

# Reviewed Preprint

v1 • January 19, 2026

Not revised

## For correspondence:

jo.lemaitresamra@gmail.com

bruno.lemaitre@epfl.ch

## Competing interests:

No competing interests declared

Funding: See page 25

Reviewing editor: Peter Rodgers, eLife, United Kingdom

This is an open-access article, free of all copyright, and may be freely reproduced, distributed, transmitted, modified, built upon, or otherwise used by anyone for any lawful purpose. The work is made available under the [Creative Commons CC0 public domain dedication](#).

# A retrospective analysis of 400 publications reveals patterns of irreproducibility across an entire life sciences research field

Joseph Lemaitre<sup>1</sup>✉, Désirée Popelka<sup>2</sup>, Blandine Ribotta<sup>2</sup>, Hannah Westlake<sup>2</sup>, Sveta Chakrabarti<sup>3</sup>, Li Xiaoxue<sup>4</sup>, Mark A Hanson<sup>5</sup>, Haobo Jiang<sup>6</sup>, Francesca Di Cara<sup>7</sup>, Estee Kurant<sup>8</sup>, Fabrice David<sup>9</sup>, Bruno Lemaitre<sup>2</sup>✉

<sup>1</sup>Department of Epidemiology, Gillings School of Global Public Health, University of North Carolina at Chapel Hill, Chapel Hill, United States • <sup>2</sup>Global Health Institute, School of life science, EPFL, Lausanne, Switzerland • <sup>3</sup>Manipal Institute of Regenerative Medicine, Bengaluru, Manipal Academy of Higher Education, Manipal, India • <sup>4</sup>College of Plant Science and Technology, Huazhong Agricultural University, Wuhan, China • <sup>5</sup>Centre for Ecology and Conservation, University of Exeter, Penryn Campus, Penryn, United Kingdom • <sup>6</sup>Department of Entomology and Plant Pathology, Oklahoma State University, Stillwater, United States • <sup>7</sup>Dalhousie University, Department of Microbiology and Immunology, Halifax, Canada • <sup>8</sup>Department of Human Biology, Faculty of Natural Sciences, University of Haifa, Haifa, Israel • <sup>9</sup>BioInformatics Competence Center of EPFL-UNIL, School of Life Sciences, EPFL, Lausanne, Switzerland

## eLife Assessment

This **important** study presents an impressive large-scale effort to assess the reproducibility of published findings in the field of *Drosophila* immunity. The authors analyse 400 papers published between 1959 and 2011, and assess how many of the claims in these papers have been tested in subsequent publications. In a companion article they report the results of experiments to test a subset of the claims that, according to the literature, have not been tested. The present article also explores if various factors related to authors, institutions and journals influence reproducibility in this field. The evidence supporting the claims is **solid**, but there is considerable scope for strengthening and extending the analysis. The limitations inherent to evaluating reproducibility based on the published literature should also be acknowledged.

<https://doi.org/10.7554/eLife.108403.1.sa3>

## Abstract

The *ReproSci* project retrospectively analyzed the reproducibility of 1006 claims from 400 papers published between 1959 and 2011 in the field of *Drosophila* immunity. This project attempts to provide a comprehensive assessment, 14 years later, of the replicability of nearly all publications across an entire scientific community in experimental life sciences. We found that 61% of claims were verified, while only 7% were directly challenged (not reproducible), a replicability rate higher than previous assessments. Notably, 24% of claims had never been independently tested and remain unchallenged. We performed experimental validations of a selection of 45 unchallenged claim, that revealed that a significant fraction (38/45) of them is in fact non-reproducible. We also found that high-impact journals and top-ranked institutions are more likely to publish challenged claims. In line with the reproducibility crisis narrative, the rates of both challenged and unchallenged claims increased over time, especially as the field gained popularity. We characterized the uneven distribution of irreproducibility among first and last authors. Surprisingly, irreproducibility rates were similar between PhD students and postdocs, and did not

decrease with experience or publication count. However, group leaders, who had prior experience as first authors in another *Drosophila* immunity team, had lower irreproducibility rates, underscoring the importance of early-career training. Finally, authors with a more exploratory, short-term engagement with the field exhibited slightly higher rates of challenged claims and a markedly higher proportion of unchallenged ones. This systematic, field-wide retrospective study offers meaningful insights into the ongoing discussion on reproducibility in experimental life sciences.

## Introduction

Science predominantly thrives on the continuous aggregation of knowledge, with researchers progressively building upon one another's discoveries. As results are replicated (when independent studies reach the same conclusions as the original authors of a claim), confidence in the evidence supporting the claim grows. Confidence in publishing results is critical to sustain trust within a scientific community and to allow new research to build on previous findings (Fidler and Wilcox, 2018). Over the past decades, this confidence has been on a decline.

