Addressing Unreliability Propagation in Scientific Digital Libraries

Heng Zheng*
zhenghz@illinois.edu
University of Illinois Urbana-Champaign
Champaign, Illinois, USA

Yuanxi Fu*
fu5@illinois.edu
University of Illinois Urbana-Champaign
Champaign, Illinois, USA

M. Janina Sarol
mjsarol@illinois.edu
University of Illinois Urbana-Champaign
Champaign, Illinois, USA

Ishita Sarraf†
sarrafis@grinnell.edu
Grinnell College
Grinnell, Iowa, USA

Jodi Schneider‡
jodi@illinois.edu
University of Illinois Urbana-Champaign
Champaign, Illinois, USA

Abstract

Today's scientists rely on scientific artifacts developed by others for their work. Individual scientists often have limited capacity to assess the validity of these resources. When errors are not caught, scientists produce second-generation errors. We say that a publication *propagates unreliability* when the main contribution of the publication becomes unreliable by using an unreliable source. An approach for checking whether publications propagate unreliability should satisfy three requirements, in priority order: (1) not miss any publications that propagate unreliability; (2) provide rationales; and (3) identify all publications that do not propagate unreliability. We consider three approaches: a base approach using metadata of the citing publications and the section headings of the citation contexts; and supplementing the base approach with either keyword-based or machine-learning-based modules. The base approach is the most generalizable. Approach-KW (base+keyword) provides concrete rationales, which could be important for convincing authors and editors to take action to update publications that propagate unreliability. Approach-ML (base+machine learning) has the best performance. Future work should develop a more general framework using multiple case studies. We will build a human-in-the-loop alerting system that digital library maintainers, editors, and authors could use to triage publications that may propagate unreliability, and maintain the quality of scientific digital libraries.

CCS Concepts

• **Information systems** → **Digital libraries and archives**; • **Applied computing** → *Document metadata.*

*Both authors contributed equally to the paper.
†Work done at the University of Illinois Urbana-Champaign
‡Jodi Schneider is also affiliated with Harvard University, Massachusetts, USA

JCDL '24, December 16–20, 2024, Hong Kong, China

ACM ISBN 979-8-4007-1093-3/24/12
https://doi.org/10.1145/3677389.3702526

Keywords

unreliable cited sources, knowledge maintenance, citations, scientific digital libraries, scholarly publications, reproducibility

ACM Reference Format:
Heng Zheng, Yuanxi Fu, M. Janina Sarol, Ishita Sarraf, and Jodi Schneider. 2024. Addressing Unreliability Propagation in Scientific Digital Libraries. In *The 2024 ACM/IEEE Joint Conference on Digital Libraries (JCDL '24), December 16–20, 2024, Hong Kong, China.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3677389.3702526

1 Introduction

Today's scientific research is a collaborative enterprise. Scientists rely on scientific artifacts developed by others for their work, from knowledge claims to datasets and computer programs. Individual scientists often have limited capacity to assess the validity of these resources: a biologist may not know that the cell lines they use have been contaminated [3]; an organic chemist following a well-known computational chemistry protocol may not know that one of its scripts may malfunction on their Mac computer [2]; a computer scientist may not know that their deep-learning algorithm's outstanding performance in reconstructing medical images is the result of hidden data processing pipelines applied to their dataset [16]. In unlucky circumstances, when errors are not caught in time even after comprehensive validations, scientists produce second-generation errors that unsuspecting future users may continue to rely on for their work, sometimes for years before the underlying first-generation errors are noticed.

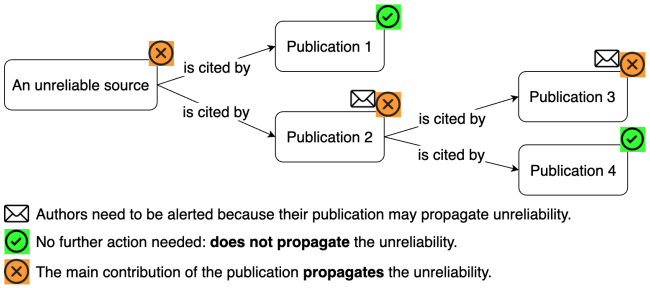

Figure 1: Not all citations propagate unreliability.

We coin the term "unreliability propagation" to refer to error propagation of this sort: First-generation errors such as the misidentified cell line and the malfunctioning script turn into second-generation errors such as false claims about diseases and incorrect chemical structures when new work incorporates these unreliable artifacts and cites them as (unreliable) sources. We say that a publication *propagates unreliability* when the main contribution of the publication becomes unreliable by using an unreliable source.

Such unreliability propagation threatens the quality of scientific digital libraries. Digital libraries must flag publications using an unreliable source and record whether their reliability was reviewed, the review date, review method (automatic or manual by authors or editors), and whether the publication needs correction.

Current approaches to handling publications that may propagate unreliability are either precise but time-consuming (through 100% manual checking in editorial offices or deep conceptual modeling [6]) or so automated (e.g., [19]) that experts are unlikely to fully rely on them. Tracing whether a given publication relies on any unreliable source is particularly important because the problems with a source could be detected at any time, even after decades [21].

Our goal in this paper is to design an approach that balances scalability and trustworthiness when checking whether a publication propagates unreliability. We experiment on designing a triage approach for the citing publications of one unreliable source [24]. Our experimentation and design process aims to answer the following questions:

RQ1: How can we triage citing publications based on their risk of propagating unreliability at scale?
RQ2: How well does each triage approach work and how explainable are the triage results?
RQ3: Can we use scientific digital library infrastructure to collect relevant data for triage?

2 Background

2.1 Existing approaches to unreliability propagation

To identify publications that propagate unreliability, researchers have developed approaches for auditing literature that cites unreliable work, relying on the fact that not all citations are equally influential in citing publications [5]. Fu and Schneider [6] developed the keystone framework and proposed its use for identifying publications whose content may be significantly affected by the unreliable sources they cited. Usman and Balke [19] developed citation intent analysis to determine whether a citing publication is dependent on retracted publications.

However, identifying publications that propagate unreliability is not well-addressed in current real-world digital libraries. Some digital library services identify publications that have been formally or informally flagged as unreliable. Very rarely, unreliable sources in bibliographies are flagged (e.g., PubMed Central labels retracted publications in red). Typically, to determine that a source publication is reliable, readers must click the CrossMark button to check whether a publication is up-to-date [12], install a plugin (e.g., PubPeer[1], RedacTek[2] [4]), or look the publication up on a website.

Among unreliable publications, retracted publications have received the most attention. One study found that biology publications citing retracted sources were more often retracted [29], perhaps in part due to the propagation of unreliable information. Yet retraction does not curtail the citation of retracted publications [14, 20]. To address this, the RetractoBot randomized controlled trial is testing whether alerting authors that publications they previously cited are retracted reduces future citations to retracted publications [22]. Meanwhile several citation management tools flag retracted publications whose DOIs are found in RetractionWatch [13]. For instance, RedacTek [4] flags publications citing retracted sources through three generations of citations.

2.2 Editorial processes

Editorial processes for quality control of manuscripts and scientific publications rely on domain experts. For peer review, responsible use of automated screening must supplement, not supplant, human editors and peer reviewers [15]. In checking for research integrity issues, automated tools can aid but not replace domain experts. For instance, Acuña et al.'s automatic detection of image reuse is designed to be checked by academic research integrity offices [1].

