

Existing screening systems perform well on gene-length orders, where hundreds or thousands of contiguous base pairs provide enough context for alignment or signature-based detection. Regulation is beginning to catch up: the EU Regulation on the responsible use of biotechnologies screens individual synthetic nucleic acids from 50 nt and, under criterion (c), requires providers to assess whether sequences in a bulk order could be assembled into a regulated sequence of 200 nt or longer (European Commission, 2025). Yet as far as we are aware, no publicly available tool exists to perform this assembly-aware assessment. Large oligonucleotide pools, orders of tens to hundreds of thousands of unique 30–80 bp sequences, are now commercially available at low per-base cost. A sequence of concern tiled across many short overlapping oligonucleotides can be reassembled by the purchaser through polymerase cycling assembly (PCA), while no individual fragment is long enough to trigger existing screening. Providers screen each order in isolation and have no visibility into how fragments from separate orders or providers might be combined after delivery.

The gap between what screening systems can detect and what oligonucleotide pools can encode has been recognised for some time. Diggans and Leproust (2019) proposed that providers computationally assemble oligonucleotide pools into longer contiguous sequences prior to screening, but to our knowledge no publicly available tool implements this approach. Demonstrating that such assembly is technically feasible and computationally efficient matters for two reasons. First, it provides synthesis providers with a practical mechanism to screen a product class they currently cannot evaluate, closing a route by which screening might be circumvented. Second, it gives policymakers evidence that extending screening requirements to short oligonucleotide orders is operationally realistic, supporting calls to lower the minimum sequence length thresholds in regulatory frameworks (Kane and Parker, 2024; Fady *et al.*, 2025).

We present OliGraph, an open-source tool that addresses this gap. Our main contributions are:

1. A screening-oriented assembly tool for oligonucleotide pools that constructs a bi-directed overlap graph and extracts contigs suitable for downstream gene-length screening.
2. An interactive web interface for exploring the overlap graph and assembled contigs, designed to support manual review workflows
3. A blinded validation across ten simulated oligonucleotide pools of varying size, sequence content, and risk category.

2. Related Work

Current approaches to synthetic DNA order screening rely on several computational strategies. Alignment-based "best match" screening, specified in the 2010 U.S. HHS guidance (Adam *et al.*, 2011) and implemented in tools such as BLiSS (Simirenko *et al.*, 2016), fragments orders into 200 bp sliding windows and runs BLAST (Camacho *et al.*, 2009) against databases of regulated pathogens. Six-frame translation allows detection of codon-optimised variants, but the approach is slow, flags conserved housekeeping genes at high rates, and can be circumvented by synonymous mutations that shift the best-match assignment away from controlled organisms (Gretton *et al.*, 2025). k-mer and signature methods such as FAST-NA (Beal *et al.*, 2021) trade sensitivity for speed through exact matching of short fragments (typically $k = 31$), but fail below approximately 80% nucleotide identity (Hoffmann *et al.*, 2023). SecureDNA takes a different approach, using cryptographic protocols to screen orders as short as 30 bp against precomputed exact-match targets including predicted functional variants (Baum *et al.*, 2025), but adoption across the global synthesis industry remains incomplete.

All of these methods perform poorly on short oligonucleotides (Figure 1). Sequences below 200 bp fall outside the operational design of alignment-based screening (Adam *et al.*, 2011), and a 50 bp oligonucleotide yields at most 20 unique 31-mers. Even a few point mutations can eliminate every exact k-mer match to a reference. Lowering screening thresholds increases false-positive rates enough to overwhelm manual review workflows that already require non-trivial human expertise and time cost (Fady *et al.*, 2025). A 2022 U.S. DHHS draft proposing screening of oligonucleotides as short as 20 nt was withdrawn after the community questioned its feasibility (Kane and Parker, 2024).