Several recent reports point out that reproducibility is lower than desirable in molecular life sciences (Amaral et al., 2025; Begley and Ioannidis, 2015; Fidler et al., 2017; Ioannidis et al., 2009; Macleod et al., 2014). In 2016, an extensive survey indicated that more than 70% of researchers have tried and failed to reproduce another scientist's experiments, and more than half have failed to reproduce their own experiments (Baker, 2016). Another study suggested that about 85% of biomedical research is 'wasted', meaning that inefficiencies and methodological flaws throughout the research process undermine the value of scientific findings (Begley and Ioannidis, 2015; Chalmers and Glasziou, 2009). In addition, 52% of scientists that were surveyed by Nature agree that we are currently facing a significant 'crisis' of replicability. Ongoing concerns about replicability in academic research carry important consequences for the private sector. Two pharmaceutical companies, Bayer and Amgen, reported replicability rates of 11% and 25% in two independent efforts to reproduce findings from dozens of groundbreaking basic studies in cancer and related areas (Prinz et al., 2011). Moreover, an extensive reproducibility project in cancer biology reveals that a very significant fraction of findings reported in high profile journals cannot be replicated or have smaller effect size than initially reported (Errington et al., 2021, 2014). Lack of replicability can have multiple causes, ranging from the presence of cryptic factors not accounted for by authors, misuse of technical tools, spoiled reagents, bias in data reporting, lack of statistical power or to even outright fraud (Devezer and Buzbas, 2023; Eisner, 2018; Lesperance and Broderick, 2021; Peng, 2015).

Misleading or erroneous findings can disrupt scientific progress, wasting valuable time and resources as researchers attempt to replicate or use the results of a published article (Stern et al., 2014). In some cases, entire research initiatives may be built upon findings that are later discovered to be inaccurate or flawed (Piller, 2022). This might also result in unfair career outcomes if sensational but irreproducible results are rewarded over rigorous studies, especially in high-impact publications (Lemaitre, 2016; Steneck, 2011). However, zero irreproducibility is neither realistic nor necessarily desirable. A certain level of irreproducibility may reflect scientific boldness and risk-taking. Cutting-edge explorations of the unknown are expected to involve occasional backtracking as the bigger picture gradually comes into focus. Thus, striving for perfect replicability could paradoxically stifle innovation and promote overly conservative science. Determining the ideal balance is currently an open question (Shiffrin et al., 2018).

Considering the prominent role played by science in contemporary society, there is a growing need to understand how academic incentives influence the reliability of published results. Indeed, the field of metascience - the scientific investigation of science itself - is gaining popularity (Fortunato et al., 2018). Analyzing how replicability affects the scientific enterprise is required to better characterize how science progresses despite mistakes. It can also help to influence policy to improve science replicability, or at least to limit its most detrimental impacts.

Research focusing on the immune system of *Drosophila* flourished in the early 1990s and is still an active field of investigation. It boasts two thousand articles published since its beginning, dealing with various facets of the immune response of this insect, such as humoral immunity, cellular immunity, response to pathogens, immuno-metabolism, and evolution of the immune system (Westlake et al., 2024 [↗](#)). While the field remains modest in size, it gained visibility in mid 1990s when researchers realized that some mechanisms regulating the *Drosophila* immune defense are conserved in mammals, including humans. The discovery of Toll-like receptors, stemming from the identification of Toll's role in the *Drosophila* immune response (Lemaitre, 2004 [↗](#)), established the value and utility of the field to immunological research, a finding that was acknowledged by a Nobel prize awarded to Jules Hoffmann in 2011.

The medium size of this research field and the lack of direct biomedical influence are ideal for a metascience analysis. The *Drosophila* immunity field is substantially smaller than other fields of research such as HIV, microbiota or cancer, where the rapid accumulation of articles forbids any comprehensive evaluation of scientific claims by an individual or a laboratory. With fewer direct translational or biomedical influences, we expect scientists working on *Drosophila* immunity to be less affected by external biases, and to prioritize recognition by peers. Moreover, *Drosophila* research is generally perceived as solid, because it relies predominantly on its powerful genetics and well-controlled experimental conditions, so the field is particularly interesting for this analysis.

In the framework of the *ReproSci* project, we retrospectively analyzed the replication of 400 articles published between 1959 and 2011 in the field of *Drosophila* immunity research—that is, nearly all the major articles published by a community of scientists in experimental life science over 50 years (Westlake et al., 2025 [↗](#)). For each article, we identified the main, major, and minor scientific conclusions, hereafter referred to as “claims”. The 14-year interval between the time the latest article was published and our study, during which scientific knowledge has advanced and other laboratories have had the opportunity to attempt replication, provides the necessary perspective to efficiently assess the validity of the claims. In a first step, we created a publicly accessible website, displaying all the information on the articles, the claims, and their validity (<https://ReproSci.epfl.ch/↗>). Second, we contacted the *Drosophila* scientific community to provide feedback via this website on the replicability of the claims. In a third step, several teams attempted to replicate 45 major and 11 minor claims that remained unchallenged in the literature at the time. The present article focuses on the meta-science findings, while the scientific consequences of the *ReproSci* project in the *Drosophila* immunity field are presented in a companion article (Westlake et al., 2025 [↗](#)).