3 Task formulation

Digital libraries should flag publications using an unreliable source, and then determine which publications are the most likely to propagate unreliability. These high-risk publications need review by authors or editors. Specific, persuasive rationales could increase authors' and editors' willingness to check and update high-risk publications that may propagate unreliability.

Not all citations propagate unreliability. However, for a given unreliable source, each citing publication needs to be checked to assess whether it propagates unreliability, as shown in Figure 1. Unreliability can propagate at least to a second generation [14], also indicated in the figure. Ideally, the entire citation network of the unreliable source would be checked for unreliability [20]. This requires a scalable triage approach.

3.1 Requirements analysis

We identify requirements for a scalable triage approach:

Input. The input is a single unreliable source.

Output. The output is a risk level and a rationale.

Performance measures. In priority order:

(1) **Do not miss any publications that propagate unreliability**: Minimize false negatives, compared to experts.
(2) **Provide rationales**: Explain the risk level.
(3) **Remove all publications that do not propagate unreliability**: Minimize false positives, compared to experts.

[1] https://pubpeer.com/
[2] https://redactek.com/

3.2 An example unreliable source

To develop a scalable triage approach, we start with a detailed examination of a single unreliable source, design and compare approaches, and envision what will generalize to other unreliable sources. The unreliable source we experiment with is a computational chemistry protocol published in 2014 [24]. The protocol is used to predict Nuclear Magnetic Resonance (NMR) spectra, which in turn, assists organic chemists in characterizing the structure of newly isolated organic compounds. But an organic chemist may not know that one of its Python scripts malfunctions on their computer. Applying the protocol in different operating systems returns different values, as shown in Figure 2, resulting in malfunctions on some UNIX-based systems, as reported by Neupane et al. [2]. The original protocol's

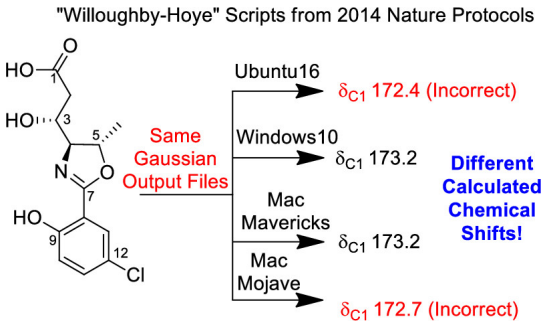

Figure 2: The calculation using some UNIX-based systems returned incorrect values. Reprinted with permission from [2]. Copyright 2019 American Chemical Society.

authors subsequently published an addendum announcing updates to the Python script that avoid the code glitch [25]. Figure 3 shows

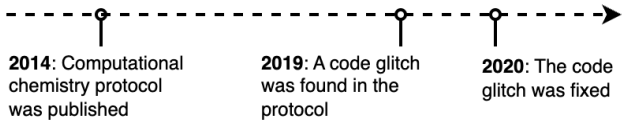

Figure 3: Timeline of reporting and correcting the code glitch related to Willoughby et al., 2014.

the timeline for reporting and correcting the code glitch.

We chose to use this example for our work because it is a concrete example of unreliability propagation. The protocol is "one of the most read and highly cited works in the field" [11] and was previously examined in a small-scale case study [6]. The code glitch in the protocol is relatively easy to understand. It is limited to one script and is solely computational, based on file-sorting mechanism differences between two families of commonly used platforms, UNIX-based systems (Linux and Mac OS) and Windows.

4 Experimentation

4.1 Data collection

On July 6, 2023, we identified publications citing Willoughby et al., 2014 and retrieved the metadata of 277 citing publications from Web of Science (WoS) and 285 citing publications from Scopus. After merging the dataset and removing duplicate items, 286 citing publications remained. We removed 2 publications that are not in English, leaving 284 citing publications in our dataset [28].

4.2 Initial manual extraction of citation contexts

One author (HZ) manually extracted citation contexts of Willoughby et al., 2014, resulting in 401 citation contexts extracted from 284 English-language full texts.

A citation context is a text (typically a clause, one or more sentences, or a paragraph) [17] in a scientific publication that contains a reference to the citation; Figure 4 shows an example. Author names, publication year, or bibliography entry number may be used as the citation marker for a publication.

Two articles [8, 25] cited the WH protocol for adopting conformation analysis, which is up-stream in the protocol from the problematic Python script D, thus not impacted by the code glitch. One [53] cited the WH protocol in the introduction as a previous successful example of characterizing molecules by using NMR chemical shift calculations.

[25] Guillen, P.O., Jaramillo, K.B., Jennings, L., Genta-Jouve, G., de la Cruz, M., Cautain, B., Reyes, F., Rodríguez, J. and Thomas, O.P. 2019. Halogenated tyrosine derivatives from the Tropical Eastern Pacific zoantharians Antipathozoanthus hickmani and Parazoanthus darwini. *Journal of Natural Products.* 82, 5 (May 2019), 1354–1360. DOI:https://doi.org/10.1021/acs.jnatprod.9b00173.

Figure 4: The publication "Fu & Schneider, 2020" uses the citation marker "25" to refer to Guillen et al., 2019 in the bibliography. The text in blue is the citation context of Guillen et al., 2019. "[8, 25]" is the full citation marker.

4.3 Manual annotation by chemistry experts

First, one chemistry domain expert (YF) spent 40+ hours annotating the corpus of 284 citing publications, using publication metadata and full text. To create a silver standard [28], we made a protocol [27], selected a representative sample of publications, then a second chemistry expert (contributor EV) annotated them, and we discussed and reconciled differences.

To select a representative sample, we (MJS) turned our full dataset of citation contexts into word embeddings, clustered them using similarity measures via BERTopic [7] with HDBSCAN algorithm, and selected representative citation contexts based on the centroids of the clusters.[3] Then the second chemistry expert (contributor EV) annotated the 77 publications associated with the citation contexts, which took 9.5 hours. Before chemistry experts YF and EV discussed differences in annotation together with JS, interannotator agreement on the double-annotated sample was .34 (fair agreement [9]), and after resolution of resolvable differences, interannotator agreement was .70 (substantial agreement [9]).

For the remaining 9 publications on which annotators still disagreed, JS made a reconciliation policy in consultation with YF, and YF updated the annotations, resulting in the silver standard we use for evaluating our triage approaches. The silver standard

[3] https://doi.org/10.5281/zenodo.13921537

categorized 86 (30.3%) publications as at risk of propagating unreliability and 198 (69.7%) publications as not at risk of propagating unreliability, with a rationale justifying each category.

5 Results for RQ1: Triage approaches

Figure 5 shows our overall triage strategy, which gives rise to three approaches: the *base approach*, *Approach-KW*, and *Approach-ML*. The base approach uses metadata and bibliographies of the citing publications (Stage 1) and the section headings of the citation contexts in which Willoughby et al., 2014 was cited (Stage 2). The other two approaches augment the base approach by analyzing the text of citation contexts (Stage 3) in two different ways, using either a keyword-based decision tree for *Approach-KW* or a machine learning-based model for *Approach-ML*.