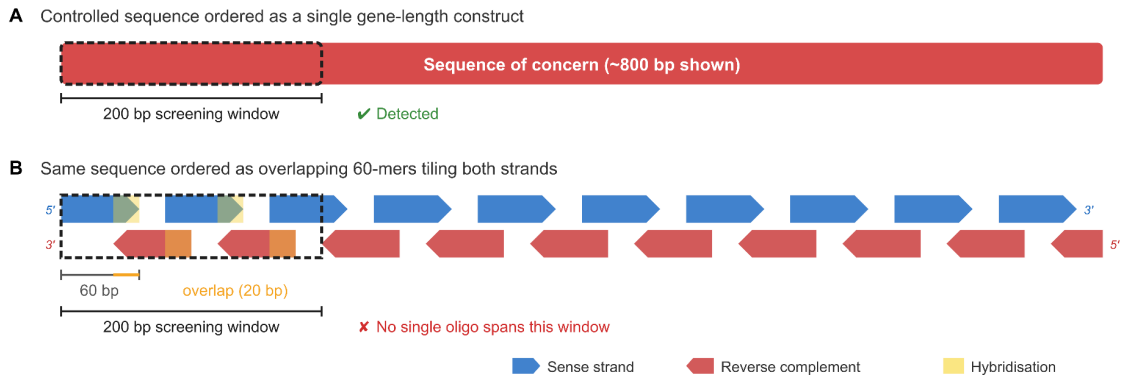

Figure 1. Evasion of alignment-based synthesis screening by oligonucleotide pool fragmentation. (A) A controlled sequence ordered as a single gene-length construct. The 200 bp sliding window used by screening tools (dashed box) fits within the order. **(B)** The same sequence is ordered as a pool of overlapping 60 bp oligonucleotides (blue tiles; overlaps highlighted in yellow). No individual oligonucleotide spans the 200 bp screening window, so none triggers alignment-based detection. Synthesis providers screen orders independently and cannot infer that a pool of short fragments tiles a controlled sequence.

This screening gap is particularly consequential because of Polymerase Cycling Assembly (PCA) (Stemmer *et al.*, 1995), a method in which carefully designed overlapping oligonucleotides anneal and are extended by a thermostable polymerase over repeated denaturation–annealing–extension cycles, remains viable for fragments as short as 25–30 bp (Fady *et al.*, 2025). Oligonucleotide pools are therefore a feasible, not merely theoretical, route past synthesis screening.

3. Methods

OliGraph is a Rust pipeline that takes a pool of oligonucleotide sequences as input, identifies pairwise overlaps between them, and assembles overlapping fragments into longer contiguous sequences (contigs) that represent what a pool could plausibly produce upon assembly, following the biological constants of PCA. The output contigs can then be input into existing gene-length screening tools.

3.1 Overlap graph construction

OliGraph represents the oligonucleotide pool as a bi-directed overlap graph, a structure commonly used in *de novo* genome assemblers (Rizzi *et al.*, 2019). Since oligonucleotide orders are primarily single stranded DNA (ssDNA), each oligonucleotide is provided in the 5' to 3' direction. This leads to two types of PCA-productive overlap:

- **Primary:** the 3' end of an oligo would be able to hybridise to the 3' end of any other oligos (Figure 2A).
- **Secondary**: 5' to 5' complementary. This is used to assemble the products of two adjacent primary fusion into a longer construct, but this alone cannot form an extendable structure from the ordered oligos (Figure 2B). However, the primary fusion creates the viable reverse complement *in situ* (Figure 2C).

Other kinds of terminal overlaps are also reported in the "all" run mode. Each oligonucleotide is a node in the graph, and an edge is added between any two nodes with complementary overlap.

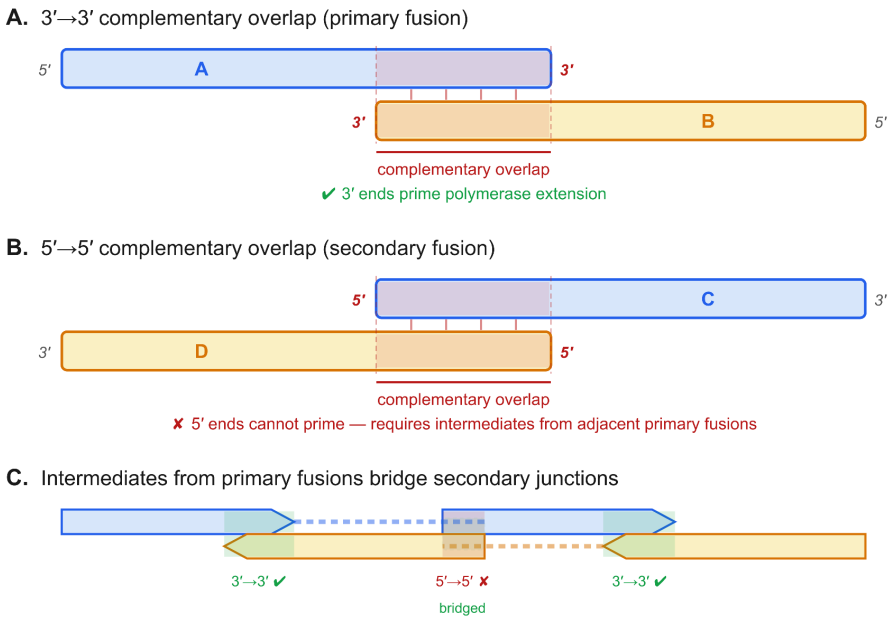

Figure 2. Overlap types in ssDNA oligonucleotide assembly graphs. (A) Primary overlap: the 3' suffixes of two oligos are reverse-complementary, allowing both 3' ends to prime polymerase extension upon hybridisation. **(B)** Secondary overlap: the 5' prefixes are reverse-complementary, but 5' ends cannot prime extension directly. **(C)** In a valid assembly path, secondary (5'→5') junctions are bridged by intermediate products generated from flanking primary (3'→3') fusions, shown by the dashed lines, yielding the alternating path structure (3'→3' · 5'→5')[bridged] · 3'→3'.

