How does a piece of unreliable code impact its citing publications? An argumentation analysis

HENG ZHENG, YUANXI FU, JODI SCHNEIDER
School of Information Sciences
University of Illinois Urbana-Champaign
501 E. Daniel St., Champaign, IL
USA
zhenghz@illinois.edu
fu5@illinois.edu, jodi@illinois.edu

Abstract: Argumentation theory can be applied in natural language contexts, such as scholarly publications. In this paper, we explore the support relationships between scholarly publications. We demonstrate how defeasible reasoning can be used to solve a practical problem: determining how a piece of unreliable code impacts its citing publications. After constructing a defeasible argument, we repurpose the exceptions from the defeasible argument as questions in a decision tree. We show two examples of citation contexts that can help us decide whether a citing publication propagates the unreliability of an unreliable source. This illustrates a practical application of argumentation theory.

Keywords: scientific publications, arguments, citation contexts, defeasible reasoning, argumentation in natural language, unreliability propagation

1. Introduction and Background

Argumentation theory can be applied in natural language contexts, such as legal documents (Prakken & Sartor, 2015), public communication of science (Oswald et al., 2022), and scholarly publications (Green 2018; Mayer et al., 2018; Wang et al., 2022). In this paper, we explore the support relationships between scholarly publications. We demonstrate how defeasible reasoning can be used to solve a practical problem: the impact of citing an unreliable computational protocol, namely, determining how unreliability propagates through citations. We say that a publication propagates unreliability when the main contribution of the publication becomes unreliable by using an unreliable source.

First, we explain terminology that may not be in the *lingua franca* of argumentation researchers. A citation context is the part of a citing publication that mentions the citation, which can vary in size from a few sentences to a paragraph to a section. Figure 1 shows an example of a citation context in a citing publication. The publication "Fu & Schneider 2020" uses the citation marker "[16]" to refer to the *Handbook of Argumentation Theory* by van Eemeren et al. (2014). In general, a citation marker may appear multiple times in the paper, each with its own citation context.

Argumentation has been used to analyze scholarly publications. Wang et al. (2022) compared the rhetorical structures of scientific publications from two domains: biomedicine and library and information science. Schneider (2023) distinguished three different kinds of arguments in empirical biomedical publications: Rhetorical moves build up the backbone of a publication, domain-specific arguments establish the inferential structure of the research presented in a publication, and citations support statements in a publication. Argument schemes have been used to extract arguments from scientific publications (Green, 2018) including clinical trial reports (Mayer et al., 2018). Mizrahi & Dickinson (2020) studied different types of arguments (deductive, abductive, and inductive) in philosophical papers.

2 RELATED WORK

2.1 Argumentation-based Curation

Argumentation theory is an interdisciplinary field with multiple branches studying persuasion, rhetoric, dialectic, defeasible reasoning, and related topics [16]. Argument schemes describe

[16] Eemeren, F.H. van, Garssen, B., Krabbe, E.C.W., Snoeck Henkemans, A.F., Verheij, B. and Wagemans, J.H.M. 2014. *Handbook of argumentation theory.* Springer Reference.

Figure 1. The publication "Fu & Schneider 2020" uses the citation marker "[16]" to refer to the *Handbook of Argumentation Theory* by van Eemeren et al. (2014)

An important application of argumentation theory is to trace unreliability propagation in scientific literature through citations. Clark et al. (2014) introduced an argumentation-based document model called micropublications for biomedical publications[1] based on Toulmin's model (Toulmin, 1958/2003). Fu & Schneider (2020) developed the Keystone Framework, which combines citation context analysis with argumentation-based document models. Under this framework, keystone citations are citations whose validity can impact the validity of the citing paper. They carried out two case studies using the micropublication model: one determined the impact of citing a computational chemistry protocol (Willoughby et al., 2014) in a small sample (more details in the next section), and the other identified all keystone citations in a single biomedical publication.

Here we develop a procedure to scale the case study of the computational chemistry protocol from Fu & Schneider (2020) so that, in the future, similar assessments of citing unreliable scholarly resources (e.g., publications, protocols, software, data) can be carried out on a large number of documents. Our procedure is based on argumentation theory. We explain the case of the computational chemistry protocol code glitch and the procedure we developed in the following sections.

