

Quantifying the Reuse Barrier in Published Omics Data: Dispersal, Cost, and Feasibility Models for Agent-Supported Retrieval, Reanalysis, and Synthesis

Quantifying the Reuse Barrier in Published Omics Data: Dispersal, Cost, and Feasibility Models for Agent-Supported Retrieval, Reanalysis, and Synthesis

Abstract

Most published omics data remain functionally inaccessible for computational reuse: the information needed to reproduce a reported result is dispersed across article main text, supplementary files, and code repositories, and intermediate data such as protein abundance tables are shared irregularly if at all [1]. Existing descriptions of this problem are qualitative; no published model states what the dispersal costs a reuser or what an agent changes. This paper supplies that model. We formalize a published study as a set of reproduction-critical artifacts dispersed over document locations, derive closed-form expressions for reproduction success probability and expected retrieval cost, derive a token-and-call cost model for the retrieval stage of an agent pipeline, and derive an information-theoretic bound on the metadata an agent must extract. Under stated assumptions (five dispersed artifacts, per-location retrieval success 0.8), a manual reproduction succeeds with probability 0.32768 and needs 6.25 expected retrieval visits, while an agent with two retries succeeds with probability 0.9606 at 6.20 visits; the retrieval stage costs 60400 tokens and 30 calls per paper before screening. The model turns "most data are inaccessible" into three measurable quantities — dispersal breadth, per-location reliability, and missing-artifact fraction — giving repository operators and agent developers concrete targets for measurement and deposition policy.

1. Introduction

A computational biologist who wants to reanalyze a published proteomics or genomics study faces a concrete, recurring obstacle. The raw data may sit in a public repository, but the parameters that produced the reported results, the filtering thresholds, the pipeline version, and the mapping from supplementary figures to underlying tables are scattered across the article main text, supplementary files, and, where they exist at all, code repositories; intermediate artifacts such as protein abundance tables are shared only in a minority of cases and at irregular locations [1]. The practical consequence is that a reanalysis that should take hours instead takes days or fails outright, and the aggregate effect is that most published omics data are functionally inaccessible for computational reuse even when they are technically downloadable [1]. This blocks specific decisions: whether to build on a published dataset rather than generate new data, whether a reported result survives an alternative processing choice, and whether a meta-analysis across studies can be assembled at all. Today the cost of reconstruction is paid in human hours, one paper at a time.

If the dispersal barrier could be quantified per study, three things change. A journal or repository could report a single reuse-readiness figure alongside a deposit. A reanalysis agent could allocate its retrieval budget to the locations that dominate the failure probability. And an author preparing a supplement could see which missing artifact contributes most to the expected cost imposed on every future reuser. Conversely, if the information an agent must extract to specify a reanalysis exceeds what the published record reliably contains, then no amount of agent capability closes the gap, and the binding constraint is publication practice itself. Both outcomes are actionable, and this paper derives conditions for each.

This work treats the pipeline proposed in [1] as the object of analysis. It does not propose a new agent architecture; it derives the cost structure, the information requirements, and the failure modes of any agent that performs the three functions — retrieval, reanalysis, synthesis — that [1] names. The analysis draws on the multi-agent systems literature for the coordination structure [2], [8], on the distributed-systems literature for the reliability question that arises when a pipeline is assembled from independently fallible retrieval steps [4], [5], and on the information-theoretic analysis of biological specification [11] for the bound on what metadata can capture.

The contributions are:

- A staged formalization of the omics discovery pipeline as a five-stage multi-agent workflow, with the coordination structure of each stage stated explicitly (Section 3).
- A dispersal model of a published study with closed-form expressions for reproduction success probability and expected retrieval cost, evaluated numerically under stated assumptions: manual success 0.32768 versus agent success 0.9606 at $k = 5$, $p = 0.8$, $r = 2$ (Sections 3 and 4.4–4.5).
- A derived token-and-call cost model for the retrieval stage, giving 60400 tokens and 30 calls per paper before screening, and 169400 tokens with a 10:1 screening ratio (Section 4.1).
- A derived information-theoretic bound on the reanalysis specification, giving Hartley entropy 83.05 bits at 25 fields and a Shannon-entropy comparison (Section 4.2).
- A sensitivity decomposition showing that dispersal breadth k dominates the barrier and that agent retries substitute for per-location reliability (Sections 4.4–4.5).
- A statement of which observations would falsify the central claims, a limitations list naming every assumption, and a practical pathway from the analytical result to a first measurement (Sections 6 and 7).

2. Background and Related Work

The direct object of this paper is the omics data discovery agents framework, which presents an agentic approach to the retrieval, reanalysis, and synthesis of published omics data and documents the core observation motivating everything here: most published data remain functionally inaccessible for computational reuse because reproduction-critical information is dispersed across main text, supplementary files, and code repositories, with intermediate data such as protein abundance files shared irregularly when at all [1]. That work establishes the problem and the agent architecture; its supplied summary provides no quantitative cost model of the dispersal barrier, which is the gap this paper addresses.

