IntuitionLabs

INTUITIONLABS / RESEARCH & PRACTICAL GUIDANCE

Claude ART Discovery Genome Mining: Reproducibility Test

> Build practical AI for pharma and biotech with IntuitionLabs. We help life-science teams turn complex information and workflows into useful software, governed knowledge systems and AI tools.

Published by IntuitionLabs · 2026-11-22 · claude art discovery · genome mining · ai genome mining · claude ai genome mining · array-associated reverse transcriptase · reproducibility in ai drug discovery · biosynthetic gene cluster mining · llm agents bioinformatics · reproducible ai pipelines · natural product discovery

Original article: https://intuitionlabs.ai/articles/claude-art-discovery-genome-mining-reproducibility



In this article

1. Executive Summary
2. Introduction and Background
3. Key Findings in the ART Report
4. How ART Fits Into Genome Mining and AI Discovery Practice
5. Implementation Considerations and Process Changes
6. Pilot Design and Worksheet
7. Data Analysis and Evidence
8. Implications and Future Directions
9. Frequently Asked Questions (FAQs)
10. Conclusion

Executive Summary

Anthropic reported on 23 September 2026 that Claude agents surfaced a previously undescribed system, **array-associated reverse transcriptase (ART)**, in bacteriophage sequence data. ART pairs a reverse transcriptase with a partner gene and a long array of evenly spaced DNA repeats, and the agents surveyed loci across 1.9 billion protein clusters (www.biorxiv.org). The technical preprint is vendor-authored, and as of 22 November 2026 the bioRxiv page "has not been certified by peer review" (www.biorxiv.org). ART is not shown to be an active RT, a gene-editing tool, a CRISPR equivalent or a drug-discovery success: the preprint states that "we have not shown that the RT is active or that the unit RNAs are its substrates" (www-cdn.anthropic.com).

The funnel is quantifiable. Agents searched **1,939,242,578 protein clusters**, scored **3,564 partner families** (www-cdn.anthropic.com), filed **19 reports** (www-cdn.anthropic.com), and confirmed 3 of 17 candidate families as new RT associations (www-cdn.anthropic.com). The campaign used 949 sessions, 215.6 million tokens and 21.5 hours of wall-clock time (www-cdn.anthropic.com). The blog's rounded figures (roughly 950 agents, 210 million tokens, 20 candidates) differ from these exact counts (www.techrepublic.com) (interestingengineering.com), so a pilot should log exact counts rather than inherit summaries.

The decisive result for pilot design is reproducibility. The vendor ran the same campaign **ten more times, and none recovered the array** (www-cdn.anthropic.com). The preprint does not report how many of the ten completed the census. In fixed-input tests, recognition is strong when the right DNA is shown, while autonomous search reliability is the weak point. This report's own calculation shows that zero recoveries in ten runs is compatible with a true per-run rate of up to about 31% (95% two-sided upper bound).

The recommended pilot fixes inputs and versions, logs worker and supervisor sessions, runs 14 to 29 independent full campaigns against a pre-registered positive control, uses blinded human triage, and applies biochemical stop/go gates before any claim of function. Candidate confirmation rate and reproducible discovery rate should be tracked separately. Access is also a constraint: Mythos 5.2 is offered to vetted organizations through trusted access programs (www.anthropic.com), with list prices of $8 and $40 per million input and output tokens (www.anthropic.com). Every cost-per-validated-candidate input should come from local data, and none is inferred from the vendor run.

Introduction and Background

This report examines a single, well-documented test of agentic genome mining: Anthropic's report that Claude agents surfaced a previously undescribed system, named **array-associated reverse transcriptase (ART)**, in bacteriophage (bacteria-infecting virus) sequence data. Anthropic describes ART as having "three parts: the RT, a partner gene beside it, and a long array of evenly spaced DNA repeat sequences" (www.anthropic.com). A reverse transcriptase (RT) is an enzyme that copies RNA into DNA. The report matters to biotech discovery leads, computational biology managers and research IT architects because it exposes both the funnel that turned billions of protein clusters into a handful of reports and a reproducibility result that vendor announcements usually omit.

The practical question is whether to pilot Claude agents for genome mining, and how to judge a candidate before allocating wet-lab capacity. The answer is conditional. The agentic search produced a testable lead, not a validated tool, and the preprint states that "we have not shown that the RT is active or that the unit RNAs are its substrates" (www-cdn.anthropic.com). ART is therefore not a gene-editing tool, a demonstrated CRISPR equivalent, a shown active RT or a drug-discovery success.

Provenance shapes how much weight the findings can bear. All authors list Anthropic as their only affiliation (www-cdn.anthropic.com), the preprint discloses that "Anthropic has filed a provisional patent application related to this work" (www.biorxiv.org), and as of 22 November 2026 the bioRxiv posting "has not been certified by peer review" (www.biorxiv.org). The only outside item located was a single-author computational follow-up that is also unreviewed (www.researchsquare.com).