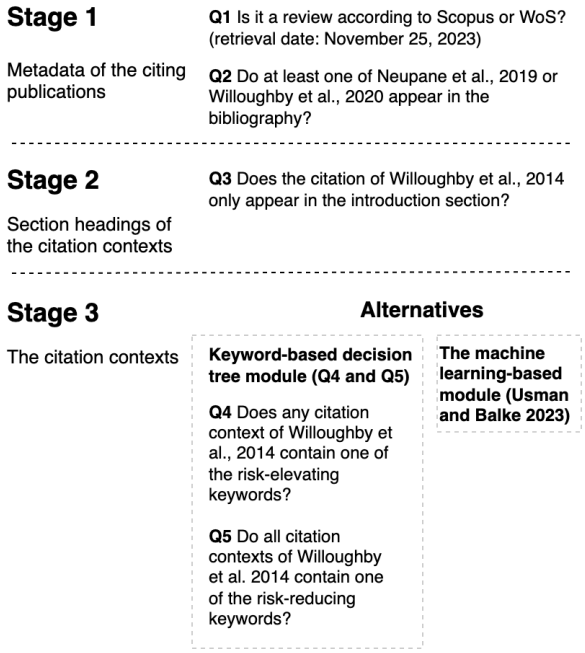

Figure 5: Our 3-stage strategy for determining the risk level.

5.1 Base approach (Stages 1 and 2)

First, we use the information easiest to retrieve from scientific digital libraries: metadata of the citing publications, including their bibliographies and publication types. Then, we extract information from the full text of citing publications: the section headings of the citation contexts and the citation contexts themselves.

Author HZ identified three questions that can be used to determine publications at negligible risk of propagating unreliability, when it is a review; when related corrective publications appear in the bibliography; or when the relevant citations appear only in the introduction section. Since these questions use limited domain knowledge, they are likely to be adaptable to other empirical research.

Q1: *Is it a review according to Scopus or WoS? (retrieval date: November 25, 2023)*
Since a review discusses published research, its main contribution is not expected to depend on the protocol's unreliable Python script. Hence we presume that review publications do not propagate unreliability due to the code glitch.

Q2: *Do at least one of Neupane et al., 2019 or Willoughby et al., 2020 appear in the bibliography?*
We presume that the main contribution does not propagate unreliability when authors show awareness of the code glitch by citing Neupane et al., 2019 (which reported the code glitch) or Willoughby et al., 2020 (which corrected the code glitch).

Q3: *Does the introduction section contain all the citations to Willoughby et al., 2014?*
A citation context in the introduction section is often used for introducing the research field or key background. Therefore, if all the Willoughby citation contexts are in the introduction section, then we presume that the publication does not apply the protocol to support its main contribution.

5.2 Approach-KW uses a keyword-based decision tree module for Stage 3

To handle citation contexts, chemistry domain knowledge could be used. From the citation contexts and full texts, author YF identified keywords that elevate (Table 1) or reduce (Table 2) the risk that a citing publication propagates unreliability. This resulted in two questions about the text of the citation contexts, used in Approach-KW:

Q4: *Does any citation context of Willoughby et al., 2014 contain one of the risk-elevating keywords in Table 1?*

Q5: *Do all citation contexts of Willoughby et al., 2014 contain one of the risk-reducing keywords in Table 2?*

The decision tree shown in Figure 6 uses binary questions from the citing publications' metadata, bibliography, and citation contexts from the full texts. The order of the nodes in the decision tree was determined by the feasibility of answering the questions.

The answers to questions determine the path through the decision tree to exactly one leaf node (labeled as "high risk", "medium risk", or "negligible risk" to propagate unreliability) and a corresponding rationale justifying the level of risk.

Using the decision tree in Figure 6, we triaged the citing publications of the protocol into 3 categories from most to least risk:

(1) The citing publication is **at high risk** of propagating unreliability
(2) The citing publication is **at medium risk** of propagating unreliability
(3) The citing publication is **at negligible risk** of propagating unreliability

5.3 Approach-ML uses a machine learning-based module for Stage 3

Approach-ML (Figure 7) swaps a machine-learning-based model into Stage 3 instead of the keyword-based citation context triage. We

Table 1: Risk-elevating keywords identified by YF's manual annotation

Keyword cluster	Keywords
Boltzmann weighting	Boltzmann
General citation	Quantum chemistry calculation; quantum chemical calculation; NMR chemical shift calculation; NMR chemical shift; NMR calculation; NMR shift prediction
Python script	Python script, script D; script

Table 2: Risk-reducing keywords identified by YF's manual annotation

Keyword cluster	Keywords
Geometry optimization	Density functional theory; DFT; geometry optimization; structural optimization; free energy calculation; structural optimization
Goodness-of-fit	MAE; CMAE; R^2; statistical error parameter
Conformational search	Conformational search; conformational space generation; conformational analysis; conformational sampling
Solvation model	PCM; solvation model
Scaling and referencing factors	Scaling; slope; intercept
Basis set	Basis set
GIAO method (an approximation)	GIAO