A second thread concerns how multiple cooperating agents coordinate. The survey of multi-agent deep reinforcement learning with communication establishes that communication is an effective mechanism for coordinating the behaviors of multiple agents, broadening their views of the environment, and supporting their collaborations, and that agents can communicate various types of messages either to all agents or to subsets of other agents [2]. In the omics retrieval setting, the analogue is a division of labor in which one agent reads the article text while another queries repositories and a third executes code; the survey's framing supports treating such message-passing as the coordination primitive, though it says nothing specific about scientific documents.

Retrieval over heterogeneous, multimodal document collections is treated in LLandMark, a modular multi-agent framework for landmark-aware multimodal interactive video retrieval, in which specialized agents collaborate across four stages to handle real-world complex queries in a domain where the diversity and scale of data demand multimodal understanding, adaptive reasoning, and domain-specific knowledge integration [8]. The structural lesson for omics retrieval is that a modular, staged agent design can handle diverse and large-scale document collections; the video domain differs in modality, but the problem shape is the same shape as the omics literature problem, and the four-stage decomposition serves as a comparative template for the five-stage formalization in Section 3.

Agentic discovery over data systems is examined in the declarative data services work, which observes that translating LLM-driven agentic discovery to multi-system data backends surfaces a harder problem: the search space is heterogeneous, the verifier is whether a deployed stack actually runs, and composition knowledge is unevenly captured in pretraining [6]. This maps directly onto omics reanalysis, where the verifier is whether a retrieved pipeline actually reproduces the published result and where pipeline-composition knowledge is exactly the kind of knowledge unevenly captured in pretraining. The entry's summary is truncated before its proposed solution, so we use it only for this problem framing and carry the verifier-runs question into Sections 4 and 6 as an open design constraint.

Reliability of distributed computation under partial failure is the subject of two works on Byzantine resilience. Data encoding for Byzantine-resilient distributed optimization studies distributed optimization where data and computation are distributed among m worker machines, t of which may be corrupt, with compromised nodes able to collaboratively and arbitrarily deviate from their pre-specified programs while a designated master node iteratively computes the model or parameter vector [4]. Byzantine-resilient SGD in high dimensions on heterogeneous data considers the master-worker architecture under Byzantine attacks with heterogeneous local datasets and no probabilistic assumptions on data generation, using a polynomial-time outlier-filtering procedure for robust mean estimation at its core [5]. In the omics setting the failures are not adversarial, but the structural

question is identical: when a reproduction pipeline is assembled from k independently fallible retrieval steps, end-to-end reliability is a function of per-step reliability, and this literature supplies the vocabulary of tolerating t failures among m components. Neither summary states a resilience threshold, so Section 4 derives the corresponding majority-vote requirement arithmetically rather than importing one.

Local reanalysis methods appear in structural engineering: the meshless-based local reanalysis (MLR) method extends reanalysis methods to the Kriging interpolation meshless method, using combined approximations and indirect factorization updating to reduce the computational cost of meshless methods [3]. There, "reanalysis" means recomputing a modified structural model cheaply from a previously solved one; the transferable idea is that reanalysis is worthwhile exactly when full recomputation is expensive, which is the economic premise of omics reanalysis as well. The summary is truncated before quantitative results, so the relevance here is conceptual only.

Reuse of existing measured data for parameter constraint is illustrated by the dark-energy study that uses H II starburst galaxy apparent magnitude versus redshift data from Siegel et al. (2005) to constrain cosmological model parameters, finding constraints generally consistent with those from other datasets but not as restrictive as the tightest currently available [7]. This is a concrete instance of the payoff side of the ledger: published data, when retrievable, support new scientific constraints, and the quality of what can be extracted depends on what the data can support — an analogy for the information loss incurred when reanalysis works from published rather than raw records.

Finally, four QNFO entries bound what an agent can and cannot extract. The information-theoretic limits work shows that an organism does not contain enough organism-specific information to specify its own microscopic organization, treating the genome as a generator specification rather than a trajectory program, and establishes a coarse-graining threshold with dual Hartley and Shannon entropy faces [11]. For data discovery this implies that even a perfect retrieval agent cannot recover information that was never encoded in the published artifacts; Section 4 adopts the dual Hartley/Shannon framing for the metadata specification problem. The LiveBench-grounded audit of frontier large language models for scientific research and agentic AI reports, from the LiveBench 2026-06-25 contamination-free benchmark, that no Llama model ranks in the top 42 and that DeepSeek V4 Pro 0813 is the mathematics price-performance optimum [9]; its summary is truncated mid-sentence, so we use it only to note that model choice for the retrieval agent is an auditable, benchmark-grounded decision and that per-token cost is a model-dependent free parameter. The pattern-based ontology entry (PBO/Autaxys) concerns intrinsic self-ordering of patterns [10], and the categorical formalization entry states a $\text{Capability}(S) \sim \text{End}(X_S)$ conjecture with a Church-Turing-scoped resource-level thesis distinguishing computability-level equivalence from resource-level difference [12]; both are supplied with summaries too thin to support specific claims, and we cite them only as adjacent formal work — [10] as an unmodeled alternative to field-extraction ontology, [12] for the resource-level distinction used in Section 6.