The report reconstructs the funnel, the ten-rerun result and the evidence ladder, then turns them into an auditable architecture, triage gates and a pilot worksheet whose local inputs are left blank. Secondary questions are answered along the way: how Claude AI genome mining differs from conventional biosynthetic gene cluster (BGC) tools, what AI genome mining means for biotech and natural product discovery, where large language model (LLM) agents fit bioinformatics workflows, and which controls make AI pipelines reproducible.

Key Findings in the ART Report

What Anthropic Reported, and What Remains Unknown

Anthropic announced the result on 23 September 2026. Its post says agents "gathered over 200,000 RTs, picked out 3,500 new candidate systems" (www.anthropic.com), and its human involvement "was limited to the initial prompt and the lab work" (www.anthropic.com). The novelty claim is narrow: previous studies had already identified the RT in a jumbo phage (interestingengineering.com), and the new elements are the array and the accessory protein. The agents surveyed RT loci across 1.9 billion protein clusters (www.biorxiv.org).

The laboratory evidence is also narrow. The announcement reports that the array is expressed as distinct short RNAs (www.anthropic.com), and the preprint's 8% figure comes from re-analysis of a public dataset, in which array-derived RNAs account for up to 8% of phage RNAs at 15 minutes after infection (www-cdn.anthropic.com). Anthropic's chief executive wrote that the system's "precise function, biotechnological

utility (if any), or level of significance is not yet clear" (interestingengineering.com). Coverage agrees that the preprint "has not yet shown that the enzyme is active" (thenextweb.com), and that ART "has not been shown to function like CRISPR" (interestingengineering.com).

The family is defined computationally. Profile hidden Markov model searches identified 95 distinct RT clusters in jumbo phages and predicted viral contigs (www-cdn.anthropic.com), 28 of which carried a detectable repeat array upstream of the RT (www-cdn.anthropic.com). Arrays consist of 3 to 21 repeat copies (thenextweb.com).

The Data-to-Report Funnel

Table 1 below reconstructs the campaign from the preprint's methods, showing the count at each stage, who made the selection, and what is not reported.

Stage	Reported count	Who or what decided	Not reported or unclear
Protein search space	1,939,242,578 clusters at 50% identity (www-cdn.anthropic.com)	Fixed data build	Freeze identifiers for local re-use
RT census	3,391,714 raw candidate sequences, clustered to about 198,290 RT clusters (www-cdn.anthropic.com)	52 profile HMMs with thresholds fixed on reference RTs, then agent filtering	Downloadable HMM files; per-step sequence loss
Partner-family scoring	3,564 families scored (www-cdn.anthropic.com)	Agent-defined statistics	Full ranked list
Confirmed new RT associations	3 of 17 candidates (www-cdn.anthropic.com) (after an agent-defined selection step)	14 rejected or set aside as annotation artifacts, parts of previously described systems, or general neighborhood residents (www-cdn.anthropic.com)	Tabulated thresholds; human adjudication record
Reports	19 (www-cdn.anthropic.com)	Agents, then human reviewers	Human review procedure
Wet lab	Array expression (www.techrepublic.com)	Human scientists	RT activity, substrate, mechanism

The largest reductions (from thousands of partner families to 17, and from 17 to 3) are reported in the preprint without a tabulated adjudication record, so a pilot should log who or what made each of those decisions and on what thresholds. The stage humans demonstrably performed, bench work, tested only expression. Human-confirmation counts per stage are not reported, so that column of a local funnel should stay blank until measured. The harness used a worker that proposed and executed a plan "while a second agent acts as the supervisor", with a shell, file access and database connectors, which defines what a pilot can log and replay.

Reconciling Rounded Blog Figures With the Preprint

The blog and press use rounded or differently defined figures:

- **Agents or sessions:** "roughly 950 agents" (www.anthropic.com) corresponds to 949 sessions, not 949 distinct agents.
- **Tokens and time:** 210 million tokens and 21 hours in the blog, against 215.6 million tokens (cache reads excluded) and 21.5 hours in the preprint (www-cdn.anthropic.com).
- **RT sequences:** "over 200,000 RTs" rounds roughly 198,290 clusters, not raw hits.
- **Candidates:** the blog's 3,500 new candidate systems compares with 3,564 candidate partner families scored (www-cdn.anthropic.com), and the preprint does not call these systems.
- **Shortlist:** the blog and one outlet say 20 candidates (interestingengineering.com), while the preprint says 19 reports (www-cdn.anthropic.com). The 19 are 16 reports covering 17 families plus three on new RT lineages, and the preprint does not explain the blog's 20.

For a pilot, the lesson is to record exact counts from logs rather than inherit rounded summaries.