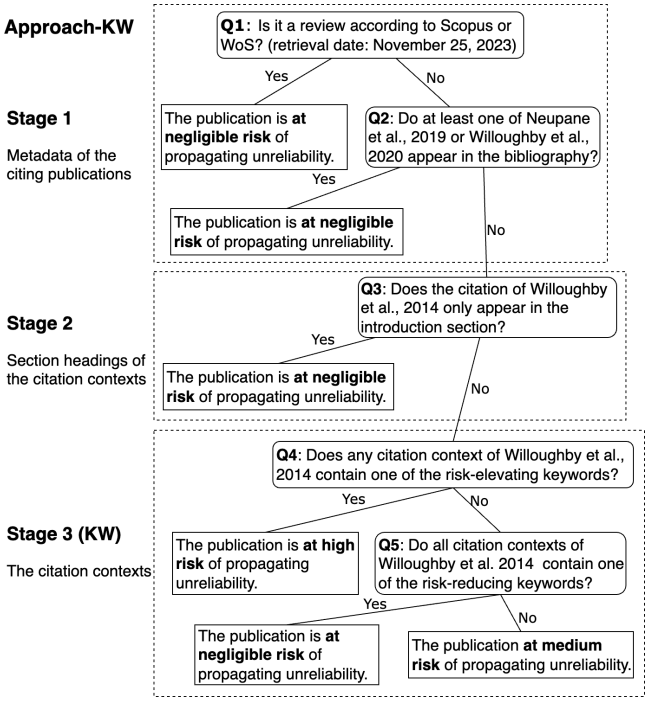

Figure 6: Risk assessment using Approach-KW

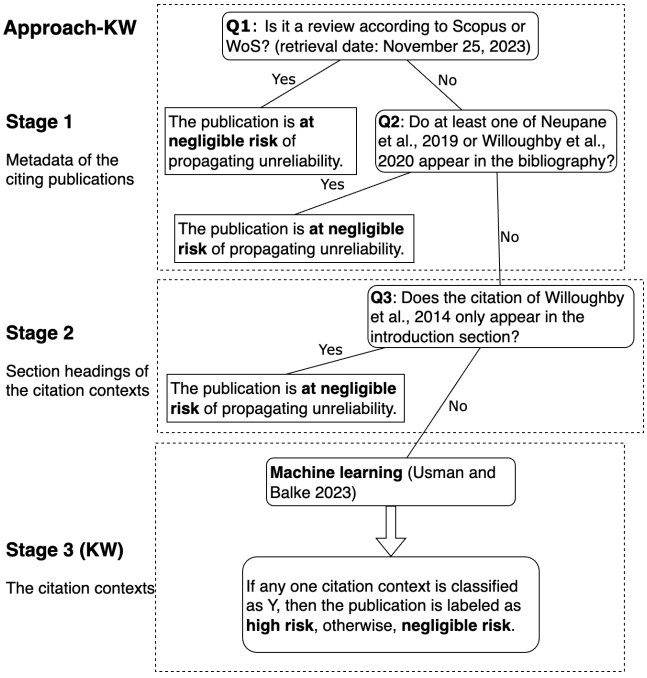

Figure 7: Risk assessment using Approach-ML

adapt and retrain publicly available classification code[4] for citation intention analysis [19]. We retrain with an 80%-20% train-test split on all 296 citation contexts from the 203 publications that were not handled in Stages 1 and 2, which the first chemistry expert had previously labeled as Y (at risk) or N (not at risk) purely based on the content of each citation context. The weighted average of precision is .70; the weighted average of recall is .70; and the F1-score is .70.[5]

To get the publication-level risk, we apply the model to all 296 citation contexts from the 203 publications to get their classifications, then interpret: If any one citation context in a publication is

[4]https://github.com/Conferences2023/TPDL

[5]The number of false negatives and false positives are the same, resulting in identical precision, recall, and F1.

classified as Y then the publication is labeled as high risk, otherwise, negligible risk.

5.4 Actions for different risk levels

Publications deemed at "high risk" are the most likely to propagate unreliability. They are likely to have used the erroneous Script D and reported incorrect chemical structures. Editors and authors must be alerted, and they need to work together to determine whether a publication's main conclusions still hold.

For publications deemed at "medium risk", the algorithms cannot rule out the propagation of unreliability, although these publications lack specific signatures of risk. Domain experts need to review these publications to determine whether they are "at risk" or "not at risk," as our chemistry experts did. Since this step requires extra expert labor, the number of publications deemed "medium risk" should be minimized if possible.

Publications deemed at "negligible risk" are unlikely to propagate unreliability. For instance, they may only cite the WJH protocol as background information, which has no direct impact on the validity of the conclusions. Authors do not need to take further action on publications deemed "negligible risk".

6 Results for RQ2: Triage performance and explainability

Next we address RQ2: *How well does each triage approach work and how explainable are the triage results?* We assess the performance of each approach using the requirements (Section 3.1). We treat the domain experts' manual annotation of the citing publications as a silver standard. Table 3 summarizes the triage performance of the three approaches introduced in Section 5.

6.1 Base approach (Stages 1 and 2)

The base approach (Stages 1 and 2) only triages out publications, determining that they are at "negligible risk".[6] Based on Section 5.4, these 203 publications are deemed of "medium risk." In reality, the 203 remaining publications would require expert review to determine their risk levels. For the purpose of evaluation, we regard the remaining 203 publications as predicted positive (e.g., at risk of propagating unreliability) for comparing the base approach against the expert annotations.

Do not miss any publications that propagate unreliability: Minimize false negatives, compared to experts. The base approach triaged out 81 publications and agreed with expert annotations on 96.6% (78/81 publications). The false negative rate is 3.5% (3/86).

Provide rationales: Explain the risk level. The base approach provided rationales for 81/284 (28.5%) publications triaged out in Stages 1 and 2. Rationales for publications at negligible risk are aimed at readers and would be displayed alongside indications that the publication was machine-checked:

(1) Human review is not needed because the publication cited at least one publication that either reported or corrected the code glitch.
(2) Human review is not needed because the publication is a review.
(3) Human review is not needed because there is no citation context of Willoughby et al., 2014 outside the introduction section.

Remove all publications that do not propagate unreliability: Minimize false positives, compared to experts. The base approach correctly detected 39.4% (78/198) publications the experts considered not at risk of propagating unreliability but yielded 60.6% (120/198) false positives.

Error analysis. Metadata and section headings are insufficient: to reduce false positives citation contexts are needed. Among false negatives: For one publication, Scopus classified it as a review but our experts think it is not a review. For two publications that cited the WJH protocol just once and only in the introduction section, we cannot rule out the possibility that Script D was used: in both cases the introduction section includes the goals of the publication (to make computational chemistry predictions of NMR) just before citing the WJH protocol.

6.2 Approach-KW

We combine categories "at high risk" and "at medium risk" from our approach to compare to the experts' "at risk" annotations.

Do not miss any publications that propagate unreliability: Minimize false negatives, compared to experts. Considering all stages, Approach-KW correctly detected 81.4% (70/86) publications that the experts considered at risk of propagating unreliability, and yielded 18.6% (16/86) false negatives. Only considering Stage 3, Approach-KW correctly detected 84.3% (70/83) publications that the experts considered at risk of propagating unreliability, and yielded 15.7% (13/83) false negatives.

Provide rationales: Explain the risk level. Approach-KW provided concrete rationales for all 284 (100%) publications, including 203 (71.4%) publications going into Stage 3. Rationales for high-risk publications would be sent to authors as a request to update the publication or certify that no update is needed:

> Please recheck the main contributions of [PUBLICATION TITLE AND LINK]. The unreliable Script D (Neupane et al., 2019; Willoughby et al., 2020) performs Boltzmann weighting. We found these words: [SPECIFIC KEYWORDS WE FOUND] in your citations to Willoughby et al., (2014), which increases the likelihood that you have used Script D, and the main contribution of your publication might be unreliable. For more information on our approach, please refer to the attached fact sheet.

Rationales for medium-risk publications would be sent to editors, to consider whether updates might be needed:

> The main contribution of [PUBLICATION TITLE AND LINK] might be unreliable due to unreliable Script D (Neupane et al., 2019; Willoughby et al., 2020) which performs Boltzmann weighting. We cannot rule out the possibility of unreliability because our validation approach could not decide based on the publication's keywords, bibliography,

[6] https://doi.org/10.5281/zenodo.14166498

Table 3: Triage performance of different approach

Approaches	False Negative Rate (%)	False Positive Rate (%)
Base approach (Stages 1 & 2)	3.5	60.6
Approach-KW (Stages 1, 2, & 3)	18.6	41.9
The keyword-based decision tree module (Q4 and Q5) in Approach-KW (Stage 3)	15.7	69.2
Approach-ML (Stages 1, 2, & 3)	17.4	10.6
The machine learning-based module in Approach-ML (Stage 3)	14.5	17.5

and article type. For more information on our approach, please refer to the attached fact sheet.

Rationales for publications at negligible risk are as shown previously in Section 6.1.

Remove all publications that do not propagate unreliability: Minimize false positives, compared to experts. Considering all stages, Approach-KW correctly detected 58.1% (115/198) publications the experts considered not at risk of propagating unreliability but yielded 41.9% (83/198) false positives. Only considering Stage 3, Approach-KW correctly detected 30.8% (37/120), and yielded 69.2% (83/120) false positives.