This article analyzes the pattern of replicability with two objectives: first, to comprehend the level of replicability within the experimental life science community, and second, to examine whether instances of non-replicability tend to aggregate among journals, first authors, and lead researchers. Such clustering may identify individual characteristics and environmental factors that influence science replicability. We open the Results section with a detailed exploratory analysis of the effect of individual categorical variables on reproducibility outcomes. Then, to assess the joint influence of all covariates and account for the clustering of multiple claims by the same author, we fit a multivariate model to account for confounding effects across our covariates. Because challenged claims are rare and the overall sample is modest, statistical power remains limited. However, taken together, i) the exploratory and ii) multivariate analyses form a single narrative, the first describing exhaustively the observed patterns, and the second testing the robustness of the associations.

## Results

### Methodology and claim assessment

Scientific claims were divided into three categories. The main (one per article), major (3-4) and minor (2-4) claims were manually extracted from 400 articles in the field of *Drosophila* immunity published before 2011. We then assessed the claims of each article to determine if they were

supported or refuted by subsequent publications. The claims were manually categorized by H.W. and B.L. into four broad groups: i) verified by subsequent studies (when later studies confirmed the claim), ii) challenged (when later studies contradicted or disproved the claim), iii) unchallenged (when no later study addressed or attempted to replicate the claim), and iv) partially verified or mixed (when later studies did not lead to a definitive conclusion (see Table 1 in Methods for classification and Westlake et al., 2025). Then, experimental work was performed in several laboratories to test the validity of 45 major and 11 minor unchallenged claims using alternate methods. These claims subjected to experimental validation were selected according to criteria such as appearing suspicious due to the absence of direct follow up or being straightforward to test experimentally. Our data and findings were put on a community-accessible website that opened to the public in early July 2023. The present article focusses on the pattern of irreproducibility and statistical analyses were performed on the 1006 major claims only (referred to as claims below) of the database as updated on December 20<sup>th</sup> 2024. In this article, challenged claims will interchangeably be called irreproducible. Our project focuses on assessing the strength of the claims themselves (i.e. indirect reproducibility or conceptual replicability) rather than testing whether the original methods produce repeatable results (direct reproducibility/replicability). Thus, our conclusions do not directly challenge the initial results leading to a claim, but rather the general applicability (i.e. generalizability) of the claim itself (Goodman et al., 2016; Meng, 2020).

## ***Drosophila* immunity claims are mostly replicable**

After an extensive literature review to assess the validity of claims published from 1959 to 2011, combined with experimental testing of 45 major claims, we found that approximately 61% of major claims (610 out of 1006 across the 400 selected articles) were fully verified (Table 2). Among these 610 verified claims, 91.7% were corroborated by independent laboratories, 7.2% were confirmed by the same laboratory, and 0.7% were verified through experimental work conducted as part of the *ReproSci* project. This verification rate is higher than that observed in other fields where reproducibility has been systematically assessed (Amaral et al., 2025; Amaral et al., 2019; Da Costa et al., 2022; Errington et al., 2021; Fidler et al., 2017; Youyou et al., 2023). These results may reflect the robust scientific standards and methodological rigor of *Drosophila* research, which benefits from the availability of diverse and powerful genetics approaches and from a collaborative culture of rapid reagent sharing, that facilitates replication across laboratories.

We also observed that 23.8% of major claims had no published follow-up (Table 2, Figure 1); these unchallenged claims are particularly interesting and will be discussed in detail in the next section. Additionally, 7.5% of the claims were categorized as partially verified. This category encompasses cases where insightful data were accompanied by incomplete interpretations, or where incomplete data were paired with insightful interpretations. Importantly, 6.8% of claims (69 out of 1006) were challenged. Among these, 44.9% (31 out of 69) were contested by published articles including 7.2% by the same authors, and 55.1% (38 out of 69) were challenged experimentally as part of the *ReproSci* project. Thus, the experimental validation performed during this project doubled the number of reported challenged claims. Finally, 1.2% of major claims were classified as “mixed” (Table 2). This category includes claims that presented particularly challenging issues (e.g., conflicting data), making it impossible to definitively categorize them as either verified or challenged.

In conclusion, contrary to the prevailing “reproducibility crisis” narrative, a significant proportion of the findings published in the field of *Drosophila* immunity before 2011 has been subsequently verified by independent studies.