The Ten-Rerun Result and Fixed-Input Tests

Anthropic launched the campaign "ten more times with the same harness and brief" (www-cdn.anthropic.com), and the preprint reports that the array was missed in every rerun, with none reading the DNA upstream of the RTs (www-cdn.anthropic.com). The preprint adds details that summaries lose:

- **Census completion:** "nearly every campaign that completed the census sampled ART loci" (www-cdn.anthropic.com), but the exact number of reruns that completed it is not given.
- **Lineage follow-up:** in two reruns workers investigated the lineage, but none read the DNA upstream of the RTs.
- **Stated cause:** the authors "attribute these results to the broad protein search space" and the non-deterministic harness (www-cdn.anthropic.com).
- **Limit of the check:** an ART locus on a contig outside the searched identifier set would not be found by the vendor's search.

In the original run, the worker read the upstream DNA of relatives and noticed the tandem pattern directly.

This suggests separating two capabilities. **Model recognition** is seeing a repeat array when the sequence is in context. **Autonomous search reliability** is the probability that a campaign puts the right sequence into context at all. In the vendor's fixed-input benchmark, the most capable models described the array in "at least 90% of attempts, compared to as low as 32%" under other conditions (www-cdn.anthropic.com). These are vendor-run measurements judged by a Claude model, so they are self-assessments, not independent benchmarks. Recognition is strong when the right DNA is shown; the rerun miss concerns which loci were examined. A pilot that only repeats a fixed-input test measures the easier property.

An Evidence Ladder for ART

Table 2 below places each rung of evidence against what is actually shown.

Rung	What it means	What is shown for ART	Status
1. Computational anomaly	Pattern unlike known elements	28 of 95 RT clusters carry an upstream array	Shown, vendor-reported

Rung	What it means	What is shown for ART	Status
2. RNA expression	The element is transcribed	Up to 8% of phage RNAs in public data; short RNAs reported by press (www.techrepublic.com)	Shown in reused data plus E. coli work
3. Biochemical activity	Enzyme acts in a defined assay	"we have not shown that the RT is active" (www-cdn.anthropic.com)	Not shown
4. Mechanism	Substrates, partner interaction	Further testing is required to establish what ARTs do (tribune.com.pk)	Not shown
5. Translational utility	Use in a product or therapy	"biotechnological utility (if any)" unclear (interestingengineering.com)	Not shown

Evidence is solid at rung 1, partial at rung 2 and absent at rungs 3 to 5. A computational preprint posted 15 October 2026 reports that the ART-associated RTs "show their strongest similarity to retron-associated RTs" (www.researchsquare.com). That is a classification hypothesis, not a replication, and it shows why rungs 3 and 4 must come from the bench.

How ART Fits Into Genome Mining and AI Discovery Practice

Conventional Genome Mining for Natural Products

Genome mining is the computational search of sequenced DNA for genes that encode useful molecules or enzymes. In natural product discovery the classic unit is the biosynthetic gene cluster (BGC). The standard detector, antiSMASH, increased detectable pathway types from 81 to 101 in version 8.0 (pmc.ncbi.nlm.nih.gov). The curated reference set is small relative to the predicted space: MIBiG reached 3,059 curated entries (pmc.ncbi.nlm.nih.gov), while BiG-FAM captured 1,225,071 predicted BGCs in 29,955 families (pmc.ncbi.nlm.nih.gov). A preprint estimated that only about 3% of bacterial biosynthetic diversity has been experimentally accessed (www.biorxiv.org). Scale is an operational constraint: clustering 1.2 million BGCs with BiG-SLiCE took about 10 days on a 36-core server (pmc.ncbi.nlm.nih.gov). For agents, the open problem is deciding what to look at.

AI and LLM Approaches to BGC and Enzyme Mining

Machine learning entered BGC detection before agents. DeepBGC offered "reduced false positive rates" (pmc.ncbi.nlm.nih.gov), and a 2026 attention-based model reports outperforming antiSMASH and DeepBGC on natural product class prediction (pmc.ncbi.nlm.nih.gov). These are models inside fixed pipelines.

BGC mining with LLMs has so far appeared mainly as orchestration; the Agent_K preprint, for example, uses "a large-language-model (LLM) agent to coordinate multiple bioinformatics tools" (pmc.ncbi.nlm.nih.gov). The ART campaign went further: agents chose HMMs and thresholds, sampled neighborhoods and wrote reports on top of standard software. HMMER3 is "100- to 1000-fold faster than HMMER2" (pmc.ncbi.nlm.nih.gov) and MMseqs2 clusters "huge protein and nucleotide sequence sets" (github.com), which explains why a CPU sandbox sufficed. For structure-level comparison, Foldseek decreases computation times by four to five orders of magnitude relative to earlier tools (pmc.ncbi.nlm.nih.gov).