Error analysis. Besides errors inherited from Stages 1 and 2, which were analyzed in Section 6.1, the experts' manual annotation and our decision tree approach, Approach-KW, disagreed on 99/284 (34.9%) publications for 2 more reasons:

(1) Some keywords (e.g., "NMR calculation") are retained in the risk-elevating dictionary even though they generate false positives. This is the cause of the high false positive rate in Approach-KW.
(2) Domain knowledge is required to determine the risk level.

6.3 Approach-ML

Do not miss any publications that propagate unreliability: Minimize false negatives, compared to experts. Considering all stages, Approach-ML correctly detected 71/86 (82.6%) and missed 15/86 (17.4%) publications that the experts annotated as at risk of propagating unreliability, resulting in 17.4% (15/86) false negatives. Only considering Stage 3, Approach-ML correctly detected 85.5% (71/83) publications that the experts considered at risk of propagating unreliability, and yielded 14.5% (12/83) false negatives.

Provide rationales: Explain the risk level. Approach-ML can provide concrete rationales for 81/284 (28.5%) publications processed by Stages 1 and 2 (the base approach) but only provides a minimal rationale for Stage 3 publications: "a machine learning model determined the risk level".

Remove all publications that do not propagate unreliability: Minimize false positives, compared to experts. Considering all stages, Approach-ML identified 89.4% (177/198) publications that the experts annotated as not at risk of propagating unreliability, resulting in 10.6% (21/198) false positives. Only considering Stage 3, Approach-ML identified 82.5% (99/120) publications that the experts annotated as not at risk of propagating unreliability, resulting in 17.5% (21/120) false positives.

7 Results for RQ3: scientific digital library infrastructure

Yes, scientific digital library infrastructure is relevant to collecting some data relevant to the 3-stage strategy for determining the risk level (Figure 5):

Data 1 Retrieve and process publication type metadata to identify reviews for Stage 1
Data 2 Retrieve and process the bibliography for Stage 1
Data 3 Retrieve full text and extract citation contexts and their section headings for Stages 2 and 3

7.1 Retrieve and process the bibliography and publication type metadata

We next describe how we collected Data 1 and Data 2 to answer the Q1 and Q2 in Stage 1.

First, we retrieved the metadata of the citing publications from WoS and Scopus to answer Q1: *Is it a review according to Scopus or WoS?*

We retrieved the bibliography of the citing publications from WoS and Scopus to answer Q2: *Do at least one of Neupane et al., 2019 or Willoughby et al., 2020 appear in the bibliography?*

Data 3 requires a more detailed explanation given below.

7.2 Retrieve full text and extract citation contexts used in Stages 2 and 3

Next, we discuss collecting Data 3, which is used in Stages 2 and 3. We retrieved full texts in XML format using the Crossref Text and Data Mining (TDM) API. We developed a Python script that automatically extracts the citation contexts of Willoughby et al., 2014 from XML files.

Step 1. Find the bibliography entry of Willoughby et al., 2014. We first find all publications in the bibliography containing authors with the surname Willoughby (e.g., Listing 1).

Listing 1: A sample bibliographic reference XML element

```
<ce:bib–reference id="b0090">
    <ce:label>[18]</ce:label>
    <sb:reference id="h0085">
        <sb:contribution langtype="en">
            <sb:authors>
                <sb:author>
                    <ce:given–name>P.H.</ce:given–name>
                    <ce:surname>Willoughby</ce:surname>
                </sb:author>
```

```
            <sb:author>
                <ce:given-name>M.J.</ce:given-name>
                <ce:surname>Jansma</ce:surname>
            </sb:author>
            <sb:author>
                <ce:given-name>T.R.</ce:given-name>
                <ce:surname>Hoye</ce:surname>
            </sb:author>
        </sb:authors>
    </sb:contribution>
```

If we only find a single bibliography entry authored by Willoughby, we assume that it pertains to Willoughby et al., 2014. If we find multiple bibliography entries authored by a person with the surname Willoughby, we check the title of each bibliography entry. If the title contains the text 'guide to small-molecule structure assignment', then we select this bibliography entry. Once we find the Willoughby et al., 2014 reference XML element, we store its reference ids, i.e., the id attribute of both the `ce:bib-reference` and `sb:reference` XML elements.

Step 2. Locate all in-text citations (i.e., citation markers) of Willoughby et al., 2014. We traverse the XML document for in-text citations using the reference ids acquired in the previous step. In-text citations are denoted by the XML elements `ce:cross-ref` (for single citations, e.g., `'[3]'`) or `ce:cross-refs` (for multiple citations, e.g., `'[1-3]'`or `'[1,2,3]'`), as shown in Figure 8. If the `ref-id` attribute of the `ce:cross-ref` or `ce:cross-refs` element matches any of the reference ids, it means that the text refers to Willoughby et al., 2014.

```
<ce:para id="p0035" view="all">To confirm the structure of compound
<ce:bold>1</ce:bold> determined on the basis of HRMS and NMR spectroscopic data, NMR
calculation was carried out at mPW1PW91/6-311+G(2d,p) (giao,scrf) <ce:cross-ref
refid="b0090" id="c0095">[18]</ce:cross-ref>. The calculated NMR data (<ce:cross-ref
refid="t0005" id="c0100">Table 1</ce:cross-ref>) matched those of the experimental
results well, which further validated the structure determination.</ce:para>
```

Figure 8: Sample paragraph and in-text citation XML elements

Step 3. Extract the citation contexts. Paragraphs are indicated by the XML element `ce:para`. Starting from the citation marker (i.e., `ce:cross-ref` or `ce:cross-refs` element), we navigate upwards through the XML structure (i.e., repeatedly obtaining the parent element) until we encounter the `ce:para` element. Once we reach this element, we extract the text enclosed within it using the XPath `string()` function.

Results of the digital library automation. We developed a Python script [7] to automatically retrieve the full text of the citing publications using Crossref's TDM API service.[8]

Detailed results are shown in Figure 9. We retrieved 74/284 (26%) full texts in at least one of the following formats: plain text (63), PDF (11), and/or XML (63). For an additional 66/284 (23%) DOIs, the files retrieved were not usable full texts but returned only an error message with HTML saying "Just a moment" or a PDF displaying binary data. Those publications were not accessible via Crossref's TDM API. For 144/284 (51%) DOIs, our pipeline returned null results, that is, nothing was downloaded.

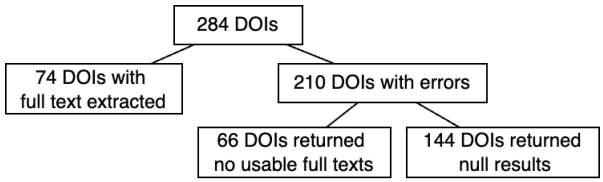

Figure 9: Results of extracting the full text of the 284 DOIs from publishers using our Python pipeline.

We automatically retrieved 63 publications with XML files, from which we were able to retrieve citation contexts for 61. For the 2 publications where the described process failed, one contained a bibliographic error (the author's given name and surname were H.W. and Patrick, respectively, instead of P.H. and Willoughby) and the other did not contain structured XML of the article text (i.e., the article text was enclosed in `<rawtext>` tags). A lack of full texts left 223/284 (79%) not processed. To make our triage approaches feasible to deploy in practice, the percentage of "not processed" publications should be significantly reduced.