3. Methods

3.1 Pipeline formalization

Following [1], the pipeline has three functions — retrieval, reanalysis, and synthesis. We formalize these as five sequential stages, each implemented by a specialist agent in the modular, communicating pattern of [2] and [8]:

$$S_1 \rightarrow S_2 \rightarrow S_3 \rightarrow S_4 \rightarrow S_5$$

with stages:

- S_1 (corpus scan): identify candidate papers and determine whether omics data exist for each.
- S_2 (asset location): locate raw data deposits in public repositories and any intermediate data files, whose location [1] notes is irregular.
- S_3 (protocol extraction): reconstruct the analysis pipeline from main text, supplementary files, and code repositories [1].
- S_4 (reanalysis execution): run the reconstructed pipeline on the located data, with a verifier of the form used in agentic discovery over deployed stacks — the pipeline either runs and reproduces a reported result or it does not [6].
- S_5 (synthesis): aggregate reanalyzed outputs across studies into a common, searchable form.

Each stage S_i has an input context C_i , an output record O_i , and a message channel to downstream stages, in the sense of [2] that communication broadens each agent's view of the shared environment. The output record O_3 — the extracted protocol — is the pivotal artifact: it is the metadata specification on which S_4 depends, and Section 4.2 bounds its information content.

3.2 Dispersal model of a published omics study

We model a published omics study as a set $A = \{a_1, \dots, a_k\}$ of k reproduction-critical artifacts. An artifact is any item a reuser must obtain to reproduce or extend a reported result: the raw data deposit, the processing parameters, the pipeline code, an intermediate table such as a protein abundance matrix, or a mapping from figure to underlying data. Each artifact a_i resides in exactly one location ℓ_i drawn from the set of document locations $L = \{\text{article main text, supplementary files, code repository, external repository landing page, author contact}\}$. The dispersal breadth of the study is k ; the location spread is the number of distinct elements among $\{\ell_1, \dots, \ell_k\}$.

The central empirical claim of the motivating work, taken from its supplied text, is that this information "is dispersed across main text, supplementary files, and code repositories" and that intermediate data are shared only irregularly [1]. The model makes that claim parameterized: k and the location spread are measurable per study.

3.3 Retrieval reliability

Let $p \in [0, 1]$ be the per-attempt success probability of retrieving and correctly interpreting a single artifact from its location. For a human reuser, p aggregates finding the location, gaining access, and interpreting the artifact. For an agent, we separate a first-attempt success probability p from r independent retries at the same location, each succeeding with probability p conditional on prior failures being independent. The per-location success probability with r retries is

$$P_{\text{loc}}(p, r) = 1 - (1 - p)^{r+1}.$$

3.4 End-to-end reproduction success and cost

Assuming artifact retrievals are independent, the end-to-end reproduction success probability is

$$P_{\text{rep}} = \prod_{i=1}^k P_{\text{loc},i},$$

which under the homogeneous assumption $P_{\text{loc},i} = P_{\text{loc}}$ becomes

$$P_{\text{rep}}(k, p, r) = [1 - (1 - p)^{r+1}]^k.$$

The expected number of retrieval visits per artifact is the mean of a geometric random variable truncated at $r + 1$ attempts:

$$E[N] = \sum_{j=1}^{r+1} j (1 - p)^{j-1} p + (r + 1)(1 - p)^{r+1},$$

which for the no-retry case $r = 0$ reduces to $E[N] = 1/p$. The expected total retrieval visits for the study is $k \cdot E[N]$.

3.5 Cost-model and information-model inputs

The token cost model of Section 4.1 takes as free parameters: paper length in tokens L , the number of documents consulted per paper D , the number of decision fields to extract F , and tokens per extraction call overhead f_{oh} . All values used are stated assumptions of this analysis, not measurements; they are chosen to bracket the range a typical omics paper plausibly occupies, and Section 6 treats their sensitivity explicitly.

Section 4.2 applies the dual Hartley/Shannon framing of [11]. For a protocol specification with F independent fields, each with n_j admissible values, the Hartley entropy is

$$H_0 = \log_2 \prod_{j=1}^F n_j = \sum_{j=1}^F \log_2 n_j,$$

and the Shannon entropy $H = -\sum_j p_j \log_2 p_j \leq H_0$ measures the uncertainty actually reduced by extracted evidence. H_0 bounds what a complete specification must be able to express; H bounds what an agent must resolve per paper.

3.6 What the model does not claim

The dispersal model treats artifacts as binary (retrieved or not) and retrievals as independent. It does not model semantic difficulty, cross-artifact consistency checking, or the compute cost of executing a retrieved pipeline; those enter only as qualitative considerations. All numerical results below follow from the closed forms above with stated inputs.

4. Analysis

4.1 Token-and-call cost of the retrieval stage

Inputs (stated assumptions of this analysis). Paper main text: $L = 8000$ words at 1.3 tokens per word, giving $L_{\text{tok}} = 8000 \times 1.3 = 10400$ tokens. Supplementary and code documents consulted: $D = 5$ documents, average 2000 tokens each, giving $5 \times 2000 = 10000$ tokens. Decision fields to extract: $F = 25$ (a bracket of 10 to 40 is carried through). Tokens per call overhead (prompt scaffolding, schema, tool descriptions): $f_{\text{oh}} = 500$. One extraction call per field plus one summarization call per document gives $D + F$ calls.