Overlaps are detected using a seed-and-extend strategy. A hash index maps the first $l_{\min}$ bases of every oriented sequence to its identifier. Each sequence is then scanned with two rolling seeds (one forward, one reverse-complement) so that a single pass discovers all overlap types. Seed matches are then verified by base-by-base extension. The minimum overlap $l_{\min}$ (default 15 bp, max 64 bp) controls sensitivity: shorter values detect more overlaps but risk spurious matches from shared motifs. The tool supports oligonucleotides up to 320 bp, exceeding the typical range of commercial pools (30 – 200 bp).

An optional mode restricts the graphs to overlaps consistent with PCA, in which alternating forward and reverse fragments anneal through their 3’ ends. When enabled, forward-to-forward edges are dropped, retaining only the cross-strand overlaps (Types B and C in Figure 2) through which PCA operates.

3.2 Contig assembly

Connected components are identified and isolated nodes (no overlaps) are discarded. Within each component, contigs are assembled by greedy bidirectional walk: starting from a tip node (one with neighbours in only one strand direction), the walk extends in both directions, always following the highest-overlap unvisited neighbour. Ties are broken by sequence identifier for reproducibility. The

consensus sequence is produced by concatenating oriented subsequences with shared overlap regions trimmed. Paths that loop back to their starting node are marked as cyclic.

In PCA mode, the walk is further constrained to follow the alternating annealing pattern of polymerase cycling assembly. Because adjacent oligonucleotides in a PCA pool anneal in a controlled way valid assembly path must strictly alternate between primary and secondary overlaps, following the motif $(3'\rightarrow3' \cdot 5'\rightarrow5')_n \cdot 3'\rightarrow3'$ where $n \geq 0$ (Figure 2C). At each step the walker only considers neighbours reachable by the next required overlap type; edges that would violate the alternation are skipped. A greedy heuristic was chosen over exact methods because the goal is not a single optimal assembly as in *de novo* genome assembly, but a practical set of contigs sufficient for downstream screening.

3.3 Interactive web interface

Because screening decisions ultimately depend on human judgement, OliGraph includes an interactive web interface (https://teojcryan.github.io/oligraph-rs/) that lets operators visually inspect the overlap structure of an oligonucleotide pool before deciding whether to escalate an order. A reviewer can upload a FASTA file and immediately see which oligonucleotides share overlaps, how they cluster into assemblable groups, and whether those groups form the long linear chains characteristic of deliberate gene construction or the short, branched fragments typical of routine molecular biology. Adjusting the minimum overlap threshold or toggling PCA mode updates the graph instantly, so an operator can test whether a suspicious cluster persists under stricter parameters or disappears, a practical check that would be difficult to perform from tabular BLAST output alone. Assembled contigs can be downloaded individually for follow-up screening. The entire application runs locally in the browser via WebAssembly; no sequence data leaves the user's machine, removing a barrier to adoption for providers handling proprietary or sensitive order data.

3.4 Model validation

To assess whether OliGraph could correctly reconstruct contigs from fragmented oligonucleotide pools, we designed a blinded validation study. Ten hypothetical sets of oligonucleotides were assembled using publicly available sequencing data, to represent a range of realistic screening scenarios, varying in pool size (70 to over 9,000 sequences), oligonucleotide length (30 - 165 bp), sequence content (innocuous or concerning sequences), and pool structure (overlapping, or non-overlapping). Each oligonucleotide set was assigned into one of four risk categories:

- **negligible:** no relation to pathogen or toxin sequences;
- **low:** contains pathogen- or toxin-derived sequences but could not realistically be used for gene assembly;
- **moderate:** contains pathogen or toxin sequences tiled into overlapping oligonucleotides suitable for gene assembly; or
- **high:** contains sequences categorised under bioterrorism legislation, tiled for assembly.

Oligonucleotide sets in FASTA format were analysed in OliGraph with a minimum overlap size (l_{min}) of 15 bp. The operator was blinded to the input data, contents and risk category of each file. Top contigs detected by the model were interrogated using nucleotide BLAST. Descriptors of oligonucleotide sets are shown in Table 1.

4. Results

The ten oligonucleotide sets described previously were analysed with OliGraph in PCA mode using the default minimum overlap of 15 bp, then screened by nucleotide BLAST against the NCBI core nucleotide database at both the individual oligonucleotide level and the OliGraph-assembled contig level. Results are summarised in Table 2.

4.1 OliGraph assembles large oligonucleotide pools quickly and accurately

OliGraph completed assembly for all ten sets in under 0.2 seconds, including the three pools of 9,184 oligonucleotides (SET1, SET2, SET10), which each required approximately 0.16 s on a single CPU core. Eight of ten sets yielded between 1 and 120 contigs, with the longest reaching 1,905 bp (SET6) from 51 input oligonucleotides of 71–74 bp. SET1 and SET8 produced no contigs, as no PCA-consistent overlaps were detected among their sequences. These sets served as true negatives: pools containing no assemblable structure and no sequences of concern.