2. Background of the Computational Chemistry Protocol Case Study and Pilot Research

Our case study is centered around the computational chemistry protocol introduced by Willoughby et al., 2014, which became known to be unreliable in 2019 when a code glitch was determined to impact some operations in the protocol, as reported by Neupane et al. (2019). In response, the code glitch was fixed in 2020 by an addendum (Willoughby et al., 2020). The code glitch impacts only one part of the protocol, Script D, which is used in a procedure called Boltzmann analysis. Because the protocol is modular, although it has multiple steps, not all steps are necessarily used by a paper citing the protocol. Some papers might cite the protocol without using it at all—for instance, to mention the influence of computational chemistry approaches in modern chemistry research.

The pilot case study (Fu & Schneider, 2020) found that only some citing publications applied unreliable steps in the computational chemistry protocol. A protocol in the context of scientific research refers to a written document reporting the experimental procedure. Protocols help scientists ensure methodological rigor across different labs and are prevalent in disciplines where experiments require lengthy and complex procedures, such as biology. The particular protocol we study here, the protocol introduced by Willoughby et al., 2014, was intended for bench

[1] This model, although developed for biomedical publications, can also be used for empirical research papers such as those in chemistry.

chemists with little computational chemistry knowledge. It teaches them how to compute theoretical Nuclear Magnetic Resonance (NMR) spectra, which can be compared to the NMR spectra obtained from experiments to determine the structure of newly isolated organic compounds.

3. Determining the impact of citing the protocol

For this project, we developed a different procedure than the one used by Fu & Schneider (2020), which was only suitable for domain experts. The current procedure aims to enable a non-expert or computer to perform the assessment.

First, we notice that when we use a citation context to determine the impact of citing the protocol in a collection of citing publications (Zheng & Fu, 2024), we follow a pattern that resembles defeasible reasoning. As we discussed above, authors of a citing publication may not use the protocol to support the main contribution of their publication. The authors may have applied other steps in the protocol rather than Boltzmann analysis. Even when they performed Boltzmann analysis, authors may have used other tools instead of Script D. Since the code glitch only impacts Script D, if a citing publication did not use Script D, then the main contribution in the publication is not at risk of propagating the unreliability. Also, when a citing publication also cited Neupane et al., 2019 (which reported the code glitch) or Willoughby et al., 2020 (which corrected the code glitch), we assume the authors of the citing publication are aware of the code glitch because these publications either reported or corrected the code glitch.

We thus distilled a defeasible argument (Figure 2). The exceptions are undercutting defeaters of the argument. Some defeaters can be identified by non-experts, such as citing the Neupane paper or the addendum, which demonstrate the authors' awareness of the code glitch, and, therefore, presumably rule out the possibility that they would repeat the mistake. Yet often, identifying defeaters requires domain knowledge, for example, citing publications that used the protocol to support the main contribution. Therefore, constructing something like Figure 2 requires that either (1) domain experts sample a significant set of the citing publications to identify the defeaters or (2) a person familiar with the paper provides a list of the defeaters that they can think of. Option (1) is more labor-intensive but is likely more rigorous. Option (2) is useful when time is of the essence, yet it is limited by one person's view and knowledge.

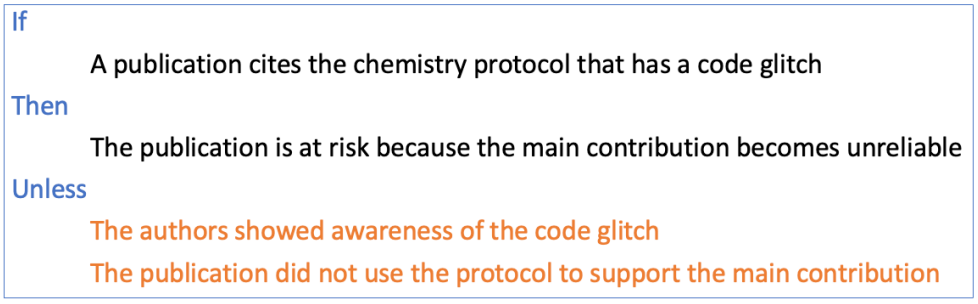

Figure 2. A defeasible argument for determining whether the publication is at risk because the main contribution becomes unreliable.