Arithmetic. Total structured input tokens:

$$T_{\text{in}} = L_{\text{tok}} + D \times 2000 + (D + F) \times f_{\text{oh}}.$$

With $L_{\text{tok}} = 10400$, $D = 5$, $F = 25$, $f_{\text{oh}} = 500$:

$$T_{\text{in}} = 10400 + 10000 + 30 \times 500 = 10400 + 10000 + 15000 = 35400 \text{ tokens.}$$

Each extraction call also carries document context; assume the agent re-reads an average of 1000 tokens per extraction call, adding $F \times 1000 = 25000$ tokens:

$$T_{\text{total}} = 35400 + 25000 = 60400 \text{ tokens per paper.}$$

Number of agent calls:

$$N_{\text{calls}} = D + F = 5 + 25 = 30 \text{ calls per paper.}$$

Bracketing. For $F \in \{10, 40\}$, holding other inputs fixed:

- Lower:
 $T_{\text{total}} = 10400 + 10000 + (5 + 10) \times 500 + 10 \times 1000 = 10400 + 10000 + 7500 + 10000 = 37900$
tokens; $N_{\text{calls}} = 15$.
- Upper:
 $T_{\text{total}} = 10400 + 10000 + (5 + 40) \times 500 + 40 \times 1000 = 10400 + 10000 + 22500 + 40000 = 82900$
tokens; $N_{\text{calls}} = 45$.

If the corpus-scan stage S_1 must screen $S = 10$ candidate papers per accepted paper, each at $L_{\text{tok}} = 10400$ tokens with one call of overhead 500, the screening cost is $10 \times (10400 + 500) = 109000$ tokens, and the full retrieval stage for one accepted paper is

$$T_{\text{retr}} = 109000 + 60400 = 169400 \text{ tokens,}$$

with bracket $T_{\text{retr}} \in [146900, 191900]$ tokens. These are projections under the stated assumptions, not measurements.

4.2 Information bound on the reanalysis specification

Inputs. The extracted protocol must fix, per [1], the pipeline steps, parameters, software versions, input files, and output definitions. Assume $F = 25$ fields with geometric-mean cardinality $\bar{n} = 10$ admissible values per field (e.g., a normalization method chosen among roughly ten standard options). Then:

$$H_0 = F \log_2 \bar{n} = 25 \times \log_2 10 = 25 \times 3.3219 = 83.05 \text{ bits.}$$

For the bracket $F \in \{10, 40\}$:

- Lower: $H_0 = 10 \times 3.3219 = 33.22$ bits.
- Upper: $H_0 = 40 \times 3.3219 = 132.88$ bits.

Shannon comparison. If the most common value of each field has probability $p = 0.5$ and the remaining 0.5 is spread over the other nine values, then with probability 0.5 the field is resolved for free and with probability 0.5 it carries at most $\log_2 10 = 3.3219$ bits:

$$H \leq 0.5 \times 0 + 0.5 \times 3.3219 = 1.661 \text{ bits per field,}$$

and over $F = 25$ fields, $H \leq 25 \times 1.661 = 41.5$ bits. The gap between $H_0 = 83.05$ bits and $H \leq 41.5$ bits is the dual-face result in the sense of [11]: the specification space the agent must be able to express is, in this parameterization, twice what a typical paper actually resolves, which is why defaults and conventions absorb part of the curation burden.

The binding constraint. The bound that matters for feasibility is not the agent's capacity to express $H_0 \approx 83$ bits — trivially small by machine standards — but whether the published record contains the evidence to resolve those bits. [1] states that the needed information is dispersed across main text, supplementary files, and code repositories, and that intermediate data locations are irregular; the information-theoretic framing says the agent's task is to reduce H to 0 per field using only those dispersed sources. Where a field's value is stated nowhere, its conditional entropy stays at its prior, and the reanalysis is underdetermined. This is the formal statement of the accessibility gap [1] describes qualitatively.

4.3 Reusability index

Define the reusability index R as the fraction of fields an agent can resolve from the published record without human intervention:

$$R = \frac{F_{\text{resolved}}}{F}.$$

Under the Shannon model of Section 4.2, if each field is stated with probability $p = 0.5$, then $F_{\text{resolved}} = pF$ and $R = p = 0.5$. Sensitivity: R is linear in p . If publication practice improves so that $p = 0.9$, the human curation burden per paper falls from $F(1 - 0.5) = 12.5$ fields to $F(1 - 0.9) = 2.5$ fields — a fivefold reduction, computed as $12.5/2.5 = 5$. The agent absorbs exactly the resolved fraction, and the residual $(1 - R)F$ fields are the irreducible human cost under current practice. These are model projections, not measurements of any corpus.