## **A significant fraction of unchallenged claims is non-reproducible**

Our analysis highlights a significant proportion of “unchallenged” claims that were not followed by subsequent studies and whose validity was unknown prior to our experimental evaluation. Most reproducibility studies have focused on verified or challenged claims so little is known about

**Table 1. Verification categories used to assess claims.**

| Category                                     | Characteristics                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                   |
|----------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <i>Verified</i>                              | <ul style="list-style-type: none"> <li>An independent lab or lab(s) have generated results that directly support this claim using similar or alternate methods.</li> </ul>                                                                                                                                                                                                                                                                                                                                                                                                                                                                        |
| <i>Verified by same authors</i>              | <ul style="list-style-type: none"> <li>Subsequent work by the same lab or authors supports or extends this claim.</li> </ul>                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                      |
| <i>Verified by reproducibility project</i>   | <ul style="list-style-type: none"> <li>This claim is supported by experiments executed specifically to assess it in the Lemaitre Lab or other laboratories in collaboration with the <i>ReproSci</i> project.</li> </ul>                                                                                                                                                                                                                                                                                                                                                                                                                          |
| <i>Partially verified</i>                    | <ul style="list-style-type: none"> <li>Part of the claim is verified by additional data, but part of it is contradicted</li> </ul> <p><b>OR</b></p> <ul style="list-style-type: none"> <li>The claim is not supported by subsequent research, but the data used to justify the claim are consistent with current knowledge. The data were interpreted incorrectly either due to lack of context at the time or other confounding factors.</li> </ul> <p><b>OR</b></p> <ul style="list-style-type: none"> <li>The claim is supported by subsequent research, but the data used to justify the claim in the original publication are not</li> </ul> |
| <i>Unchallenged</i>                          | <ul style="list-style-type: none"> <li>No additional data is available that significantly supports or contradicts this claim.</li> </ul>                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                          |
| <i>Unchallenged, logically consistent</i>    | <ul style="list-style-type: none"> <li>No additional data is available that directly supports this claim, but the claim is consistent with other well-supported related data in the field.</li> </ul>                                                                                                                                                                                                                                                                                                                                                                                                                                             |
| <i>Unchallenged, logically inconsistent</i>  | <ul style="list-style-type: none"> <li>No additional data is available that directly contradicts this claim, but the claim is inconsistent with other well-supported related data in the field.</li> </ul>                                                                                                                                                                                                                                                                                                                                                                                                                                        |
| <i>Mixed</i>                                 | <ul style="list-style-type: none"> <li>Both data that supports and contradicts this claim have been published in the field. It is unclear based on the published data whether the claim is correct, incorrect, or whether our understanding of the topic is simply incomplete.</li> </ul>                                                                                                                                                                                                                                                                                                                                                         |
| <i>Challenged</i>                            | <ul style="list-style-type: none"> <li>An independent lab or lab(s) have generated results that directly contradict this claim using similar or alternate methods.</li> </ul>                                                                                                                                                                                                                                                                                                                                                                                                                                                                     |
| <i>Challenged by same authors</i>            | <ul style="list-style-type: none"> <li>Subsequent work by the same lab or authors contradicts this claim.</li> </ul>                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                              |
| <i>Challenged by reproducibility project</i> | <ul style="list-style-type: none"> <li>This claim is contradicted by experiments executed specifically to assess it in the Lemaitre Lab or other laboratories in collaboration with the <i>ReproSci</i> project.</li> <li>When complete, a link to a writeup of the experimental results is included.</li> </ul>                                                                                                                                                                                                                                                                                                                                  |

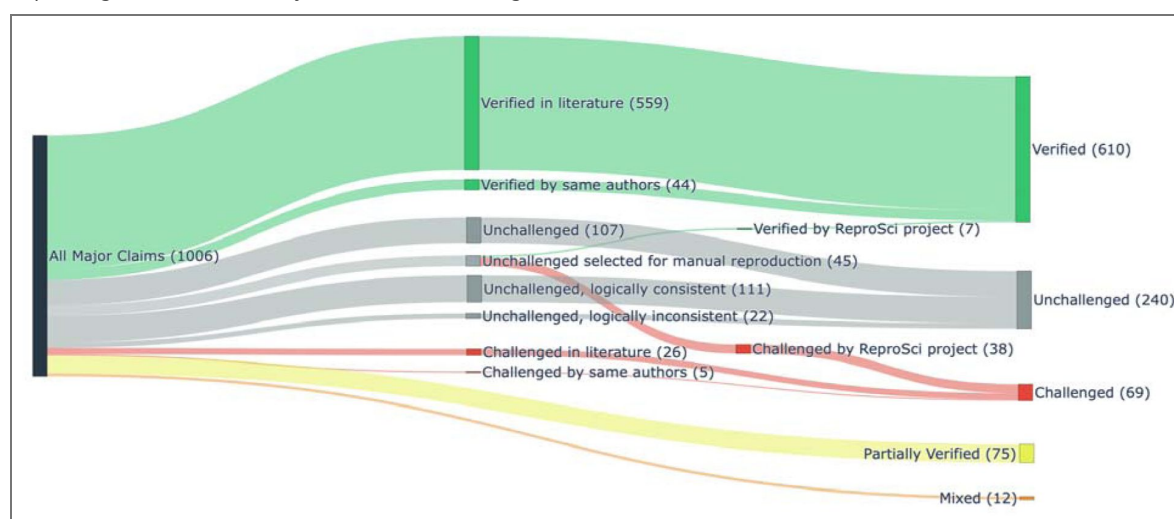
**Table 2. Proportion of claims in broad reproducibility assessment categories before and after experimental validation by the *ReproSci* project.**