4.2 Per-oligonucleotide BLAST misses or obscures sequences of concern

BLAST screening of individual oligonucleotides directly produced weak, absent, or noisy results across the majority of sets. Two sets (SET8, SET9) returned zero hits, rendering their contents entirely invisible to per-oligonucleotide screening. For the large pools (SET1, SET2, SET10; 9,184 oligonucleotides each), fewer than 3% of sequences matched any database entry, and the majority of hits were to common laboratory vectors and plasmids. Where sequences of concern (SOCs) were present, they were a minority among vector noise: in SET2, only 26 of 236 (11.0%) oligonucleotides with hits matched an SOC, while the remaining 210 matched vectors. Even where most oligonucleotides returned SOC hits (SET6, SET7), alignment lengths were constrained by oligonucleotide size, reaching at most 74 bp and 300 bp respectively, with correspondingly modest E-values.

4.3 Contig-level screening recovers high-confidence identifications

After OliGraph assembly, BLAST against contigs produced substantially longer alignments and more confident assignments. SET6 exemplifies the contrast (Figure 3): the best per-oligonucleotide alignment was 74 bp ($E = 1.4 \times 10^{-28}$), whereas the largest assembled contig matched the same SOC over 1,875 bp at 99.9% identity ($E = 0.0$). SET2 showed a comparable improvement (Figure 4), with its single contig of 1,393 bp matching an SOC at 98.4% identity over 1,400 bp ($E = 0.0$), an unambiguous identification compared with the 26 scattered oligonucleotide-level hits of at most 109 bp. For SET10, three contigs (332–457 bp) matched an

SOC at 97–100% identity, where individual oligonucleotide screening had detected only 16 hits among 226 predominantly vector-matching queries.

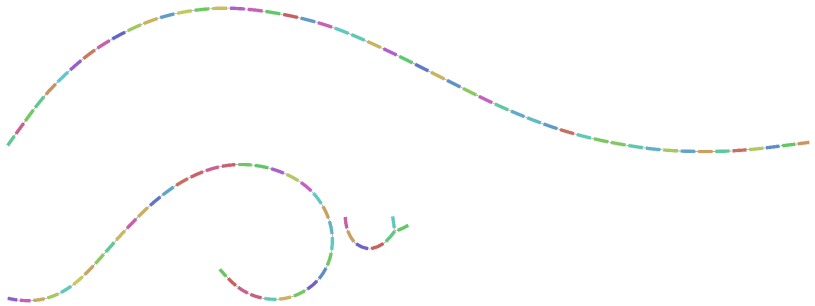

Figure 3. Overlap graph for SET6 (95 oligonucleotides, 71–74 bp). The dominant linear chain comprises 51 oligonucleotides assembling into a 1,905 bp contig matching a SOC at 99.9% identity over 1,875 bp. A second component (37 oligos, 1,367 bp) and a short 6-oligo fragment are also visible. Per-oligonucleotide BLAST identified the SOC but with alignments constrained to 74 bp ($E = 1.4 \times 10^{-28}$); assembly extended this to an unambiguous match ($E = 0.0$).

Assembly also recovered signals where per-oligonucleotide screening had failed entirely:

- SET5, which returned zero oligonucleotide-level hits, yielded 20 contigs after assembly; the largest (665 bp) matched synthetic construct sequences at 75.7% identity.
- SET9, also invisible at the oligonucleotide level, produced a single 36 bp contig matching an SOC.
- The number of BLAST results requiring operator review fell correspondingly: for SET2, from 1,634 hits across 236 query sequences to 10 hits from a single contig, all unambiguously assigned to the same SOC.