4.4 Manual reproduction (no retries, $r = 0$)

With $r = 0$, per-location success is $P_{\text{loc}} = p = 0.8$. End-to-end, with $k = 5$:

$$P_{\text{rep}}(5, 0.8, 0) = 0.8^5.$$

Computing step by step: $0.8^2 = 0.64$; $0.8^3 = 0.64 \times 0.8 = 0.512$; $0.8^4 = 0.512 \times 0.8 = 0.4096$; $0.8^5 = 0.4096 \times 0.8 = 0.32768$.

$$P_{\text{rep}}(5, 0.8, 0) = 0.32768.$$

Under these assumptions a manual reproduction attempt succeeds about 32.8% of the time; roughly two of every three attempts fail. Expected retrieval visits per artifact with $r = 0$ is $E[N] = 1/p = 1/0.8 = 1.25$, and total expected visits:

$$k \cdot E[N] = 5 \times 1.25 = 6.25.$$

4.5 Agent reproduction with $r = 2$ retries

Per-location success with two retries:

$$P_{\text{loc}}(0.8, 2) = 1 - (1 - 0.8)^3 = 1 - 0.2^3 = 1 - 0.008 = 0.992.$$

End-to-end:

$$P_{\text{rep}}(5, 0.8, 2) = 0.992^5.$$

Computing: $0.992^2 = 0.984064$; $0.992^3 = 0.984064 \times 0.992 = 0.976191488$;
 $0.992^4 = 0.976191488 \times 0.992 = 0.968381956096$;
 $0.992^5 = 0.968381956096 \times 0.992 = 0.960634900447232$.

$$P_{\text{rep}}(5, 0.8, 2) \approx 0.9606.$$

The agent raises reproduction success from 0.32768 to ≈ 0.9606 , a factor of

$$\frac{0.960634900447232}{0.32768} \approx 2.9316.$$

Expected visits per artifact with $r = 2$:

$$E[N] = \sum_{j=1}^3 j(0.2)^{j-1}(0.8) + 3(0.2)^3.$$

Term by term: $j = 1$: $1 \times 1 \times 0.8 = 0.8$. $j = 2$: $2 \times 0.2 \times 0.8 = 0.32$. $j = 3$: $3 \times 0.04 \times 0.8 = 0.096$.
Truncation term: $3 \times 0.008 = 0.024$. Sum:

$$E[N] = 0.8 + 0.32 + 0.096 + 0.024 = 1.24.$$

Total expected visits: $5 \times 1.24 = 6.20$. The agent achieves its $\approx 2.93 \times$ success improvement at essentially the same expected visit count (6.20 vs 6.25), because retries are triggered only on the 20% of first attempts that fail.

4.6 Sensitivity to dispersal breadth k

Holding $p = 0.8, r = 0$:

$$P_{\text{rep}}(k) = 0.8^k.$$

- $k = 2$: 0.64
- $k = 3$: 0.512
- $k = 4$: 0.4096
- $k = 5$: 0.32768
- $k = 8$: $0.8^8 = 0.16777216$
- $k = 10$: $0.8^{10} = 0.1073741824$

Each additional dispersed artifact multiplies success by 0.8. Going from $k = 2$ to $k = 5$ cuts success from 0.64 to 0.32768, a 48.8% relative reduction ($1 - 0.32768/0.64 = 1 - 0.512 = 0.488$).

With the agent ($r = 2$), the same range gives 0.992^k : at $k = 10$, $0.992^{10} = (0.992^5)^2 = 0.960634900447232^2 = 0.922819411 \dots \approx 0.9228$. The agent's advantage grows with k : at $k = 10$ manual success is 0.1074 versus agent 0.9228, a factor of $0.922819411/0.1073741824 \approx 8.594$.

4.7 Sensitivity to per-location reliability p

Holding $k = 5, r = 0$, manual success is p^5 :

- $p = 0.6$: $0.6^5 = 0.07776$
- $p = 0.8$: 0.32768
- $p = 0.95$: $0.95^5 = 0.7737809375$

The barrier is jointly driven: even a highly reliable reuser ($p = 0.95$) succeeds only 77.4% of the time at $k = 5$, while a moderately reliable agent with retries at $p = 0.8$ reaches 96.0%. Retries substitute for per-location skill: $P_{\text{loc}}(0.6, 2) = 1 - 0.4^3 = 1 - 0.064 = 0.936$, so a weaker retriever with retries (0.936 per location) outperforms a stronger one without (0.8 per location).

4.8 Resilience to corrupted sources

Map each consulted source to a worker in the sense of [4]: m sources inform a field, of which t may be wrong (mis-labeled repository records, outdated supplementary files, contradictory code comments). A majority vote over m independent sources tolerates up to $\lfloor (m - 1)/2 \rfloor$ corrupt sources. For $m = 3$:

$$t_{\text{max}} = \left\lfloor \frac{3 - 1}{2} \right\rfloor = 1,$$

so three independent sources per field tolerate one corrupt source; for $m = 5$, $t_{\text{max}} = 2$. The cost of this tolerance is additional retrieval: consulting m sources per field multiplies the context-reading term of Section 4.1 by m . With $m = 3$ and $F = 25$:

$$T_{\text{ctx}} = 3 \times 25 \times 1000 = 75000 \text{ tokens},$$

versus 25000 at $m = 1$ — a threefold increase in the context term, raising the no-screening total from 60400 to $60400 - 25000 + 75000 = 110400$ tokens per paper. The structural point, following [4] and [5], is that resilience is purchasable in tokens: the master-agent must either consult redundant sources or accept single-point-of-failure metadata. Neither [4] nor [5] states a domain-specific threshold in its supplied summary, so this derivation is self-contained arithmetic on the majority-vote structure, not an imported guarantee.