The manual and automated extraction on these 61 publications differed little: 90 citation contexts[9] were manually extracted, compared to 89 automatically extracted. There were only four differences: Besides the one citation context missed, the automated method returned longer paragraphs for 3 citation contexts.

8 Discussion

We experimented with different approaches to identify the publications that propagate unreliability. Much of our analysis is specific to a single case: the citing publications of a computational chemistry protocol with a code glitch. Yet based on our experimentation we can analyze the advantages and limitations of the three approaches, as shown in Table 4.

Our design process demonstrates the importance of expert involvement. First, our system is impossible to build without experts. HZ, who is not an expert in chemistry, proposed Q1, Q2, and Q3 that triaged out 81/284 (28.5%) in Stages 1 and 2, leaving more than two-thirds of publications for manual review. The first expert (YF) generated Q4 and Q5 and associated keyword dictionaries to triage the remaining publications. The machine learning model also relies on experts' annotation of the citation contexts. Second, expert involvement aims to maximize authors' and editors' trust. Using the keyword approach, our expert can produce more specific rationales to persuade authors and editors to check and update publications.

Yet we expect the time commitment and roles for expert involvement and the choice between machine learning versus keyword approach to be situation dependent. Sometimes educated professionals without domain expertise can write rules to triage out most publications. In such a case, experts could verify the rules written by the non-expert and manually check the remaining publications, avoiding the need to analyze citation contexts (our Stage 3). When

[7] https://doi.org/10.5281/zenodo.14015039
[8] https://www.crossref.org/documentation/retrieve-metadata/rest-api/text-and-data-mining/
[9] Among the 61 publications, 14 cited the Willoughby publication two or more times.

Table 4: Advantages and limitations of different approaches to propagating unreliability. Case: 284 citing publications of a computational chemistry protocol with a code glitch

Approach	Involvement	Advantages	Limitations
Base approach (Stages 1 & 2)	Only non-experts	- No expert is needed	- High false positive rate - No rationales for 203/284 (71.5%) publications
Approach-KW (Stages 1, 2, & 3)	Experts + non-experts	- Concrete rationales for 284/284 (100%) publications	- High false positive rate - Experts cannot identify all possible keywords
Approach-ML (Stages 1, 2, & 3)	Experts + non-experts	- Low false positive rate - Low false negative rate	- Minimal rationales for 203/284 (71.5%) publications - Data labeling still requires expert labor

non-experts cannot triage most of the publications, experts are needed to design computational routines analogous to our Stage 3, which might use either a keyword approach or a machine learning approach. Machine learning requires a large amount of labeled data, which will require expert labor. Its lack of explainability is also a severe deficiency. Thus, the machine learning approach is likely to be suitable for high impact cases with hundreds or even thousands of citing publications to be triaged where experts do not have the confidence to craft keyword dictionaries even after reviewing a fraction of the citing publications. Future work should use multiple case studies to develop pragmatic advice for how to approach triage in different situations.

8.1 Limitations

We only studied one case of unreliability propagation in this paper. Our silver standard was not fully double-annotated: only 77/284 (27%) publications were annotated by both chemistry experts, while the remaining 207/284 (73%) publications were single-annotated then adjusted based on analysis of the disagreements with an adjudication policy crafted by JS with input from YF.

In order to systematically triage all publications citing the unreliable protocol, two challenges remain:

(1) A lack of full texts to automatically determine the questions in the decision tree;
(2) The decision tree cannot completely reflect human experts' decision-making process.

The Crossref TDM API from which we retrieved full texts does not retrieve Elsevier or Wiley publications because both publishers have their own licensed text and data mining APIs[10]. Metadata from the Crossref TDM API returns information in a JSON format that includes a tag containing the full text URL (when available) but it is up to the user to download the full text using the URL either manually or automatically. Metadata returned from the Crossref TDM API sometimes has PDF files tagged as "unspecified" instead of "pdf" (see Figure 10). The Python script we developed for full-text retrieval is programmed to download files tagged as "pdf", "xml", or "txt", not for "unspecified" files.

Currently, our digital library implementation requires XML format, yet 11 of 74 (15%) items were retrieved only in PDF format. For PDFs, other citation context extraction methods are needed.

[10]see https://www.elsevier.com/about/policies-and-standards/text-and-data-mining
https://onlinelibrary.wiley.com/library-info/resources/text-and-datamining
https://community.crossref.org/t/tdm-click-through-service/1533

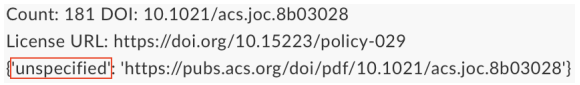

Figure 10: A citing publication with its PDF file tagged as "unspecified" in the metadata.

Our citation context extraction process has been tested on different publishers' XML files retrieved via Crossref. More variance could be expected in future collections of tagged fulltext due to different XML Document Type Definitions or different uses of the Journal Article Tag Suite (JATS) [8].

We could not readily identify which citing publications used the glitched versions of the Python script to support their main contribution, because these citations did not follow software citation standards [18]. We used the citation of the protocol as a proxy for the citation of the Python script with the code glitch. Further we extracted sentence-level citation contexts, although in practice a citation context can extend beyond the sentence containing the citation marker. Because authors could discuss the protocol without citing Willoughby et al., 2014 at all, citation context analysis cannot truly suffice to analyze the propagation of the unreliability due to the code glitch.

8.2 Future work

We will double annotate the remaining 207 publications to create a gold standard dataset. We will re-evaluate our approaches on the new dataset. We will also test whether our triage approaches work on publications beyond our corpus: the citing publications of Willoughby et al., 2014 that were published after July 6, 2023 when we collected our dataset. We will study more cases of unreliability propagation to improve these approaches.

We need more full text in computable formats (XML and EPUB), and better methods for automatically extracting citation contexts from citing publications, including identifying the corresponding section heading. When only PDF full text is available, we will test methods for converting PDF to XML using software such as GROBID [10]

In the future, we will develop a framework that could work for any input citation. We envision a future human-in-the-loop alert system that could send alerts to the authors and editors of citing publications deemed to be at high risk (see Figure 11), while publications at medium risk can be flagged for review by other domain experts for examination before alerting authors. Accurate alerts

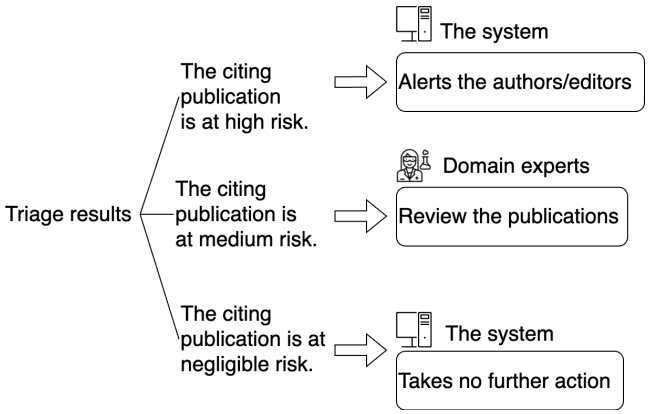

Figure 11: Future research: an envisioned alerting system to handle unreliability propagation

will be needed to avoid false positives that drive alert fatigue [23] for authors and editors. Future work must investigate under what conditions authors and editors are willing to check whether their publications propagate the unreliability.