Exceptions in the defeasible argument from Figure 2 become questions in the decision tree shown in Figure 3:

- Question 1: Did the authors show awareness of the code glitch?

- Question 2: Did the citing publication use the protocol to support its main contribution?

We then used the resulting decision tree (Figure 3) to determine whether the code glitch potentially impacts the main contribution in a citing publication.

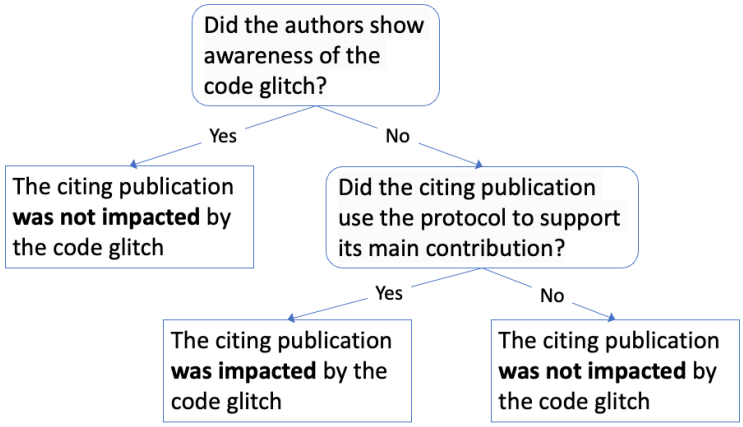

Figure 3. Determining whether a citing publication was impacted by the code glitch.

Beyond determining whether a citing paper is "impacted" (and its main contribution is presumed to be invalid) or "not impacted" (and its main contribution is presumed to be valid), we want to specify the justification for "the citing paper is presumed to be reliable", based on the path in the decision tree:

- Scenario 1: Based on the citation contexts referring to the protocol, the authors of the citing publication showed awareness of the code glitch.
- Scenario 2: Based on the citation contexts referring to the protocol, the authors of the citing publication did not show awareness of the code glitch, but the publication did not use the protocol to support its main contribution.
- Scenario 3: Based on the citation contexts referring to the protocol, the authors of the citing publication did not show awareness of the code glitch, and the publication used the protocol to support its main contribution.

In the first two scenarios, the citing publication is not impacted. Only in Scenario 3 the validity of the main contribution was impacted.

4. Distinguishing the scenarios using keywords, phrases, and the bibliography

Only the citing publications categorized as Scenario 3 are impacted by the code glitch. We use a list of words and phrases to determine which scenario a publication belongs to, based on its citation contexts: each paragraph contains a citation marker referring to the protocol. We identified the keywords and phrases from the text of Willoughby et al., 2014 describing the steps related to Boltzmann analysis, such as "Operation IV: Boltzmann-weighting of shielding tensors and conversion to chemical shifts" and "Assemble and Boltzmann-average the NMR and free energy data (using Script D)".