For **Claude for life sciences research**, Anthropic launched Claude Science in beta for Pro, Max, Team and Enterprise users (www.anthropic.com). A vendor benchmark from the earlier Claude for Life Sciences launch reported 0.83 on Protocol QA against a human baseline of 0.79 (www.anthropic.com), but it is a multiple-choice test with 10-shot prompting (www.anthropic.com), so it measures protocol knowledge, not genome mining reliability.

Precedents That Reached the Bench

Array-association mining is not new. FLSHclust found about 130,000 clusters associated with CRISPR-like repeats (pmc.ncbi.nlm.nih.gov), identified 188 previously unreported CRISPR-linked modules (pmc.ncbi.nlm.nih.gov), and then experimentally characterized three HNH nuclease-containing CRISPR systems (pmc.ncbi.nlm.nih.gov). The Fanzor study concluded that "Fz is a eukaryotic OMEGA system" after biochemical, structural and human-cell work (pmc.ncbi.nlm.nih.gov). The common feature is wet-lab characterization before anything was called a tool, and ART is before that step.

A census will also contain better-understood RT elements. Diversity-generating retroelements "use reverse transcription to generate vast numbers of sequence variants in specific target genes" (pmc.ncbi.nlm.nih.gov). Retron defense units have three components: the RT, a non-coding RNA and an effector (www.pubmed.gov). In several type III CRISPR systems, Cas1 is naturally fused to an RT (pmc.ncbi.nlm.nih.gov). CRISPR arrays store spacers inserted between direct repeats. Structural resemblance is a reason to look, not evidence of CRISPR-like function.

Data Scale and Public Sequence Resources

Input data matter as much as the agent. UniRef100 had 453,950,711 clusters and UniRef50 had 69,290,910 in release 2025_01 (academic.oup.com), the MGnify protein database exceeded 2.4 billion non-redundant sequences (pmc.ncbi.nlm.nih.gov), a 2023 analysis of 26,931 metagenomes identified 1.17 billion protein sequences (pmc.ncbi.nlm.nih.gov), and Logan assembled 44.1 petabases of raw data into 0.9 petabases of contigs (pmc.ncbi.nlm.nih.gov). The ART campaign searched 1.9 billion clusters (www.biorxiv.org), the same order as the largest public resources. For a local pilot, a smaller, frozen, accession-listed subset is the realistic input, fixed before the first run.

Implementation Considerations and Process Changes

An Auditable Architecture for Agentic Genome Mining

The ten-rerun result implies that a pilot cannot rely on a single successful transcript. Each campaign should be an experiment with recorded inputs and outputs. These controls translate **reproducible AI pipelines** practice for biotech into agent-specific requirements.

- **Immutable source sequences:** freeze the build, list accessions and store checksums, and cite public accessions for every input.
- **Versioned retrieval and models:** "Archive the Exact Versions of All External Programs Used" (pmc.ncbi.nlm.nih.gov), including HMMER, MMseqs2, profile HMM files and thresholds.

- **Result lineage:** "For Every Result, Keep Track of How It Was Produced" (pmc.ncbi.nlm.nih.gov). For agents, store the plan, tool calls and outputs, not only the final report.
- **Seeds and sampling:** record random seeds, model version, system prompt and brief.
- **Worker and supervisor logs:** keep both sides of the loop, because revision history after supervisor review is itself an audit artifact.
- **Independent reruns:** repeat the full campaign, as the vendor did ten times, and score against a pre-registered hit list.
- **Blinded candidate review:** strip agent confidence language and rank so the ranking itself can be tested.

Existing tooling carries much of this. Nextflow's authors identified software and database practice as the "main source of computational irreproducibility" in large omics analyses (www.nature.com). W3C PROV defines provenance as "information about entities, activities, and people involved in producing a piece of data or thing" (www.w3.org), RO-Crate offers "a lightweight approach to packaging" research data (www.researchobject.org), and the nf-prov plugin generates provenance for Nextflow runs (www.researchobject.org). The vendor also describes a reviewer agent that "checks citations and calculations" (www.anthropic.com), and Claude Science's product page says outputs include "the exact code, environment, and conversation that produced them" (claude.com), but a pilot should still export logs to a system the team controls.

Why Run-to-Run Variance Is Expected

The reruns are consistent with the nondeterminism literature. A study of five LLMs on eight tasks found accuracy varying "up to 15% across naturally occurring runs" in settings meant to be deterministic (arxiv.org). Sampling 1,000 completions at temperature 0 produced 80 unique outputs (thinkingmachines.ai), which one analysis attributed to batch-size variation under server load (thinkingmachines.ai), and none of the tested models was repeatable across all tasks (arxiv.org). Agent trajectories "diverge early, often within the first few percent of tokens" (arxiv.org), with run-to-run standard deviations above 1.5 percentage points even at temperature 0 (arxiv.org), and a September 2026 preprint on 584 coding-agent runs found that identical runs of one pairing varied more than the pairings did (arxiv.org). If variance is measurable in short tasks, a multi-hour campaign that depends on an agent choosing to read one stretch of DNA should vary more. Reporting the rerun rate, as the vendor did, is the correct practice.