Rationales targeting authors or editors would be customized to point out specific concerns, for instance:

> "Please recheck the main contributions of [PUBLICATION TITLE AND LINK]. The unreliable Script D (Neupane et al., 2019; Willoughby et al., 2020) performs Boltzmann weighting. We found these words: [SPECIFIC KEYWORDS WE FOUND] in your citations to Willoughby et al., (2014), which increases the likelihood that you have used Script D, and the main contribution of your publication might be unreliable. For more information on our approach, please refer to the attached fact sheet."

The RetractoBot trial [22], which is alerting citing authors about all past citations to retracted publications, provides an initial proof of concept that alerting authors is feasible. Even though RetractoBot emails do not request that authors check or update their publications, at least one paper has been corrected as a result of the project [26]. By reducing the number of alerts, an alerting system using triaged data would reduce alert fatigue compared to a system such as RetractoBot[11] whose intervention sends alerts to 100% of citing authors.

9 Conclusions

We investigated how to triage publications based on their risk level of propagating unreliability, starting with all publications citing a single computational chemistry protocol that had a code glitch.

We compared three triage approaches to a silver standard of manual annotations by two domain experts. The base approach, which relies on limited domain knowledge, is the most generalizable. Approach-KW, using a keyword-based decision tree on top of the base approach, provides concrete rationales, which could be important for convincing authors and editors to take action to update publications that propagate unreliability. Approach-ML, using machine learning based on citation intention on top of the base approach, has the best performance. While all three approaches are promising, none is in perfect agreement with the experts. The base approach missed 60.6% (120/198) of publications that two experts considered not at risk of propagating unreliability and incorrectly flagged 3.5% (3/86) of publications as not at risk of propagating unreliability. Approach-KW missed 41.9% (83/198) of publications that two experts considered not at risk of propagating unreliability, and incorrectly flagged 18.6% (16/86) of publications as not at risk of propagating unreliability. Approach-ML missed 10.6% (21/198) of publications two experts considered not at risk of propagating unreliability, and incorrectly flagged 17.4% (15/86) of publications as not at risk of propagating unreliability.

To improve the triage of citing publications, we would like to more fully capture the experts' decision processes, such as by identifying additional questions to use in our decision tree. We also need to explore how to better leverage scientific digital library infrastructure, to retrieve more full text, and to better extract citation contexts from full text. In the future, we will generalize our triage strategy by investigating other unreliable cited sources. We will incorporate this work into human-in-the-loop alerting that digital library maintainers, editors, and authors could use to triage publications that may propagate unreliability and maintain the reliability of scientific digital libraries.

Data and Code Availability

Our silver standard annotation with chemistry experts' annotations and human-extracted citation contexts is available at [28]. Our protocol for double-annotation is available at [27]. Our code for citation context extraction and citation clustering using similarity measures is at https://doi.org/10.5281/zenodo.13921537. Our code for triage approaches is at https://doi.org/10.5281/zenodo.14166498. Our code to automatically retrieve full text publications from the Crossref Text and Data Mining API is at https://doi.org/10.5281/zenodo.14015039.

Acknowledgments

Thanks to Muhammad Usman and Wolf-Tilo Balke for their publicly available citation intention code which we adapted in this work. Thanks to Malik Salami for merging citation records from Web of Science and Scopus; Tzu-Kun Hsiao for advice on citation context analysis; Janaynne Carvalho do Amaral for advice on scholarly communication industry peer review and editorial processing standards. Thanks also to Malik Salami and Hannah Smith for recommending relevant literature and feedback on a draft and to Stephen Downie, Dave Dubin, Daniel Evans, Michael Robert Gryk, Yuerong Hu, Dan Katz, Ted Ledford, Lan Li, Bertram Ludäscher, and Corinne McCumber for feedback on a draft.

Funding was provided by Alfred P. Sloan Foundation G-2022-19409 Reducing the Inadvertent Spread of Retracted Science II: Research and Development towards the Communication of Retractions, Removals, and Expressions of Concern and NSF 2046454 CAREER: Using network analysis to assess confidence in research synthesis. Ishita Sarraf was supported in part by the Distributed Research Experiences for Undergraduates (DREU) program, a joint project of the CRA Committee on the Status of Women in Computing Research (CRA - W) and the Coalition to Diversify Computing

[11] https://www.retracted.net

(CDC), which is funded in part by the NSF Broadening Participation in Computing program (NSF BPC - A #1246649). Jodi Schneider was supported in part as the 2024–2025 Perrin Moorhead Grayson and Bruns Grayson Fellow, Harvard Radcliffe Institute for Advanced Study.

CRediT

- Heng Zheng – Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Project administration, Software, Validation, Visualization, Writing – original draft, Writing – review & editing
- Yuanxi Fu – Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Validation, Writing – original draft, Writing – review & editing
- M. Janina Sarol – Software, Validation, Writing – original draft, Writing – review & editing
- Ishita Sarraf – Software, Validation, Writing – original draft, Writing – review & editing
- Jodi Schneider – Conceptualization, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Supervision, Writing – original draft, Writing – review & editing
- Ellie Vandel – Data curation, Formal analysis

References

[1] Daniel E. Acuña, Paul S. Brookes, and Konrad P. Kording. 2018. Bioscience-scale automated detection of figure element reuse [Preprint]. bioRxiv. https://doi.org/10.1101/269415

[2] Jayanti Bhandari Neupane, Ram P. Neupane, Yuheng Luo, Wesley Y. Yoshida, Rui Sun, and Philip G. Williams. 2019. Characterization of Leptazolines A–D, polar oxazolines from the cyanobacterium Leptolyngbya sp., reveals a glitch with the "Willoughby–Hoye" scripts for calculating NMR chemical shifts. *Organic Letters* 21, 20 (Oct. 2019), 8449–8453. https://doi.org/10.1021/acs.orglett.9b03216

[3] Amanda Capes-Davis, Amos Bairoch, Tanya Barrett, Edward C. Burnett, Wilhelm G. Dirks, Erin M. Hall, Lyn Healy, Douglas A. Kniss, Christopher Korch, Yuqin Liu, Richard M. Neve, Raymond W. Nims, Barbara Parodi, Rebecca E. Schweppe, Douglas R. Storts, and Fang Tian. 2019. Cell lines as biological models: practical steps for more reliable research. *Chemical Research in Toxicology* 32, 9 (Sept. 2019), 1733–1736. https://doi.org/10.1021/acs.chemrestox.9b00215

[4] Dalmeet Singh Chawla. 2021. New bot flags scientific studies that cite retracted papers. *Nature Index* (Feb. 2021). https://www.nature.com/nature-index/news/new-bot-flags-scientific-research-studies-that-cite-retracted-papers

[5] Daryl E. Chubin and Soumyo D. Moitra. 1975. Content analysis of references: Adjunct or alternative to citation counting? *Social Studies of Science* 5, 4 (Nov. 1975), 423–441. https://doi.org/10.1177/030631277500500403