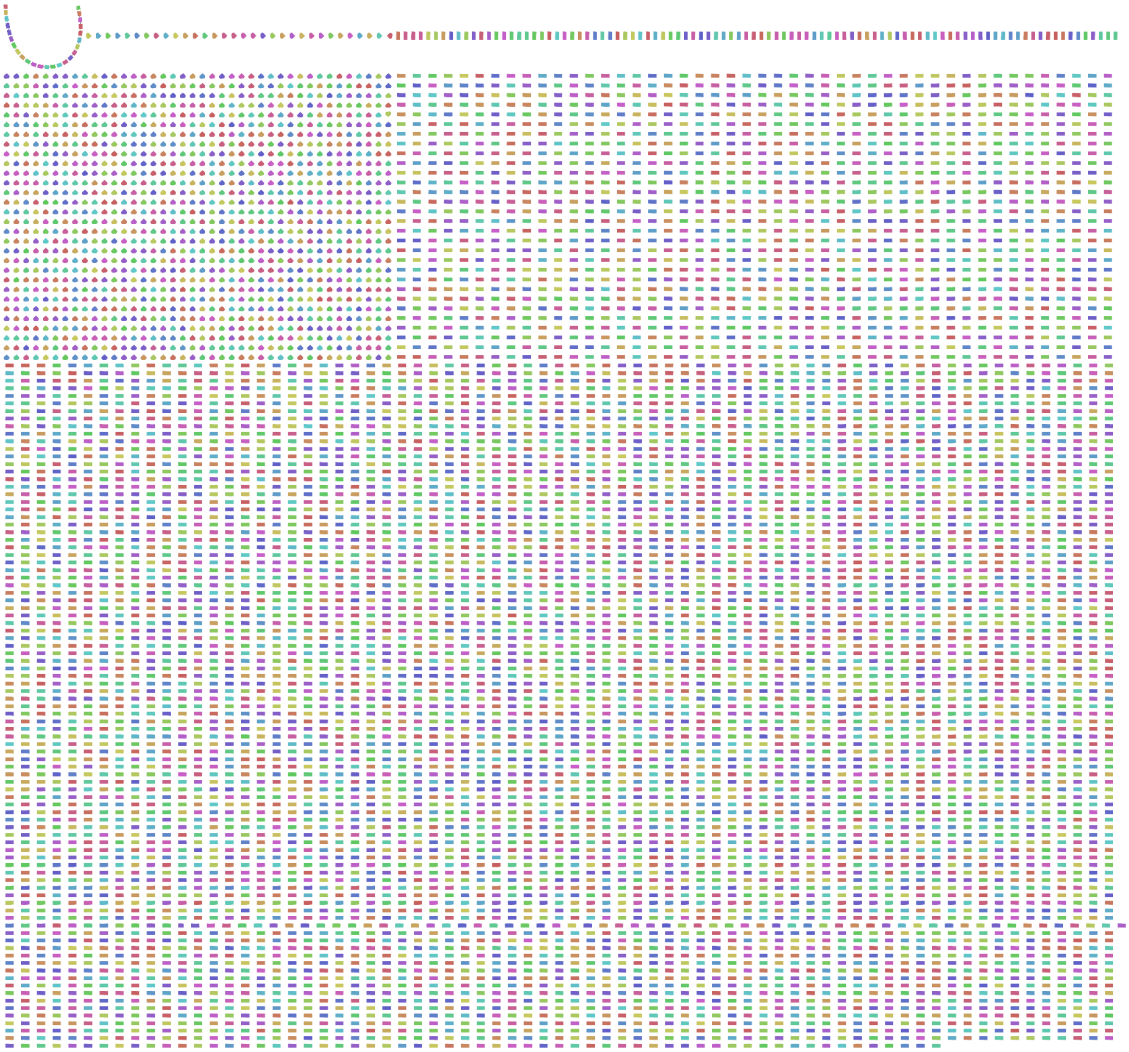

Figure 4. Overlap graph for SET2 (9,184 oligonucleotides). A single connected component of 25 oligonucleotides (top-left) assembles into a 1,393 bp contig matching a sequence of concern at 98.4% identity. The remaining 9,159 non-overlapping background oligonucleotides are shown as isolated nodes. Per-oligonucleotide BLAST returned 1,634 hits dominated by vector noise; contig-level screening collapsed these to 10 hits from one contig, all assigned to the same source organism.

4.4 PCA mode correctly rejects non-assemblable SOC fragments

SET7 illustrates the effect of PCA-specific edge filtering (Figure 5). The set contains 64 oligonucleotides (104–300 bp) that individually match an SOC with high identity, tiling contiguously along the reference genome. However, these oligonucleotides share no cross-strand (3′-to-3′) overlaps and therefore remain isolated nodes under PCA mode, as they cannot be assembled *in vitro* via PCA. A separate subset of 14 shorter oligonucleotides (71–85 bp) does

exhibit PCA-consistent overlaps and assembles into an 825 bp contig matching a different SOC at 100% identity. OliGraph thus correctly distinguishes between sequences of concern that could plausibly be assembled by PCA and those that, despite appearing in the same pool, lack the strand-alternating overlap structure required for assembly.

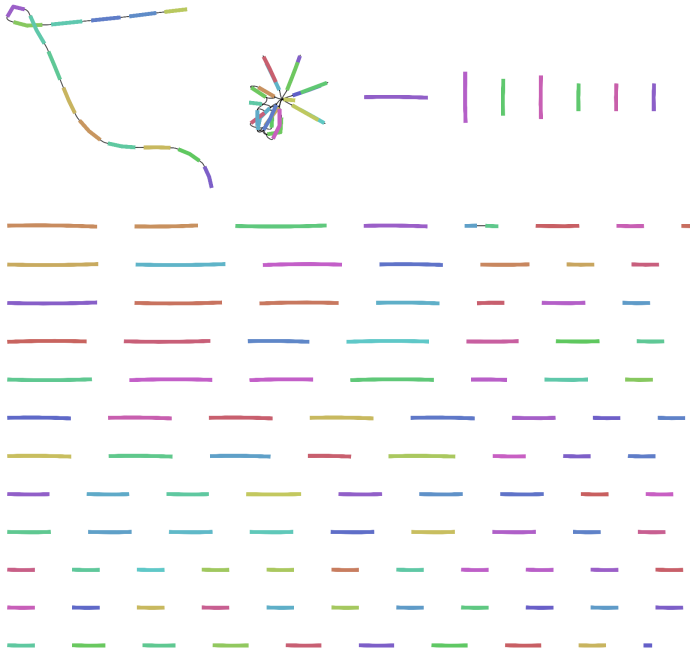

Figure 5. Overlap graph for SET7 (144 oligonucleotides). The linear chain (top-left, 14 oligos) assembles into an 825 bp contig matching a sequence of concern at 100% identity via cross-strand overlaps consistent with polymerase cycling assembly. The dense cluster (top-centre) represents 64 oligonucleotides that individually match a different SOC but share only forward-to-forward overlaps, which are dropped under PCA mode as they cannot support assembly. Isolated nodes (grid) have no overlaps at the 15 bp threshold.