See description of each category in Material and Methods. In this article, we are using indirect reproducibility, also called conceptual replicability, to assess the validity of a claim using alternate methods to those used in the original study.

| Major claim        | Nb (%)      | Nb (%) after experimental validation |
|--------------------|-------------|--------------------------------------|
| Verified           | 603 (59.9%) | 610 (60,6%)                          |
| Unchallenged       | 285 (28,3%) | 240 (23,8%)                          |
| Partially verified | 75 (7,5 %)  | 75 (7,5%)                            |
| Mixed              | 12 (1,2%)   | 12 (1,2%)                            |
| Challenged         | 31 (3,1%)   | 69 (6,8%)                            |
| Total              | 1006        | 1006                                 |

**Figure 1. Flow of the study.**

Claims were first classified in five categories based on literature. 45 unchallenged claims were selected to be experimentally tested and classified either as challenged or as verified. We further classify some unchallenged as *consistent* or *inconsistent* depending on their consistency with current knowledge.s



the vast number of statements generated in a scientific community without subsequent direct validation.

Despite the lack of validation, these “unchallenged claims” are often assumed valid and contribute to the broader construction of scientific knowledge. An important challenge is to estimate the percentage of unchallenged major claims that are actually non-reproducible.

Prior to our experimental validation, we identified a total of 285 unchallenged claims, representing 28.3% of all major claims published up to 2011. We conducted experimental evaluations of 45 unchallenged major claims with the assistance of multiple laboratories (Westlake et al., 2025 [DOI](#)). The selection of unchallenged claims for testing was intentionally biased toward two categories: (i) those that appeared suspicious in light of the existing literature (31 claims), and (ii) those that were relatively straightforward to test experimentally (14 claims). Among the 45 unchallenged major claims tested, 84% (38 out of 45) could not be experimentally reproduced and were then reclassified as “challenged by the *ReproSci* project.” Conversely, only 16% (7 out of 45) were successfully reproduced and reclassified as “verified by the *ReproSci* project” (Figure 1 [DOI](#)).

Notably, all 31 suspicious unchallenged claims were challenged, while 7 out of the 14 unchallenged “easy to test” claims were also challenged. Excluding the 45 unchallenged major claims that were experimentally tested, we categorized the remaining 240 unchallenged claims into three groups: 1) unchallenged logically consistent – claims not directly tested but consistent with current knowledge of the field (111 claims); 2) unchallenged logically inconsistent – claims that appear inconsistent with current knowledge (22 claims); 3) Unclassified – claims that could not be definitively categorized (107 claims) (Figure 1 [DOI](#)). In the remainder of the text, we analyze the claims following *ReproSci* experimental validation, labelling the 45 selected claims as either challenged or verified.

This analysis reveals that a substantial number of claims remain unchallenged even a decade after publication. Moreover, our findings suggest that a considerable proportion of these claims may not be reproducible.

## Higher representation of challenged claims in trophy journals and from top universities

Articles published in high-impact journals receive high visibility, a key factor in career progression and funding security (van Wesel, 2016 [DOI](#)). Additionally, high-impact publications are increasingly used as metrics to evaluate university rankings. Consequently, scientists face significant internal and external pressures to publish in such prestigious journals (Udesky, 2025 [DOI](#)). On the other hand, high-impact journals aim to attract the most surprising findings, which are inherently more likely to be challenged than more conventional studies. Based on this, we hypothesized that articles published in high-impact journals are more likely to include non-reproducible results. To test this hypothesis, we estimated the number of challenged claims published in journals categorized as having low or medium-impact factors (<10), high-impact factors (>10 and <50), and “trophy journals” of the highest reputation (>50: *Science*, *Nature*, and *Cell*) in the *Drosophila* immunity at that time. The results are summarized in Figure 2 [DOI](#), with a more detailed breakdown by journal provided in Table S1 [DOI](#). Compared with low-impact journals, where 5.8 % of major claims were challenged, the proportion rose to 7.8 % in high-impact journals (odds ratio [OR] = 1.36, 95 % confidence interval [CI] 0.80–2.32) and more than doubled to 12.9 % in trophy journals (OR = 2.39, 95 % CI 1.06–5.40). We recall that the OR expresses how much more (OR > 1) or less (OR < 1) likely a claim is to be challenged relative to articles published in low-impact journals, while the 95 % CI indicates the range of values compatible with random sampling error; OR intervals that exclude 1 denote a statistically significant effect. Thus, our analysis show major claims published in low-impact journals are significantly more likely to be reproducible than major claims published in trophy journals. One explanation for this trend is that claims published in high-impact journals are more frequently subjected to replication attempts or used as a foundation for further research, thereby increasing the likelihood of identifying irreproducible results. Indeed, low-impact journals include a higher share of claims remaining unchallenged

(29.3 %, 95 % CI 25.9–32.9 %) compared to high-impact journals (13.9 %, 95 % CI 10.5–18.2 %) and to trophy journals (17.7 %, 95 % CI 10.2–29.0 %). However, variation in the number of unchallenged claims does explain the higher proportion of challenged claims in trophy journals compared to high-impact journals. If we take into consideration the percentage of verified, unchallenged and challenged claims, Figure 2B shows that high-impact but not trophy or low-impact journals contain the highest proportion of verified claims. Thus, we conclude that while trophy journals may indeed attract some of the best science, they also appear more likely to include irreproducible claims.