[6] Yuanxi Fu and Jodi Schneider. 2020. Towards knowledge maintenance in scientific digital libraries with the keystone framework. In *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries in 2020 (JCDL '20)*. ACM, Virtual Event China, 217–226. https://doi.org/10.1145/3383583.3398514

[7] Maarten Grootendorst. 2022. BERTopic: Neural topic modeling with a class-based TF-IDF procedure [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2203.05794

[8] Tzu-Kun Hsiao and Vetle I. Torvik. 2023. OpCitance: Citation contexts identified from the PubMed Central open access articles. *Scientific Data* 10, 1 (April 2023), 243. https://doi.org/10.1038/s41597-023-02134-x

[9] J. Richard Landis and Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics* 33, 1 (1977), 159–174. https://doi.org/10.2307/2529310

[10] Mario Lipinski, Kevin Yao, Corinna Breitinger, Joeran Beel, and Bela Gipp. 2013. Evaluation of header metadata extraction approaches and tools for scientific PDF documents. In *Proceedings of the 13th ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL '13)*. Association for Computing Machinery, New York, NY, USA, 385–386. https://doi.org/10.1145/2467696.2467753

[11] Maribel O. Marcarino, María M. Zanardi, and Ariel M. Sarotti. 2020. The risks of automation: A study on DFT energy miscalculations and its consequences in NMR-based structural elucidation. *Organic Letters* 22, 9 (May 2020), 3561–3565. https://doi.org/10.1021/acs.orglett.0c01001

[12] Carol Anne Meyer. 2011. Distinguishing published scholarly content with CrossMark. *Learned Publishing* 24, 2 (2011), 87–93. https://doi.org/10.1087/20110202

[13] Gary Price. 2021. EndNote adds Retraction Watch notification integration, similar service available for Zotero and Papers. https://www.infodocket.com/2021/11/10/endnote-adds-retractionwatch-integration-similar-service-also-available-from-zotero/

[14] Jodi Schneider, Di Ye, Alison M. Hill, and Ashley S. Whitehorn. 2020. Continued post-retraction citation of a fraudulent clinical trial report, 11 years after it was retracted for falsifying data. *Scientometrics* 125, 3 (Dec. 2020), 2877–2913. https://doi.org/10.1007/s11192-020-03631-1

[15] Robert Schulz, Adrian Barnett, René Bernard, Nicholas J. L. Brown, Jennifer A. Byrne, Peter Eckmann, Małgorzata A. Gazda, Halil Kilicoglu, Eric M. Prager, Maia Salholz-Hillel, Gerben ter Riet, Timothy Vines, Colby J. Vorland, Han Zhuang, Anita Bandrowski, and Tracey L. Weissgerber. 2022. Is the future of peer review automated? *BMC Research Notes* 15, 203 (June 2022), 203. https://doi.org/10.1186/s13104-022-06080-6

[16] Efrat Shimron, Jonathan I. Tamir, Ke Wang, and Michael Lustig. 2022. Implicit data crimes: Machine learning bias arising from misuse of public data. *Proceedings of the National Academy of Sciences* 119, 13 (2022), e2117203119. https://doi.org/10.1073/pnas.2117203119

[17] Henry Small. 1982. Citation context analysis. In *Progress in Communication Sciences*. Vol. 3. Ablex Publishing Corporation, Norwood, NJ, 287–310.

[18] Arfon M. Smith, Daniel S. Katz, and Kyle E. Niemeyer. 2016. Software citation principles. *PeerJ Computer Science* 2 (Sept. 2016), e86. https://doi.org/10.7717/peerj-cs.86

[19] Muhammad Usman and Wolf-Tilo Balke. 2023. On retraction cascade? Citation intention analysis as a quality control mechanism in digital libraries. In *Linking Theory and Practice of Digital Libraries (Lecture Notes in Computer Science)*. Springer Nature Switzerland, Cham, 117–131. https://doi.org/10.1007/978-3-031-43849-3_11

[20] Paul E. van der Vet and Harm Nijveen. 2016. Propagation of errors in citation networks: A study involving the entire citation network of a widely cited paper published in, and later retracted from, the journal Nature. *Research Integrity and Peer Review* 1 (Dec. 2016), 3. https://doi.org/10.1186/s41073-016-0008-5

[21] Peiling Wang. 2023. Rising of retracted research works and challenges in information systems: Need new features for information retrieval and interactions. In *Proceedings of the 2023 Conference on Human Information Interaction and Retrieval*. ACM, Austin TX USA, 69–82. https://doi.org/10.1145/3576840.3578281

[22] Karolina Wartolowska, Francis Irving, Nicholas DeVito, Benjamin Feakins, Helen Curtis, Seb Bacon, Liam Smeeth, Carl Heneghan, and Ben Goldacre. 2023. RetractoBot: a protocol for a randomised controlled trial to assess the impact of notifying authors that they have cited a retracted paper [Revised Stage 1 Registered Report Protocol (Oct 2023)]. (October 2023). https://doi.org/10.6084/m9.figshare.24468391.v1 Protocol for a BMC Medicine Registered Report.

[23] Christopher D. Wickens, Justin G. Hollands, Simon Banbury, and Raja Parasuraman. 2012. Alarm and alert systems [Section 5.3]. In *Engineering Psychology and Human Performance*. Taylor & Francis Group, London, 23–27.

[24] Patrick H. Willoughby, Matthew J. Jansma, and Thomas R. Hoye. 2014. A guide to small-molecule structure assignment through computation of (^{1}H and ^{13}C) NMR chemical shifts. *Nature Protocols* 9, 3 (March 2014), 643–660. https://doi.org/10.1038/nprot.2014.042

[25] Patrick H. Willoughby, Matthew J. Jansma, and Thomas R. Hoye. 2020. Addendum: A guide to small-molecule structure assignment through computation of (^{1}H and ^{13}C) NMR chemical shifts. *Nature Protocols* 15, 7 (July 2020), 2277. https://doi.org/10.1038/s41596-020-0293-9

[26] Inbal L. Zak, Santosh C. Gadekar, and Anat Milo. 2024. Erratum - Designing the Secondary Coordination Sphere in Small-Molecule Catalysis. *Synlett* 35, 07 (April 2024), 832. https://doi.org/10.1055/s-0043-1763761

[27] Heng Zheng, Yuanxi Fu, M. Janina Sarol, and Jodi Schneider. 2024. PROTOCOL: Annotation by chemistry expert #2 for Addressing Unreliability Propagation in Scientific Digital Libraries [Protocol]. https://hdl.handle.net/2142/123380

[28] Heng Zheng, Yuanxi Fu, Ellie Vandel, and Jodi Schneider. 2024. Dataset of 286 publications citing the 2014 Willoughby-Jansma-Hoye protocol [Dataset]. https://doi.org/10.13012/B2IDB-4610831_V3

[29] Hongmei Zhu, Yongliang Jia, and Siu-wai Leung. 2024. Citations of microRNA biomarker articles that were retracted: A systematic review. *JAMA Network Open* 7, 3 (March 2024), e243173. https://doi.org/10.1001/jamanetworkopen.2024.3173

Received 10 August 2024; accepted 25 September 2024; revised 7 November 2024