5. Discussion and Limitations

OliGraph demonstrates that computational assembly of oligonucleotide pools prior to screening is both technically feasible and practically efficient, addressing the countermeasure first proposed by Diggans and Leproust (2019) but, to our knowledge, never publicly implemented. By reconstructing contigs from short overlapping fragments, the tool recovers screening signals that individual oligonucleotides are too short to trigger: alignment-based tools such as BLAST, whose sensitivity improves with query length, can operate on the assembled output where they would fail on any single input oligonucleotide. The assembly step therefore acts as both signal recovery (surfacing sequences that were invisible to gene-length screening) and signal amplification, since longer contigs produce higher-scoring, more specific alignments. Noise is simultaneously reduced because spurious short matches are consolidated into coherent contiguous sequences whose taxonomic assignments are less ambiguous. Processing pools of over 100,000 oligonucleotides in

under one second, OliGraph is fast enough to integrate into real-time order-screening workflows without impeding commercial turnaround. The explicit PCA assembly model correctly restricts the overlap graph to cross-strand annealing overlaps, faithfully reflecting the biophysical mechanism through which PCA operates. The PCA mode can also detect splint based ligase cycling reaction (LCR), omitting the PCA filter broadens detection to all LCR strategies and other assembly methods, making the tool applicable across multiple assembly paradigms. These results align with recent policy recommendations that oligonucleotide orders should fall within the scope of synthesis screening, and provide evidence that the technical barriers cited in the withdrawal of the 2022 U.S. DHHS proposal can be overcome (Kane and Parker, 2024).

5.1 Limitations

Several limitations qualify these findings:

1. No robust external comparator was available: we could not integrate our tool with and therefore benchmark against SecureDNA (Baum *et al.*, 2025) or the Common Mechanism (Wheeler, Carter, *et al.*, 2024) due to technical limitations, so our evaluation relies on BLAST interrogation of assembled contigs rather than head-to-head comparison with an established screening tool.
2. The greedy assembly heuristic does not guarantee optimal contig paths; complex or highly branched graphs may yield suboptimal assemblies, though even partial contigs can reveal concerning sequences.
3. OliGraph requires exact sequence matching at overlap junctions, so oligonucleotides with mismatches in their overlap regions will not be joined, potentially missing true positives where inexact overlaps would still support assembly.
4. The validation study used simulated pools derived from public sequence data from academic studies rather than real commercial orders, which may exhibit additional complexity such as adapter sequences, barcodes, or deliberate obfuscation.
5. The tool assumes that a single pool contains all fragments needed for assembly; an adversary distributing fragments across providers or time points would evade any per-order analysis, a limitation shared by all provider-side screening approaches.

5.2 Future Work

Natural extensions include incorporating biophysical constraints such as predicted melting temperature and secondary structure stability to filter thermodynamically implausible overlaps, improving specificity without sacrificing sensitivity. Support for additional assembly chemistries (Golden Gate, Gibson, and other Type IIS restriction-ligation methods) would broaden applicability, as each method leaves characteristic signatures in graph topology (e.g., fixed-length sticky-end overlaps) that could serve as assembly-type classifiers.

Beyond screening, the overlap graph itself encodes rich structural information: graph topology metrics such as component diameter, branching factor, and cycle frequency may distinguish routine molecular biology orders from anomalous pool designs, and could support applications in source

organism identification or the detection of DNA data-storage payloads. Finally, the bi-directed overlap graph may have utility as a lossless compressed representation of oligonucleotide pools, storing redundant overlap regions only once while preserving full sequence recoverability. This could have the potential to serve as a lossless and compressed data structure aiding in the comprehensive and long-term storage of synthetic DNA orders by providers.

6. Conclusion

OliGraph shows that assembling oligonucleotide pools before screening is computationally cheap and recovers sequences of concern that individual oligonucleotides are too short to trigger. Across ten blinded test sets varying in pool size and risk category, assembly prior to BLAST screening surfaced sequences that were invisible or buried in vector noise at the per-oligonucleotide level, while reducing the number of hits requiring manual review. The PCA-specific overlap model correctly distinguished fragments whose strand-alternating structure supports polymerase cycling assembly from those that match regulated sequences individually but lack the overlap geometry required for reassembly.