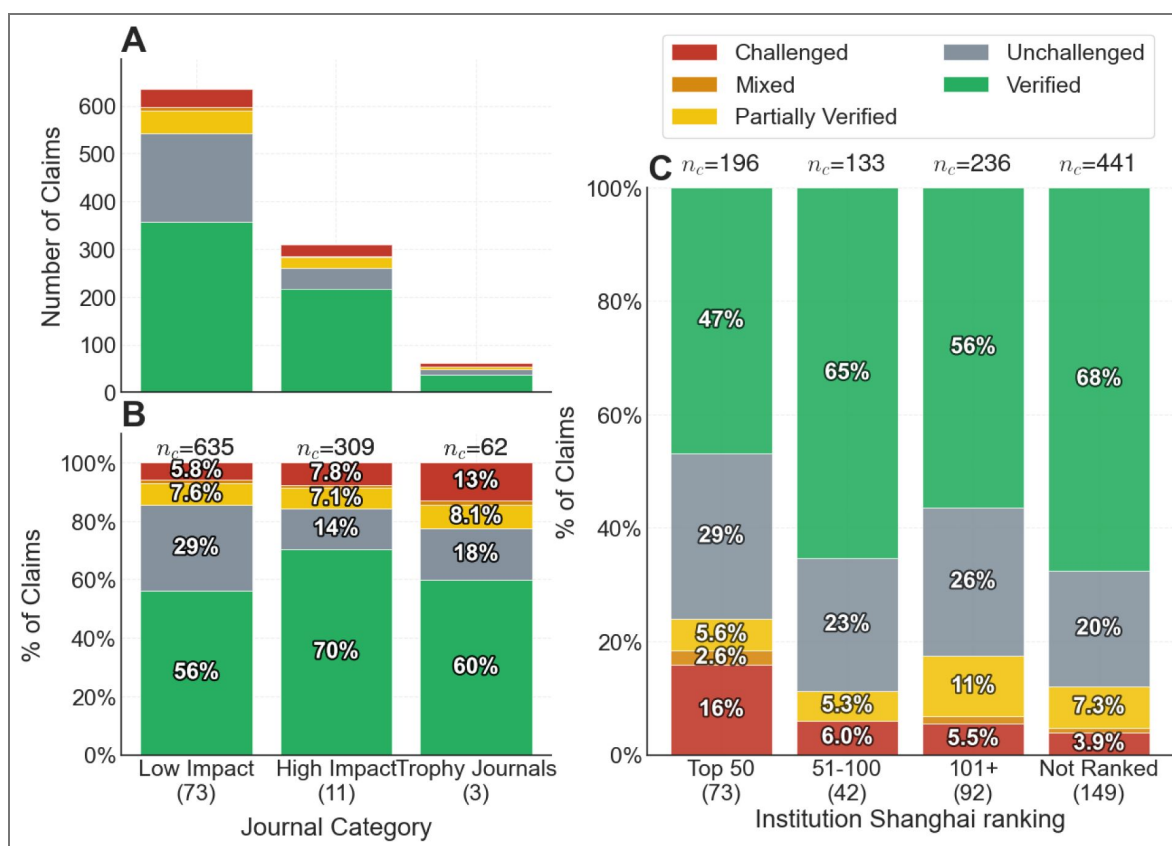
We then investigated whether the percentage of irreproducibility varies among laboratories affiliated according to the prestige of their affiliate universities. To test this hypothesis, we classified all the major claims published according to the Shanghai ranking of the last author university. Figure 2C shows that claims from laboratories at top-50 universities were challenged far more often 15.8 % (95 % CI 11.4–21.6 %) than other institutions, up to a four-fold increase in the odds of being questioned (OR = 4.69, 95 % CI 2.53–8.70) versus 3.9 % for unranked institutions, which notably include hospitals and research institutes such as the French CNRS. Even after adjusting for journal tier (see model SI1), university prestige remains the dominant signal: claims from top-50 institutions are 3.6 times more likely to be challenged than those from unranked universities (OR 3.57, 95 % HDI 1.67–7.92). By comparison, publishing in a “trophy” journal adds only a notable and uncertain risk (OR 1.63, 0.66–3.88), while other journal or ranking bands show no clear effect. Put simply, elite affiliation more than journal prestige best predicts irreproducibility in our sample.