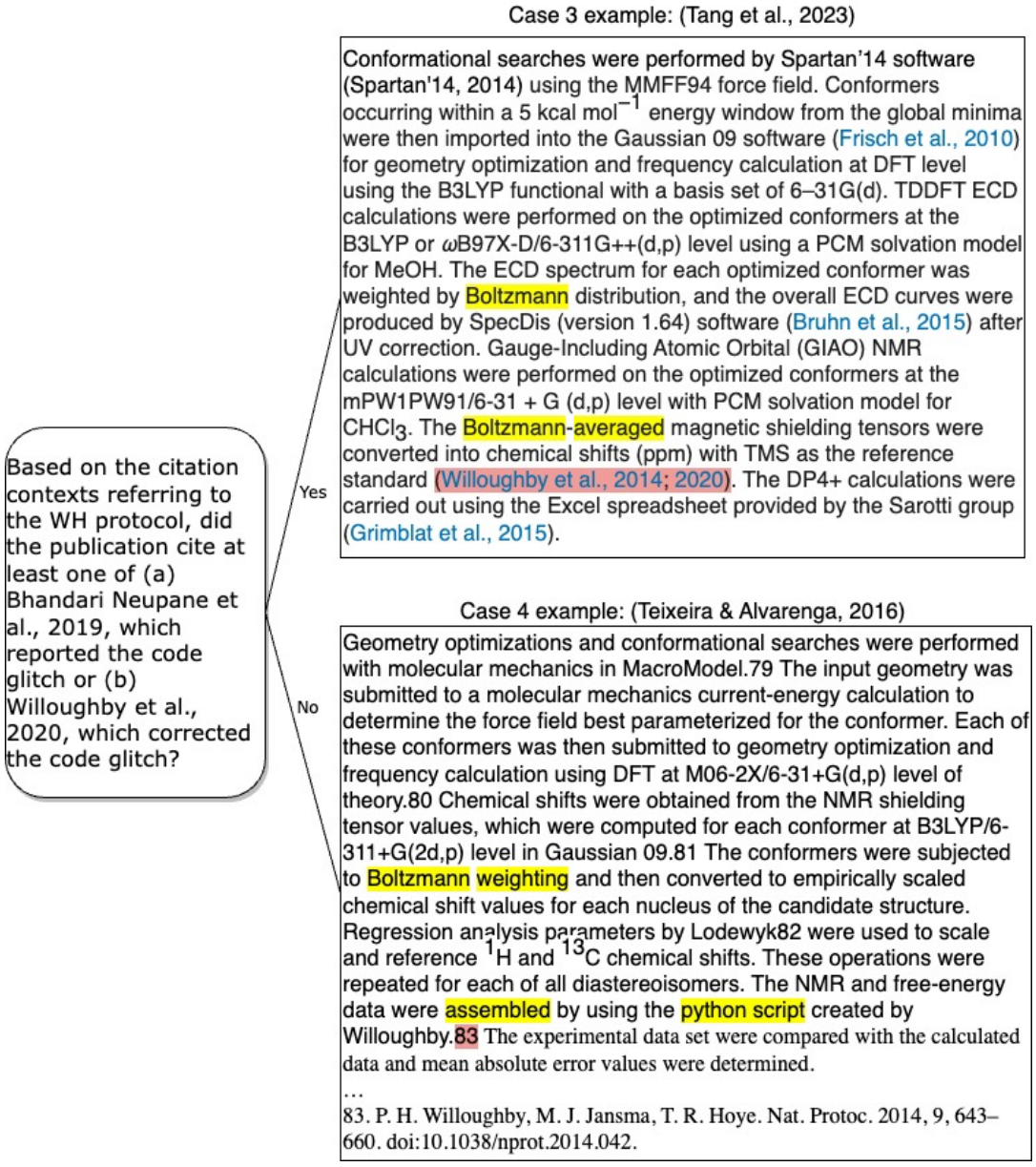

Figure 4. Examples of citation contexts from Tang et al., 2023 (top) and Teixeira & Alvarenga, 2016 (bottom) help us decide whether a citing publication was impacted by the code glitch. Citation markers are highlighted in red. Yellow highlights show the keywords we used to decide which scenario the citation context belongs to.

Figure 4 shows two examples of citation contexts that can help us decide whether a citing publication was impacted by the code glitch. First, in Tang et al., 2023, even though we see keywords such as "Boltzmann" and "Boltzmann-averaged" that suggest the publication used the protocol to support its main contribution, the citation marker "(Willoughby et al., 2014;2020)" indicates that the authors cited the addendum Willoughby et al., 2020 which fixed the code glitch. Therefore, the answer to Question 1 is "Yes" and we categorize Tang et al., 2023 as Scenario 1, meaning that their main contribution is presumably not be impacted by the code glitch.

Second, in Teixeira & Alvarenga, 2016, the citing publication did not cite Neupane et al., 2019 or Willoughby et al., 2020, so the answer to Question 1 is “No”, and we can exclude Scenario 1. Keywords and phrases such as “Boltzmann weighting” and “assembled” lead us to presume that the publication used the protocol to support its main contribution ("Yes" to Question 2), so we categorize the publication as Scenario 3, and its main contribution may be unreliable because of the code glitch.