Candidate Triage and Wet-Lab Validation Gates

Gates should be pre-registered, because pre-registration "clarifies the distinction between planned and unplanned analyses" (pmc.ncbi.nlm.nih.gov), which matters when an agent proposes and tests hypotheses on the same data.

- **Gate 0, data and logs complete:** stop if the run cannot be replayed from logs.
- **Gate 1, second computational method:** confirm the anomaly with an independent tool. An established repeat finder is the natural second method.
- **Gate 2, human novelty check:** the agents' own literature search is a filter, and a human should confirm novelty.

- **Gate 3, expression evidence:** RNA-seq or small-RNA sequencing on a defined construct, with replicates stated in advance.
- **Gate 4, biochemical activity:** an assay with positive and negative controls, such as a catalytic-site mutant. This is the rung ART has not passed (see Table 2).
- **Gate 5, biosafety and IP review:** a human decision. The preprint discloses a provisional patent application (www.biorxiv.org). Consult biosafety officers and patent counsel; this report offers no legal opinion.

The blinded step matters because the preprint ranked its 19 reports in 342 pairwise games with a model judge (www-cdn.anthropic.com). Model-ranked reports are a convenient filter, but humans should score a blinded sample, including low-ranked reports, to test whether rank predicts bench outcomes.

Access, Safeguards and Pricing

Access to the campaign model is restricted. Mythos 5 was "currently only available to a small group of vetted partners" (www.anthropic.com), and Mythos 5.2 "is available to vetted organizations that work in these dual-use areas" (www.anthropic.com). The Life Sciences Verification Program offers "a refined set of safeguards more permissive for biology-related work" (anthropic.com), and Mythos models require 30-day data retention by default (www.anthropic.com), which research IT should weigh against local governance rules. Reproducing the campaign is thus a question of program eligibility, not only budget, and standard grants cover Mythos 5.1, Opus 5 and Sonnet 5 (anthropic.com), while classifiers still cover "most tasks in research biology" under the public Fable 5.2 model (www.anthropic.com).

As of 22 November 2026, list prices are $2 per million input tokens and $10 per million output tokens for Sonnet 5.5 (platform.claude.com), $4 and $20 for Opus 5.5 (platform.claude.com), and $8 and $40 for Mythos 5.2 (www.anthropic.com), with a 50 percent Batch API discount (platform.claude.com). These are inputs to the local cost worksheet, not an estimate of the vendor run.

Pilot Design and Worksheet

Design Principles

A defensible pilot answers three questions in order: does the pipeline find a pre-specified positive, how often does it do so across independent reruns, and what fraction of its candidates survive the bench. The design below keeps every threshold local and every unreported input blank.

1. **Plant a positive control.** Choose a known but under-annotated element in the frozen dataset, record it in a pre-registered hit list, and do not mention it in the brief. The original ART run was not told about arrays, and a pilot should reproduce that blindness deliberately.
2. **Fix everything that can be fixed.** Frozen sequences, HMM versions, thresholds, model version, brief and harness configuration.
3. **Run enough independent full campaigns.** The sample-size arithmetic below is this report's own calculation, not a published figure.
4. **Score with humans first.** Model-judged rankings may be logged, but go/no-go decisions rest on human review and bench data.

5. **Separate recognition tests from search tests.** Run a fixed-input test (cheap, high-signal) and a full-campaign test (expensive, the one that matters).

On sample size: if all N independent full reruns miss the pre-registered hit, the exact one-sided 95% upper bound on the per-run success probability is 1 minus 0.05 to the power 1/N. For N = 10, as in the vendor's rerun set, that is about 26%; the two-sided exact interval's upper limit is about 31%. Put differently, ten misses are compatible with a true per-run recovery rate as high as roughly one in four. To bound the rate below 10% with the same confidence after zero recoveries, a team needs at least 29 reruns; to bound it below 20%, at least 14. These figures assume the reruns are independent and the hit definition is fixed in advance, which the vendor's rerun set cannot fully guarantee because the original success was identified after the fact.

Pilot Worksheet

Table 3 below is the worksheet. It gives each metric's formula and leaves every local input blank. The right-hand column states what the ART report publishes, so a team can see which inputs it must measure itself.