## The irreproducibility rate has increased over time as the field has grown in popularity

The reproducibility crisis narrative suggests that irreproducibility has increased over recent decades, posing a significant challenge to the life sciences (Baker, 2016; Hunter, 2017). Our dataset spans 50 years, from the onset of *Drosophila* immunity research as a relatively marginal field to its raise to prominence, with a notable increase in publications since the mid-1990s driven by growing interest in innate immunity. We analyzed fluctuations in irreproducibility over time by grouping major claim assessments into time windows: pre-1991, followed by 5-year intervals until 2011. The percentage of irreproducibility was calculated for each time window (Figure 3). Across the five windows examined, the proportion of challenged claims climbed steadily from 0% in the small corpus prior to 1991 (0/29 articles), to 4.6% during 1992–1996, and then further increased to 7.7% between 1997 and 2001. This upward trend peaked at 8.3% during the 2002–2006 period, before slightly declining to 6.1% between 2007 and 2011 (Figure 3). Despite these apparent temporal shifts, none of the pairwise comparisons between periods reached statistical significance, with p-values consistently  $\geq 0.15$ . Thus, while the trends highlighted in our study are consistent with a raise of irreproducibility, we cannot exclude that the observed upward trajectory might reflect sampling variation rather than a definitive sustained temporal trend. Further studies focusing on articles published after 2011 could clarify whether the mild decrease we observed marks the beginning of a downward trend or reflects fluctuations at a consistently high level.

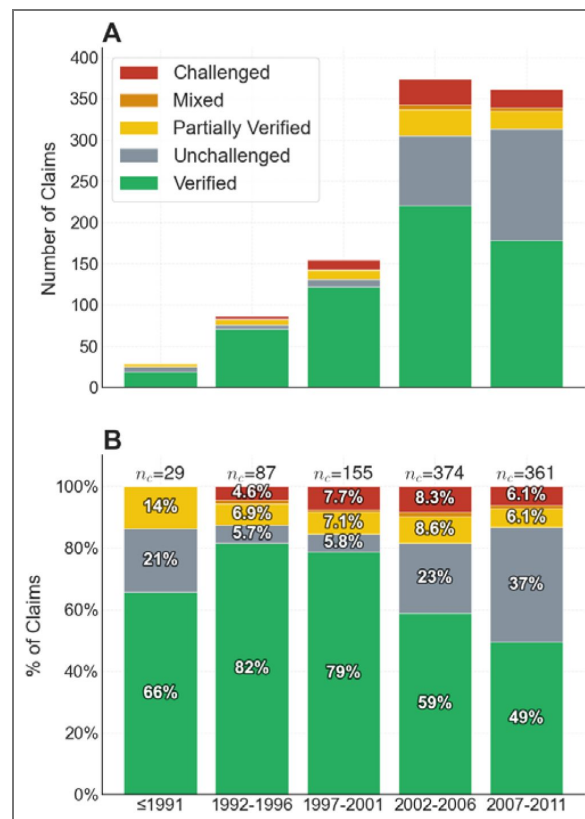
Interestingly, we observed a peak in verified claims between 1992 and 2001, accompanied by a lower percentage of unchallenged claims (Figure 3). This period coincided with the early expansion of the field when many researchers worked on similar research questions with a relatively narrow conceptual framework of innate immunity. The subsequent increase in unchallenged claims may reflect the rapid conceptual expansion of the field, which likely outpaced the growth in the number of researchers, reducing opportunities for overlapping or validating research efforts. An alternative explanation for this trend is that in the early stages of the field, researchers uncovered the largest, easily detectable effects. As the field matured, focus shifted to subtler phenotypes and complex regulatory mechanisms, making replication more challenging.

Overall, our time-course analysis suggests that irreproducibility has increased over the decades, aligning with the reproducibility crisis narrative. However, an alternative hypothesis is that irreproducibility varied alongside the field’s popularity, peaking once when the field became



**Figure 2. Percentage of claims by category according to journal-impact and institution ranking.**

(A, B) The distribution of claims across the five assessed categories for articles published in low-impact, high-impact, and trophy journals. The number of claims is indicated at the top. Panel A presents raw counts, while Panel B shows normalized data. (C) The distribution of claims across the five assessed categories according to the ranking of the last author institution in Shanghai Ranking's 2010 Academic Ranking of World Universities. The number of claims is indicated at the top, while the number of institutions in each category is indicated at the bottom. The Not Ranked categories encompass research institutes (ex. CNRS), hospitals and other institution that are not included in the Shanghai ranking. In this figure, and in the rest of the paper, the number  $n_c$  indicates the number of claims in each category, while the number at the bottom highlight the count in each category (e.g there are 11 high-impact journals, and 73 Top-50 institutions).



**Figure 3. Evolution of claim replicability over time.**

(A, B) The distribution of claims by categories of assessment over time is shown in A, and the proportion normalized over the total number of claims per time period is shown in B.

popular and had attracted newcomers who may not have been fully acquainted with its established standards or who joined the field by opportunistically. Furthermore, the rise in the percentage of challenged claims was accompanied by an increase in unchallenged claims, indicating a reduced degree of validation within the field.