5. Discussion

We used defeasible reasoning to model exceptions in deciding whether the main contribution in a publication may be unreliable. We used a decision tree to distinguish how the citing publications used the protocol. Sometimes, defeaters can be identified by non-experts, such as citing the Neupane paper or the addendum. But often, identifying defeaters requires domain knowledge, for example, the authors use the protocol to support the main contribution. Our procedure relied on a domain expert (the second author, YF) to identify defeaters. Our defeasible reasoning analysis was then used to construct a decision tree with which a non-expert (first author, HZ) performed the analysis. Addepalli et al. (2022) provides a prior example of how non-professionals developed and used a decision tree to determine the impact of citing a retracted medical article. In ongoing work, we are testing an automatic process for categorizing the publications citing the protocol into the scenarios we identified in this paper (Zheng et al, under review).

For the empirical science community, identifying when an unreliable publication can impact a citing publication helps understand how the unreliability of information may spread. For the argumentation community, our case study demonstrates how argumentation theory can be applied to solve a real-world issue in scholarly publications: the impact of citing an unreliable publication.

Using defeasible reasoning to describe possible worlds has been fruitful. To model crime investigations, Bex & Verheij (2012) used hypothetical stories and generalized abstract story schemes such as “beginning—middle—end”. Bex and Verheij also proposed a list of critical questions to assist in decision-making during an investigation. The questions we used to construct the decision tree can be considered critical questions in an argumentation scheme. Using our experience of discovering exceptions, we will systematically design a list of critical questions to help people analyze the impact of other cases of unreliable code on their citing publications.

The formal argumentation community may be interested in examining how different uses of citations impact the validity of a citing paper. Verheij distinguishes three kinds of argument validity strengths ranging from strong to weak (Verheij, 2018). We are more confident about the inference made by certain words and phrases: a citation context containing “Boltzmann analysis” and “Python script” more likely reflects the use of Script D for Boltzmann analysis (and hence impact on the citing paper) than a citation context only mentioning “Boltzmann analysis”.

6. Conclusions

Our case study applies argumentation theory to empirical scientific publications. We used defeasible reasoning to determine whether citing papers are impacted by a code glitch in a computational chemistry protocol, which enabled us to construct a decision tree operationalizing the defeasible reasoning process. We analyzed under which conditions the code glitch may impact the main contribution of a citing publication. While the construction of the decision tree requires

domain expertise and analysis of the reasoning structure, the decision tree can be used by non-experts. In future work, we will test an automatic process for applying the decision tree.

Data availability: A bibliography with the 286 citing publications of Willoughby et al., 2014 is available in the following dataset:

Zheng, Heng; Fu, Yuanxi (2024): Dataset of 286 publications citing the 2014 Willoughby-Jansma-Hoye protocol. University of Illinois at Urbana-Champaign Databank. https://doi.org/10.13012/B2IDB-4610831_V2

Acknowledgements: 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. We thank Malik Salami for helping us prepare the dataset of citing papers. We thank Frank Zenker, Liliana Giusti Serra, Shiyang Yu, and Lev Frank for providing feedback on a draft. We thank Fabio Paglieri for his commentary on this paper.

CRediT:

- Heng Zheng: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Project administration, Validation, Writing – original draft (lead), Writing – review & editing
- Yuanxi Fu: Conceptualization, Data curation, Funding acquisition, Investigation, Methodology, Validation, Writing – original draft, Writing – review & editing
- Jodi Schneider: Conceptualization, Funding acquisition, Project administration, Supervision, Writing – original draft, Writing – review & editing

References

Addepalli, A., Subin, K. A., & Schneider, J. (2022). Testing the keystone framework by analyzing positive citations to Wakefield's 1998 paper. In M. Smits (Ed.), *Information for a Better World: Shaping the Global Future* (pp. 79–88). Springer International Publishing. https://doi.org/10.1007/978-3-030-96957-8_9

Bex, F., & Verheij, B. (2012). Solving a murder case by asking critical questions: an approach to fact-finding in terms of argumentation and story schemes. *Argumentation*, *26*(3), 325–353. https://doi.org/10.1007/s10503-011-9257-0

Bhandari Neupane, J., Neupane, R. P., Luo, Y., Yoshida, W. Y., Sun, R., & Williams, P. G. (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), 8449–8453. https://doi.org/10.1021/acs.orglett.9b03216