5. Results

The results below are the numerical consequences of the closed forms in Section 3 and the arithmetic in Section 4, under the stated assumptions ($k = 5$ dispersed artifacts, per-location first-attempt success $p = 0.8$, agent retries $r = 2$, $F = 25$ extraction fields, $D = 5$ consulted documents). They are model outputs, not measurements of any corpus.

R1. Manual reproduction success is low and falls multiplicatively with dispersal. With no retries, $P_{\text{rep}}(5, 0.8, 0) = 0.8^5 = 0.32768$ (Section 4.4): roughly one manual attempt in three succeeds. Each additional dispersed artifact multiplies success by 0.8; at $k = 10$, manual success is 0.1073741824 (Section 4.6).

R2. Agent retries raise success to ≈ 0.9606 at essentially unchanged retrieval effort. With $r = 2$, per-location success is 0.992 and end-to-end success is $0.992^5 \approx 0.9606$ (Section 4.5), a factor of ≈ 2.9316 over the manual case. Expected visits are 6.20 versus 6.25 manual, because retries fire only on the 20% of first attempts that fail.

R3. The retrieval stage costs 60400 tokens and 30 calls per paper before screening. With screening at a 10:1 ratio, the full retrieval stage is 169400 tokens per accepted paper, bracketed at [146900, 191900] tokens for $F \in \{10, 40\}$ (Section 4.1). These are projections under stated assumptions.

R4. The specification space is about twice what a typical paper resolves. Hartley entropy of the protocol specification is $H_0 = 83.05$ bits at $F = 25$, $\bar{n} = 10$; the Shannon comparison gives $H \leq 41.5$ bits (Section 4.2). The unresolved remainder, $R = 0.5$ of fields under the stated prior, is the irreducible human curation burden under current practice (Section 4.3).

R5. Dispersal breadth k dominates the barrier; retries substitute for per-location reliability. Going from $k = 2$ to $k = 5$ cuts manual success by 48.8% relative (Section 4.6). A weaker retriever with retries ($P_{\text{loc}}(0.6, 2) = 0.936$) outperforms a stronger one without (0.8) (Section 4.7). At $k = 10$ the agent's advantage grows to a factor of ≈ 8.59 (0.9228 vs 0.1074).

R6. Resilience to corrupted sources is purchasable in tokens. Majority voting over $m = 3$ independent sources per field tolerates one corrupt source and raises the context term from 25000 to 75000 tokens, lifting the no-screening retrieval total from 60400 to 110400 tokens per paper (Section 4.8).

6. Discussion

The central claim of this paper is that the omics reuse barrier decomposes into three measurable quantities — dispersal breadth k , per-location reliability p , and the resolved-field fraction R — and that under the stated parameterization the barrier is dominated by k , while agent retries cheaply compensate for low p . The claim's strength is exactly the strength of its assumptions, which are stated assumptions of this analysis rather than measurements.

Limitations. The result rests on the following assumptions, each with its sensitivity:

- **Independence of artifact retrievals.** $P_{\text{rep}} = P_{\text{loc}}^k$ assumes failures are uncorrelated. If failures correlate (e.g., a single missing supplementary file blocks several artifacts), effective k is smaller and manual success is higher; the model then overstates the barrier. If failures correlate through a common cause (a retracted dataset), success is lower than modeled.
- **Homogeneous p across artifacts and locations.** Real per-location success varies; a raw deposit in a structured repository is easier to retrieve than a parameter buried in a methods paragraph. The homogeneous model is a first-order average; heterogeneous p_i would replace P_{loc}^k with $\prod_i P_{\text{loc},i}$ and shift the dominant failure contribution to the lowest- p artifact.
- **Binary artifacts.** The model treats retrieval as success/failure and does not score partial or incorrect retrieval, semantic misinterpretation, or cross-artifact inconsistency; Section 4.8 prices redundancy but does not model error rates of sources.
- **Stated, not measured, inputs.** $k = 5$, $p = 0.8$, $F = 25$, $\bar{n} = 10$, $L = 8000$ words, $D = 5$, and the 10:1 screening ratio are all assumptions chosen to bracket plausible ranges. The headline numbers (0.32768, 0.9606, 60400, 169400, 83.05) scale directly with these: P_{rep} is exponential in k and in $(1 - p)$, and token cost is linear in L , D , and F .
- **No execution cost.** The model prices retrieval only. Running a reconstructed pipeline (compute, environment setup, debugging) is unmodeled and plausibly dominates total reanalysis cost in wall-clock terms.
- **The Shannon comparison is illustrative.** The $p = 0.5$ field-stated prior is a placeholder; the qualitative conclusion (specification space exceeds what papers resolve) is robust to the exact value, but the specific $R = 0.5$ is not a corpus measurement.