Metric	Formula	Local input (leave blank until measured)	ART report status
Candidate confirmation rate	Independently validated candidates divided by lab-tested candidates	Validated: ____; tested: ____	Not applicable: no candidate has been validated beyond expression
Reproducible discovery rate	Independent full-run repeats recovering the pre-registered hit divided by full reruns	Recovering: ____; reruns: ____	0 of 10 reruns recovered the array (www-cdn.anthropic.com); completed-census count not reported
Recognition rate (fixed input)	Attempts describing the target feature divided by attempts	Described: ____; attempts: ____	High for the most capable models under favorable conditions (vendor-run)
Cost per validated candidate	(Local inference + compute + expert review + wet-lab cost) divided by validated candidates	Inference: ____; compute: ____; review hours: ____; wet-lab: ____; validated: ____	Token and session counts are published; costs are not
Funnel retention by stage	Candidates at stage k divided by candidates at stage k minus 1	Counts per stage: ____	Stage counts published in part (see Table 1)
Human-reviewer agreement	Share of blinded reports with concordant reviewer decisions	Concordant: ____; reviewed: ____	Not reported

The worksheet emphasizes that the two headline rates are different quantities. The confirmation rate is a property of the candidates that reach the bench, and it can be high even when the discovery process is unreliable. The reproducible discovery rate is a property of the process, and it can be low even when every individual finding is real. The vendor report supplies a numerator and denominator for the second (0 of 10 reruns) but not for the first. A go decision should require adequate values on both.

Suggested Stop/Go Structure

- **Go to wet lab** only if the second method reproduces the computational anomaly, a human confirms novelty, and the biosafety and IP reviews are complete.
- **Go to scale-up** only if the reproducible discovery rate on the planted positive has an upper confidence bound the team pre-agreed is acceptable.
- **Stop** if provenance cannot be reconstructed, if reviewers cannot distinguish agent-ranked candidates from controls, or if the confirmation rate at the bench matches the rate for random candidates drawn from the same census.
- **Do not extrapolate** the vendor's 21.5-hour, 949-session, 215.6-million-token run into a productivity benchmark: the preprint describes one campaign, not a throughput measurement, and no published protocol compares it with expert time.

Data Analysis and Evidence

This section consolidates the quantitative material and separates vendor-run measurements from independent ones. All ART figures are vendor-reported, and no independent replication of the campaign has been located as of 22 November 2026; the only outside analysis found is an unreviewed computational preprint (www.researchsquare.com).

Quantities Reported for the ART Campaign

Table 1 holds the stage counts, and the effort and family-definition counts appear in the sections above. The rerun set produced zero recoveries of the array in ten attempts, as described above.

From 3,564 partner families to 3 confirmed associations is a retention of about 0.08%, and from 17 candidates to 3 is about 18% (computed by this report from the preprint's counts in Table 1). A team should budget for low yield at every stage and recognize that bench-level yield of the whole pipeline is unknown until assays run.

Reproducibility Statistics

With zero recoveries in 10 independent runs, the exact one-sided 95% upper bound on the per-run recovery probability is about 26%, and the two-sided upper limit about 31% (own calculation). The vendor checked the reruns by searching thousands of session transcripts for identifiers of the original contigs. The fixed-input benchmark is a separate quantity: 100 attempts per model at each of five input levels, judged by a Claude model, with recognition highest under the most informative condition.

Context From Other Agent Evidence

Broader agent evidence gives a calibration range. BixBench has nearly 300 open-answer questions (arxiv.org), and its abstract reports that frontier models "only achieve 17% accuracy in the open-answer regime" (arxiv.org) (a different paper version gives a slightly different figure, so the abstract wording is used). The Biomni paper averaged results "across three independent runs" (pmc.ncbi.nlm.nih.gov) but evaluated only a subset of the benchmark. FutureHouse's Robin describes its approach as "semi-autonomous" (arxiv.org). The Virtual Lab study designed 92 nanobodies computationally, and two showed improved binding to recent variants (www.nature.com), a yield of about 2% for the highlighted property (computed from the reported counts).

These tasks differ from genome mining, but they share a pattern: agents produce candidates at scale and the bench confirms a minority. A June 2026 review argues that AI drug discovery platforms "remain constrained by biological complexity, limited reproducibility, and insufficient clinical validation" (www.mdpi.com).

What the Data Do Not Support

- **No productivity benchmark:** one 21.5-hour campaign with a model-judged shortlist does not measure expert time saved.
- **No activity or editing claim:** RT activity was not shown, and ART has not been shown to function like CRISPR (interestingengineering.com).
- **No cost-per-discovery estimate:** token counts are published, but compute, review and bench costs are not.

Implications and Future Directions

For discovery leads. The report shows that an agentic pipeline can generate a testable anomaly from public data at modest compute, and that doing so was not repeatable across ten reruns. The reasonable position is to pilot agents as lead generators feeding a human-gated validation process. Wire coverage reports that Anthropic stressed further testing is required to establish what ARTs do (tribune.com.pk).

For computational biology managers. The decisive variable is search policy, because recognition is strong when the right DNA is read. Teams can constrain agents to specified loci, run several campaigns with different briefs, and log which sequences ever entered context.