The tool gives synthesis providers a way to screen a product class that current pipelines do not evaluate, at a cost compatible with real-time order processing. It also provides regulators with evidence that the technical barriers previously cited against lowering minimum screening thresholds are tractable, supporting the assembly-aware assessment required under the EU Regulation on the responsible use of biotechnologies and calls to bring oligonucleotide orders within the scope of synthesis screening.

Code and Data

- **Code repository**: https://github.com/teojcryan/oligraph-rs
- **Demo link:** https://teojcryan.github.io/oligraph-rs/
- **Data/Datasets**: The sequencing datasets used in this study were curated from existing publicly available data, but due to informational hazards around sharing sequences of concern, these will not be available. Descriptions of dataset compositions are provided instead in Table S1.

References

Adam, L. *et al.* (2011) 'Strengths and limitations of the federal guidance on synthetic DNA', *Nature Biotechnology*, 29(3), pp. 208–210. Available at: https://doi.org/10.1038/nbt.1802.

Baum, C. *et al.* (2025) 'A system capable of verifiably and privately screening global DNA synthesis'. arXiv. Available at: https://doi.org/10.48550/arXiv.2403.14023.

Beal, J. *et al.* (2021) *Development and Transition of FAST-NA Screening Technology*. Available at: https://jakebeal.github.io/Unpublished/2021-BBN-FunGCATFinalReport.pdf.

Camacho, C. *et al.* (2009) 'BLAST+: architecture and applications', *BMC Bioinformatics*, 10(1), p. 421. Available at: https://doi.org/10.1186/1471-2105-10-421.

Diggans, J. and Leproust, E. (2019) 'Next Steps for Access to Safe, Secure DNA Synthesis', *Frontiers in Bioengineering and Biotechnology*, 7, p. 86. Available at: https://doi.org/10.3389/fbioe.2019.00086.

European Commission (2025) *Annexes 1 to 3 to the proposal for a regulation of the European Parliament and of the Council on establishing a framework of measures for strengthening Union's biotechnology and biomanufacturing sectors particularly in the area of health and amending Regulations (EC) No 178/2002, (EC) No 1394/2007, (EU) No 536/2014, (EU) 2019/6, (EU) 2024/795 and (EU) 2024/1938 (European Biotech Act)*. Strasbourg: European Commission.

Fady, P.-E. *et al.* (2025) *Annexes to "Cost-Benefit Analysis of Synthetic Nucleic Acid Screening for the UK"*. The Centre for Long-Term Resilience. Available at: https://doi.org/10.71172/q750-y70v.

Gretton, D. *et al.* (2025) 'Exact-match search with functional variant prediction enables automated DNA screening'. bioRxiv, p. 2024.03.20.585782. Available at: https://doi.org/10.1101/2024.03.20.585782.

Hoffmann, S.A. *et al.* (2023) 'Safety by design: Biosafety and biosecurity in the age of synthetic genomics', *iScience*, 26(3), p. 106165. Available at: https://doi.org/10.1016/j.isci.2023.106165.

Kane, A. and Parker, M.T. (2024) 'Screening State of Play: The Biosecurity Practices of Synthetic DNA Providers', *Applied Biosafety*, 29(2), pp. 85–95. Available at: https://doi.org/10.1089/apb.2023.0027.

Rizzi, R. *et al.* (2019) 'Overlap graphs and de Bruijn graphs: data structures for de novo genome assembly in the big data era', *Quantitative Biology*, 7(4), pp. 278–292. Available at: https://doi.org/10.1007/s40484-019-0181-x.

Rose, S. *et al.* (2024) 'Practical Questions for Securing Nucleic Acid Synthesis', *Applied Biosafety*, 29(3), pp. 159–171. Available at: https://doi.org/10.1089/apb.2023.0028.

Simirenko, L. *et al.* (2016) 'Bliss: The Black List Sequence Screening Pipeline'. *2016 Synthetic Biology: Engineering, Evolution & Design (SEED)*. Available at:

https://proceedings.aiche.org/sbe/conferences/synthetic-biology-engineering-evolution-design-seed/2016/proceeding/paper/bliss-black-list-sequence-screening-pipeline (Accessed: 26 April 2026).

Stemmer, W.P.C. *et al.* (1995) 'Single-step assembly of a gene and entire plasmid from large numbers of oligodeoxyribonucleotides', *Gene*, 164(1), pp. 49–53. Available at: https://doi.org/10.1016/0378-1119(95)00511-4.

Wheeler, N.E., Carter, S.R., *et al.* (2024) 'Developing a Common Global Baseline for Nucleic Acid Synthesis Screening', *Applied Biosafety*, 29(2), pp. 71–78. Available at: https://doi.org/10.1089/apb.2023.0034.

Wheeler, N.E., Bartling, C., *et al.* (2024) 'Progress and Prospects for a Nucleic Acid Screening Test Set', *Applied Biosafety*, 29(3), pp. 133–141. Available at: https://doi.org/10.1089/apb.2023.