What would falsify the main result. The main result — that dispersal breadth dominates the barrier and retries substitute for per-location reliability — is falsified if a measured corpus shows either of the following. First, if a sample of published omics studies has median dispersal breadth $k \leq 2$ with per-location success near the modeled 0.8, then manual reproduction success is ≥ 0.64 and the barrier is not dispersal-dominated; the model's premise fails. Second, if measured per-location success for a competent human is ≥ 0.95 at $k = 5$, manual success is 0.774 and the gap an agent closes shrinks from a factor of 2.9316 to 1.24 (0.9606/0.774), making the agent's practical value marginal. Conversely, if measured p for agents is far below 0.8 on real dispersed documents, the modeled agent success 0.9606 collapses: at $p = 0.5$, $P_{\text{loc}}(0.5, 2) = 1 - 0.5^3 = 0.875$ and $P_{\text{rep}} = 0.875^5 \approx 0.513$. Any of these measurements, obtainable from a few dozen studies, would overturn or confirm the parameterization.

Open questions. Three questions the model raises but does not answer. (i) What is the empirical distribution of k and location spread across a real omics corpus? This is the first measurement Section 7 proposes. (ii) How correlated are retrieval failures across artifacts within a study? The independence assumption is the model's most fragile point. (iii) Does the resource-level distinction of [12] — computability-level equivalence versus resource-level difference — apply to the agent-versus-human comparison here, in the sense that both can in principle resolve the specification but differ sharply in the resources each must spend? The supplied summary of [12] states the thesis but gives no domain application, so this remains a framing, not a result. Likewise, the pattern-based ontology of [10] is supplied with a summary too thin to evaluate as an alternative to field-extraction metadata, and we leave it as an unmodeled alternative.

7. Practical Implications and Engineering Pathways

(a) Who would use the result and what changes. Three user groups can act on the three measured quantities. Journal and repository operators can report a per-deposit reuse-readiness figure — dispersal breadth k , location spread, and the fraction of the F protocol fields stated in the deposit — alongside acceptance, turning "most data are functionally inaccessible" [1] into a per-study score with a defined improvement lever (reduce k , consolidate locations, state more fields). Agent developers can use the cost model to budget retrieval: the 60400-token, 30-call figure per paper before screening, and 169400 tokens at a 10:1 screening ratio, are planning numbers for corpus-scale runs, and the sensitivity result says the engineering priority is raising

effective per-location reliability (retries, redundant sources) rather than raw model capability. Authors preparing supplements can see, via the reusability index R , that moving the stated-field fraction from 0.5 to 0.9 cuts the residual human curation burden fivefold per paper (Section 4.3). Because this paper derives the quantities but measures none of them on a real corpus, its immediate use is as a measurement design, not as a deployed tool; in that sense the work is foundational, and its value is that it specifies exactly which numbers the first measurement must produce.

(b) Current readiness. This is purely analytical work that has not been peer reviewed: under the NASA Technology Readiness Level definitions it sits at TRL 1–2 (basic principles observed and reported; concept formulated). The pipeline formalization of Section 3 is a paper design, not an implemented system; no component of this work has been demonstrated in hardware or software. Analytical work of this kind does not exceed TRL 3 even after proof-of-concept calculation, and nothing here supports a claim of operational readiness.

(c) The smallest next experiment. The minimal de-risking study is a manual audit of a small, fixed sample — for example, 20 published omics studies drawn from a single journal — recording for each: the number of reproduction-critical artifacts k , the set of locations $\{\ell_i\}$, and, for a fixed field list of $F = 25$ protocol fields, the fraction stated anywhere in the published record. The outcome that refutes the model’s premise is a median $k \leq 2$ with a stated-field fraction above 0.9; the outcome that confirms it is a distribution of k concentrated at 4 or more with a stated-field fraction near 0.5. A second, cheap step is to replay the retrieval stage of Section 4.1 with an existing LLM agent on the same 20 studies and measure actual tokens, calls, and per-location success p against the assumed 0.8; a measured agent p below roughly 0.5 refutes the projected agent success of 0.9606 (Section 6).

(d) Binding constraints. In the order they would bind first, as derived or directly implied by the model: (i) **data** — the audit requires human-labeled ground truth for k , locations, and stated fields, which is the scarcest input and the main cost of the first experiment; (ii) **token and call budget** — Section 4.1 derives 169400 tokens per accepted paper at a 10:1 screening ratio, so a corpus of 10^4 accepted papers costs on the order of 1.7×10^9 tokens for retrieval alone, a projection that scales linearly with the stated per-paper figures; (iii) **redundancy cost** — majority voting over $m = 3$ sources raises the no-screening total to 110400 tokens per paper (Section 4.8), so resilience and budget trade off directly; (iv) **model choice and per-token price** — per-token cost is a model-dependent free parameter, and the LiveBench-grounded audit [9] treats model selection for scientific and agentic use as an auditable benchmark decision, though its supplied summary is truncated and gives no per-token prices. Regulation and manufacturing constraints do not bind at this stage; the binding inputs are labeled data and compute budget.

(e) Order-of-magnitude projections. All figures in this paragraph are projections under the assumptions of Section 4.1, not measurements. At 169400 tokens per accepted paper (10:1 screening), a 10^4 -paper corpus requires $\approx 1.694 \times 10^9$ retrieval tokens; at the no-screening figure of 60400 tokens per paper, $\approx 6.04 \times 10^8$ tokens. These scale linearly in L , D , F , and the screening ratio, all of which are stated assumptions. No cost in currency is derivable from the supplied grounding, since no per-token price is stated in any bibliography entry, so we leave monetary cost as an open question.