For research IT architects. Treat the harness as a validated computational system. FDA's January 2025 draft guidance proposes a "risk-based credibility assessment framework" tied to a "particular context of use" (www.fda.gov), and it states that AI used for drug discovery is not in its scope (www.fda.gov), so it serves as vocabulary for credibility and risk, not as a genome-mining requirement. Its idea that model risk combines influence and consequence (www.fda.gov) helps rank which agent outputs need strict review. **Reproducible AI pipelines in biotech** need frozen inputs and logs so a second team can re-run a campaign against a pre-registered hit list.

Open questions as of November 2026. Peer review has not been located; the bioRxiv page still carries the notice that the work has not been certified by peer review (www.biorxiv.org). Wet-lab follow-up on RT activity, release of agent transcripts and release of the ART sequence set are not documented in the sources reviewed, and no data-availability or code-availability statement for newly generated materials was found in the preprint text. Model access is also moving: Mythos 5.2 is offered to vetted organizations working in dual-use areas (www.anthropic.com), so a replication on a current model would not be a strict re-run of a campaign that used Mythos 5 (thenextweb.com). Anthropic's vendor description that every output carries "an auditable history of how it was made" (www.anthropic.com) is a feature claim to test, not a substitute for local logs.

Evidence that would change this assessment includes independent biochemical characterization, a pre-registered rerun series by a third party, and publication of the briefs and logs. Until then, the evidence ladder in *Table 2* should govern how ART is discussed internally.

Frequently Asked Questions (FAQs)

Can Claude do AI genome mining in biotech today?

Claude agents completed a campaign over 1,939,242,578 protein clusters and produced 19 reports (thenextweb.com), but none of ten reruns recovered the key array (thenextweb.com). The supported claim is lead generation under human supervision.

Is ART a gene-editing tool or a CRISPR equivalent?

No evidence supports that. ART has not been shown to function like CRISPR (interestingengineering.com), and the preprint has not yet shown that the enzyme is active (thenextweb.com).

Has the ART result been peer reviewed or independently replicated?

Not according to the sources reviewed. The bioRxiv version has not been certified by peer review (www.biorxiv.org), and an independent computational follow-up is itself unreviewed (www.researchsquare.com).

What does reproducibility in AI drug discovery require?

Frozen inputs, pinned versions, retained logs and repeated independent runs scored against pre-registered outcomes. Reviews cite limited reproducibility as a constraint (www.mdpi.com), and LLM output varies across runs even in settings meant to be deterministic (arxiv.org).

How do LLM agents differ from antiSMASH for natural product discovery?

Conventional tools run fixed algorithms, such as antiSMASH 8.0 with 101 detectable pathway types (pmc.ncbi.nlm.nih.gov). Agents choose tools, thresholds and neighborhoods during a run, so logging and reruns matter. Agent frameworks for BGC work are early; one is a preprint that coordinates existing tools (pmc.ncbi.nlm.nih.gov).

What does a pilot cost?

List prices are $2 and $10 per million input and output tokens for Sonnet 5.5 (platform.claude.com) and $4 and $20 for Opus 5.5 (platform.claude.com). Cost per validated candidate depends on local compute, review and wet-lab costs, which the worksheet leaves blank.

Conclusion

The ART report is a useful test case because it publishes both a result and its limits. Claude agents working from public sequence data surfaced a repeat array beside a reverse transcriptase that earlier literature had not described, and the array is reported to be expressed as short RNAs. The same preprint reports that ten full reruns did not recover the array, that RT activity has not been shown, and that no peer review had been completed at the time of writing.

For a team deciding whether to pilot Claude agents for genome mining, the report supports a disciplined, measured approach. Fix inputs and versions, log every session, run full campaigns repeatedly against a pre-registered positive control, review candidates blind, and move to the bench only after a second computational method and a human novelty check agree. Track the confirmation rate and the reproducible discovery rate separately, set stop and go thresholds from local data before the first run, and keep biosafety and intellectual property review with qualified people. Under that structure, an agentic pipeline can add leads to a discovery program without turning a pattern into a claim the data do not support.

Source links

Links cited in this article, in order of first appearance. Full addresses are provided for readers using extracted text.