Set	Oligos	Risk	Description
SET1	9,184	Negligible	Large non-overlapping innocuous oligonucleotide set
SET2	9,184	High	Oligonucleotides for constructing gene(s) from a UK ATCSA Schedule 5 virus. Inserted into a large non-overlapping background oligonucleotide set to mimic an attempt to conceal malicious intent
SET3	547	Negligible	Panel of oligonucleotides for synthesising an innocuous gene
SET4	231	High	Oligonucleotides for constructing a gene for a virulence factor from a UK ATCSA Schedule 5 bacterium. Spiked into a combination of overlapping innocuous sequences and randomly-generated oligonucleotides
SET5	70	Moderate	Oligonucleotide set from published literature relating to a non-Schedule 5 virus, subset for constructing gene(s)
SET6	95	High	Oligonucleotide set from published literature used for partially constructing a synthetic UK ATCSA Schedule 5 virus
SET7	144	High	Oligonucleotides for constructing genes for two toxins covered under UK ATCSA Schedule 5. Spiked into a combination of overlapping innocuous sequences and randomly-generated oligonucleotides
SET8	100	Negligible	Randomly-generated 80bp oligos
SET9	73	Low	Primers and other non-assembling sequences from published literature used for genotyping of UK ATCSA Schedule 5 pathogens
SET10	9,184	Moderate	Oligonucleotides for constructing gene(s) from a virus closely related to a UK ATCSA Schedule 5 virus. Inserted into a large non-overlapping background oligonucleotide set to mimic an attempt to conceal malicious intent

Table 1. Summary of oligonucleotide sets used in the blinded validation study. Ten sets spanning a range of pool sizes (70–9,184 oligonucleotides), sequence origins, and assembly contexts were each assigned a risk category prior to analysis.

			OliGraph		Per-oligonucleotide BLAST			Per-contig BLAST		
Set	**Oligos**	**Lengths (bp)**	**Contigs**	**Max contig (bp)**	**Hits (%)**	**Best align (bp)**	**Top hit**	**Hits (%)**	**Best align (bp)**	**Top hit**
SET1	9,184	152	0	–	2.30	108	cloning vector	–	–	–
SET2	9,184	74–152	1	1,393	2.57	109	cloning vector	100	1,400	SOC
SET3	547	22–165	120	1,597	85.6	150	cloning vector	88.3	1,593	innocuous
SET4	231	17–88	9	415	10.4	88	SOC	100	214	SOC
SET5	70	35–56	20	665	5.71	54	synthetic construct	20	622	synthetic construct
SET6	95	71–74	4	1,905	100	74	SOC	100	1,875	SOC
SET7	144	17–300	9	825	66.0	300	SOC (unassembleable)	100	803	SOC (assembleable)
SET8	100	80	0	–	0	–	–	–	–	–
SET9	73	20–39	1	36	0	–	–	100	36	SOC
SET10	9,184	64–152	3	457	2.46	108	cloning vector	100	457	SOC

Table 2. Summary of OliGraph assembly, and BLAST screening results. Each set was screened by nucleotide BLAST (E-value $\leq 10^{-5}$, max 10 targets) against the NCBI core nucleotide database at both the individual oligonucleotide level and the assembled contig level. "Queries with hits" indicates the number of query sequences returning at least one BLAST alignment. Best alignment length refers to the single highest-scoring hit across all queries. "SOC" denotes a sequence of concern (pathogen- or toxin-associated); specific identities are reported in Table S1. Dashes indicate no hits or no contigs produced.

Appendix

Dual-Use Considerations

This work describes a screening gap in which overlapping oligonucleotide fragments, individually too short to trigger existing detection thresholds, can be reassembled into regulated sequences through polymerase cycling assembly. The underlying route is already documented in the published literature and exploits commercially available reagents and services; OliGraph is designed to close it by enabling synthesis providers to evaluate a product class that current pipelines cannot screen.

Nonetheless, a formal description of the gap necessarily draws attention to a circumvention strategy whose barrier to entry is already low, and we have taken several steps to mitigate information hazard. We omit step-by-step assembly protocols, specific tiling parameters that would maximise evasion probability, and searchable identifiers for the sequences in our test sets, and we do not cite sources that would permit direct replication of the evasion strategy. A formal red-team assessment of whether the oligonucleotide-pool route bypasses existing commercial screening controls end-to-end was beyond the scope of the present study; such an evaluation, conducted under appropriate institutional biosafety and ethical oversight, is a priority for follow-on work. Provider-side or cross-provider countermeasures not captured by our analysis may already mitigate aspects of the threat described here. We recommend that researchers extending this work engage dual-use review processes before publishing assembly-level detail and coordinate disclosure with synthesis providers and regulatory bodies where findings reveal screening gaps.

LLM Usage Statement

Claude was used during software development to refine the initial overlap-graph construction code for performance and to generate the WebAssembly port of the screening engine. The model was not used to design experimental methodology, select test sequences, or interpret results. All code produced with LLM assistance was reviewed by the authors, and tool outputs were independently verified against the manually curated validation datasets described in Section 3.4.