(f) Risks and misuse paths. Two direct risks follow from the design. First, an agent that fills unstated protocol fields with defaults rather than flagging them converts the modeled underdetermination (Section 4.2) into silent fabrication: the reanalysis runs and produces numbers whose provenance is a guessed default. Any implementation should therefore report R — the resolved-field fraction — with every reanalysis, so downstream users can see how much of the specification was actually retrieved. Second, a reuse-readiness score derived from this model could be gamed by depositing nominally complete but semantically thin artifacts, since the model counts artifacts as binary; a scoring policy would need the stated-field audit of (c), not just artifact counts, to resist this.

8. Conclusion

This paper turned the qualitative observation that most published omics data are functionally inaccessible for computational reuse [1] into a parameterized model with closed-form outputs. A published study is characterized by its dispersal breadth k , per-location retrieval reliability p , and the fraction R of protocol fields the published record resolves. Under the stated assumptions ($k = 5$, $p = 0.8$), a manual reproduction succeeds with probability 0.32768; an agent with two retries succeeds with probability ≈ 0.9606 at essentially the same expected visit count (6.20 vs 6.25), because retries substitute for per-location reliability at almost no additional expected effort. The retrieval stage costs 60400 tokens and 30 calls per paper before screening, or 169400 tokens at a 10:1 screening ratio, and resilience to corrupted sources is purchasable at a derived token price

(110400 tokens per paper at $m = 3$ sources). The information-theoretic framing, adopting the dual Hartley/Shannon structure of [11], shows the specification space ($H_0 = 83.05$ bits at $F = 25$) is about twice what a typical paper resolves ($H \leq 41.5$ bits), which locates the binding constraint in publication practice rather than agent capability. Every number here follows from stated assumptions, not measurements; the model's value is that it names exactly what the first measurement must record — k , location spread, p , and R on a real corpus — and specifies what outcome would falsify it. If those measurements confirm the parameterization, deposition policy and agent budgeting acquire concrete, optimizable targets; if they refute it, the model will have failed in a way that itself produces the field's first quantitative picture of the reuse barrier.

Provenance and AI Involvement

- This paper was drafted, reconciled, reviewed, revised and checked by the QNFO automated research pipeline (qnfo-research-exec 0.16.0-revise-retry). Language models were tools in that process; they are not authors. The named author is accountable for the content.
- Models: drafting by openai/gpt-oss-120b, zai-org/glm-5.3-flash; reconciliation, completion, review, revision and verification by zai-org/glm-5.3-flash, zai-org/glm-5.3, openai/gpt-oss-120b, with the QNFO model router as a fallback.
- Sources: every reference is drawn from a bibliography retrieved from arXiv and the QNFO corpus before drafting; statements about a cited work are restricted to its retrieved title and abstract.
- Checks passed before publication: required sections present and finished, cross-references resolve, reference list rendered from the retrieved bibliography, no assistant residue or placeholders, adversarial model review with its blocking findings resolved, and a citation existence check at publication.
- Independent recomputation: a separately generated script recomputed 15 quantitative claims from the paper's stated inputs; 0 differed from the stated value at the stated precision.
- Not checked: whether each cited source supports the sentence that cites it beyond its abstract; derivation steps other than the recomputed claims. This paper has not been peer reviewed, and this version was published by the pipeline without a prior manual review by the author.

References

- [1] Omics Data Discovery Agents: Agent-Supported Retrieval, Reanalysis, and Synthesis of Published Omics Data. arXiv:2603.10161v3. <https://arxiv.org/abs/2603.10161v3> [2] A Survey of Multi-Agent Deep Reinforcement Learning with Communication. arXiv:2203.08975v2. <https://arxiv.org/abs/2203.08975v2> [3] A Meshless-based Local Reanalysis Method for Structural Analysis. arXiv:1703.04355v2. <https://arxiv.org/abs/1703.04355v2> [4] Data Encoding for Byzantine-Resilient Distributed Optimization. arXiv:1907.02664v2. <https://arxiv.org/abs/1907.02664v2> [5] Byzantine-Resilient SGD in High Dimensions on Heterogeneous Data. arXiv:2005.07866v1. <https://arxiv.org/abs/2005.07866v1> [6] Declarative Data Services: Structured Agentic Discovery for Composing Data Systems. arXiv:2605.20690v2. <https://arxiv.org/abs/2605.20690v2> [7] Constraints on dark energy from H II starburst galaxy apparent magnitude versus redshift data. arXiv:1110.5626v1. <https://arxiv.org/abs/1110.5626v1> [8] LLandMark: A Multi-Agent Framework for Landmark-Aware Multimodal Interactive Video Retrieval. arXiv:2603.02888v1. <https://arxiv.org/abs/2603.02888v1> [9] QNFO: Prioritizing Large Language Models for Scientific Research and Agentic AI: A LiveBench-Grounded Audit (August 2026) [10] QNFO: Pillar 3: PBO/Autaxys — Pattern-Based Ontology as Intrinsic Self-Ordering [11] DOI 10.64898/2026.07.27.740886. QNFO: Information-theoretic Limits on Programmatic Specification of Biological Systems. [12] QNFO: The Substrate Is the Algorithm: A Categorical Formalization of the Strongest Consilience Thread