Clark, T., Ciccarese, P. N., & Goble, C. A. (2014). Micropublications: A semantic model for claims, evidence, arguments and annotations in biomedical communications. *Journal of Biomedical Semantics*, *5*(1), 28. https://doi.org/10.1186/2041-1480-5-28

Fu, Y., & Schneider, J. (2020). Towards knowledge maintenance in scientific digital libraries with the keystone framework. *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries in 2020*, 217–226. https://doi.org/10.1145/3383583.3398514

Green, N. L. (2018). Towards mining scientific discourse using argumentation schemes. *Argument & Computation*, *9*(2), 121–135. https://doi.org/10.3233/AAC-180038

Mayer, T., Cabrio, E., Lippi, M., Torroni, P., & Villata, S. (2018). Argument mining on clinical trials. *Computational Models of Argument*, 137–148. https://doi.org/10.3233/978-1-61499-906-5-137

Mizrahi, M., & Dickinson, M. (2020). Argumentation in philosophical practice: An empirical study. *OSSA Conference Archive. 3.* https://scholar.uwindsor.ca/ossaarchive/OSSA12/Saturday/3

Oswald, S., Lewiński, M., Greco, S., & Villata, S. (Eds.). (2022). *The Pandemic of Argumentation* (Vol. 43). Springer International Publishing. https://doi.org/10.1007/978-3-030-91017-4

Prakken, H., & Sartor, G. (2015). Law and logic: A review from an argumentation perspective. *Artificial Intelligence*, *227*, 214–245. https://doi.org/10.1016/j.artint.2015.06.005

Schneider, J. (2023). How do empirical biomedical research articles argue? Examining the layers of rhetorical, domain-specific, and citation-based argumentation. *10th Conference of the International Society for the Study of Argumentation*. https://jodischneider.com/pubs/issa2023layers.pdf

Tang, S.-Y., Tan, C.-H., Sim, K.-S., Yong, K.-T., Lim, K.-H., Low, Y.-Y., & Lim, S.-H. (2023). Polyneurines A−H, iboga alkaloids from Tabernaemontana polyneura. *Phytochemistry*, *208*, 113587. https://doi.org/10.1016/j.phytochem.2023.113587

Teixeira, M. G., & Alvarenga, E. S. (2016). Characterization of novel isobenzofuranones by DFT calculations and 2D NMR analysis. *Magnetic Resonance in Chemistry*, *54*(8), 623–631. https://doi.org/10.1002/mrc.4411

Toulmin, S. E. (1958/2003). *The Uses of Argument* (2nd ed.). Cambridge University Press. https://doi.org/10.1017/CBO9780511840005

van Eemeren, F. H., Garssen, B., Krabbe, E. C. W., Snoeck Henkemans, A. F., Verheij, B., & Wagemans, J. H. M. (2014). *Handbook of Argumentation Theory*. Springer. https://doi.org/10.1007/978-90-481-9473-5

Wang, X., Song, N., Zhou, H., & Cheng, H. (2022). The representation of argumentation in scientific papers: A comparative analysis of two research areas. *Journal of the Association for Information Science and Technology*, *73*(6), 863–878. https://doi.org/10.1002/asi.24590

Willoughby, P. H., Jansma, M. J., & Hoye, T. R. (2014). A guide to small-molecule structure assignment through computation of (^{1}H and ^{13}C) NMR chemical shifts. *Nature Protocols*, *9*(3), Article 3. https://doi.org/10.1038/nprot.2014.042

Willoughby, P. H., Jansma, M. J., & Hoye, T. R. (2020). Addendum: A guide to small-molecule structure assignment through computation of (^{1}H and ^{13}C) NMR chemical shifts. *Nature Protocols*, *15*(7), Article 7. https://doi.org/10.1038/s41596-020-0293-9

Zheng, H., Fu, Y. Sarol, M. J., Sarraf, I., & Schneider, J. (2024). Addressing unreliability propagation in scientific digital libraries. Under review.

Zheng, H., Fu, Y. (2024). Dataset of 286 publications citing the 2014 Willoughby-Jansma-Hoye protocol. University of Illinois at Urbana-Champaign Data Bank. https://doi.org/10.13012/B2IDB-4610831_V2