1. https://www.biorxiv.org/content/10.64898/2026.10.05.755809v1
2. https://www-cdn.anthropic.com/22573675ada52a8ca8a97a1a4b4326b2f208a071.pdf
3. https://www.techrepublic.com/article/news-anthropic-claude-ai-agents-enzyme-discovery
4. https://interestingengineering.com/ai-robotics/claude-discovers-crispr-like-enzyme-system
5. https://www.anthropic.com/claude/mythos
6. https://www.anthropic.com/news/claude-discovers-novel-enzyme-system
7. https://www.researchsquare.com/article/rs-11367685/v1
8. https://thenextweb.com/news/anthropic-claude-enzyme-system-crispr-like-repeats
9. https://tribune.com.pk/story/2631536/anthropic-says-claude-discovered-novel-enzyme-system
10. https://pmc.ncbi.nlm.nih.gov/articles/PMC12230676
11. https://pmc.ncbi.nlm.nih.gov/articles/PMC11701617
12. https://pmc.ncbi.nlm.nih.gov/articles/PMC7778980
13. https://www.biorxiv.org/content/10.1101/2021.08.11.455920.full.pdf
14. https://pmc.ncbi.nlm.nih.gov/articles/PMC7804863
15. https://pmc.ncbi.nlm.nih.gov/articles/PMC6765103
16. https://pmc.ncbi.nlm.nih.gov/articles/PMC12744964
17. https://pmc.ncbi.nlm.nih.gov/articles/PMC13738595
18. https://pmc.ncbi.nlm.nih.gov/articles/PMC3197634
19. https://github.com/soedinglab/MMseqs2
20. https://pmc.ncbi.nlm.nih.gov/articles/PMC10869269
21. https://www.anthropic.com/news/claude-science-ai-workbench
22. https://www.anthropic.com/news/claude-for-life-sciences
23. https://pmc.ncbi.nlm.nih.gov/articles/PMC10910872
24. https://pmc.ncbi.nlm.nih.gov/articles/PMC10432273
25. https://pmc.ncbi.nlm.nih.gov/articles/PMC5758913
26. https://www.pubmed.gov/33157039
27. https://pmc.ncbi.nlm.nih.gov/articles/PMC4898656

28. https://academic.oup.com/nar/article/53/W1/W547/8126256
29. https://pmc.ncbi.nlm.nih.gov/articles/PMC9825492
30. https://pmc.ncbi.nlm.nih.gov/articles/PMC10584684
31. https://pmc.ncbi.nlm.nih.gov/articles/PMC12424806
32. https://pmc.ncbi.nlm.nih.gov/articles/PMC3812051
33. https://www.nature.com/articles/nbt.3820
34. https://www.w3.org/TR/prov-overview
35. https://www.researchobject.org/ro-crate
36. https://claude.com/product/claude-science
37. https://arxiv.org/abs/2408.04667
38. https://thinkingmachines.ai/blog/defeating-nondeterminism-in-llm-inference
39. https://arxiv.org/abs/2602.07150
40. https://arxiv.org/pdf/2609.33812
41. https://pmc.ncbi.nlm.nih.gov/articles/PMC7725296
42. https://www.anthropic.com/claude-fable-and-mythos-5-2
43. https://anthropic.com/news/life-sciences-verification-program
44. https://platform.claude.com/docs/en/about-claude/pricing
45. https://arxiv.org/abs/2503.00096
46. https://pmc.ncbi.nlm.nih.gov/articles/PMC12157518
47. https://arxiv.org/abs/2505.13400
48. https://www.nature.com/articles/s41586-025-09442-9
49. https://www.mdpi.com/1424-8247/19/6/916
50. https://www.fda.gov/regulatory-information/search-fda-guidance-documents/considerations-use-artificial-intelligence-support-regulatory-decision-making-drug-and-biological
51. https://www.fda.gov/media/184830/download

THE PUBLISHER

About IntuitionLabs

IntuitionLabs is an AI consulting, custom software development and data engineering firm serving pharmaceutical, biotechnology, medical-device and other life-science organizations. We work with clinical, regulatory, medical-affairs, commercial, quality and IT teams to connect technology decisions with the work people need to accomplish.

AI consulting and adoption

Our AI enablement services cover readiness assessments, use-case selection, governance and policies, team workshops, adoption measurement and ongoing advisory support. We help organizations structure the information layer behind AI: source material, context, permissions and maintained knowledge that make generated answers useful and reviewable. Private LLM inference and hosted AI options support teams evaluating how to operate AI with appropriate control over their data and infrastructure.

Software, data and life-science workflows

IntuitionLabs develops custom software for pharma and biotech, integrates enterprise systems, and builds data engineering and business intelligence solutions. Areas of focus include AI agents, regulatory research, medical writing, medical affairs, CMC information, competitive intelligence and clinical-document workflows. Our eTMF intelligence work includes cross-system reconciliation and inspection-readiness support.

Enterprise platforms and regulated delivery

We provide Veeva services, application support, managed services, integrations and custom applications, alongside enterprise content work involving platforms such as Egnyte. For regulated workflows, our services include GxP enablement, computer-system validation and software development addressing 21 CFR Part 11 requirements. The applicable controls, validation responsibilities and acceptance criteria are defined for each engagement.

Work with IntuitionLabs

Explore AI enablement, pharma and biotech software development, data engineering and BI, and Veeva services. Contact IntuitionLabs to discuss your workflow, information sources and implementation needs.

IntuitionLabs publishes educational research to help life-science teams make informed technology decisions. Coverage of a product or organization does not imply a client relationship, endorsement or partnership.

Website: https://intuitionlabs.ai

This article is educational. Verify consequential medical, legal, financial, regulatory and technical decisions with appropriate professionals and the cited primary sources. Company services are described separately from the article's research findings.