## First-author patterns of irreproducibility

Having analyzed the pattern of irreproducibility across a large number of articles, representing contributions from multiple laboratories over five decades, we next investigated the distribution of irreproducibility among authors and laboratories. A central question is whether irreproducibility is evenly distributed across the scientific community, or if hubs — specific authors or laboratories — are responsible for producing more invalid statements. First, we explored whether reproducibility is associated with the first author, who typically conducts the work, or with the last author, who manages the laboratory and may foster a research environment that produces irreproducible results.

To explore the relationship between irreproducibility and first authorship, we computed the percentage of “challenged” articles attributed to each first author. Additional information on first author was obtained through interviews with principal investigators (PIs) or via public LinkedIn profiles, including variables such as gender and position in the laboratory.

In total, 289 first authors contributed to the 400 articles analyzed by the *ReproSci* project, with an average of 3.45 claims per author. Notably, 47 authors (16.3%, 95% CI: 11.4–19.6%) published at least one non-reproducible claim, with one author reaching up to 83% irreproducibility (5 challenged claims out of 6 from two articles). [Figure 4A](#) displays the distribution of challenged claims among first authors. A statistical dispersion analysis revealed a Gini index of 0.88 (inequality index ranging from 0 -perfect equality- to 1 -extreme inequality-), where the top 10% of first authors accounted for 73.9% of all challenged claims, and the top 20% accounted for all 100% of challenged claims ([Figure 4B](#)), suggesting a highly unequal distribution of irreproducibility according to first author.

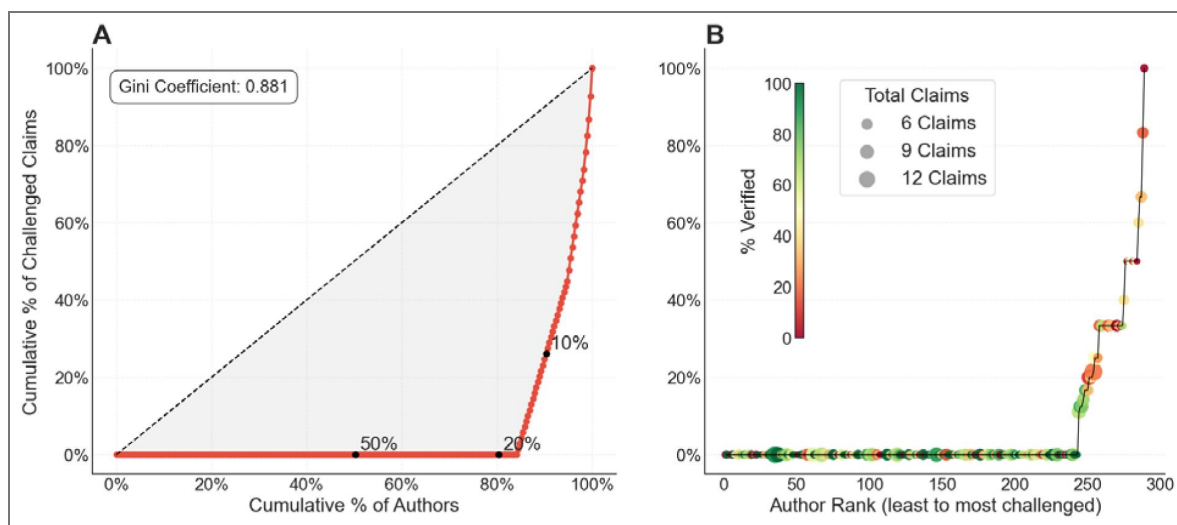
We next investigated whether irreproducibility varied by the gender of the first author. The proportions of challenged articles were similar for male and female first authors: 7.3% and 6.7%, respectively, and this difference was not statistically significant (OR = 1.09, 95% CI: 0.66–1.78, Fisher’s exact  $p = 0.802$ ; [Figure 5A](#)).

We were able to determine the status of most of first authors at the time of the article was published, most being PhD students ( $n=168$ ) or postdoctoral researchers ( $n=87$ ), with smaller numbers identified as technicians ( $n=5$ ), Masters students ( $n=5$ ), PIs ( $n=8$ ), or permanent staff ( $n=6$ ) ([Figure 5B](#)). No major differences in irreproducibility were observed between PhD students and post-docs. While we did observe a higher proportion of challenged claims among PIs and Master’s students, and a lower proportion among technicians and permanent staff, these trends are difficult to interpret due to small sample sizes.

Finally, we tested the relation between the number of articles published by a first author and the proportion of irreproducible research. No significant trend was observed (linear regression;  $r^2=0.01$ ,  $p=0.088$ ; [Figure 5C](#)), suggesting the level of author productivity is independent of irreproducibility. In summary, we observed considerable variability in irreproducibility among first authors, suggesting that individual authors may substantially contribute to reproducibility issues.

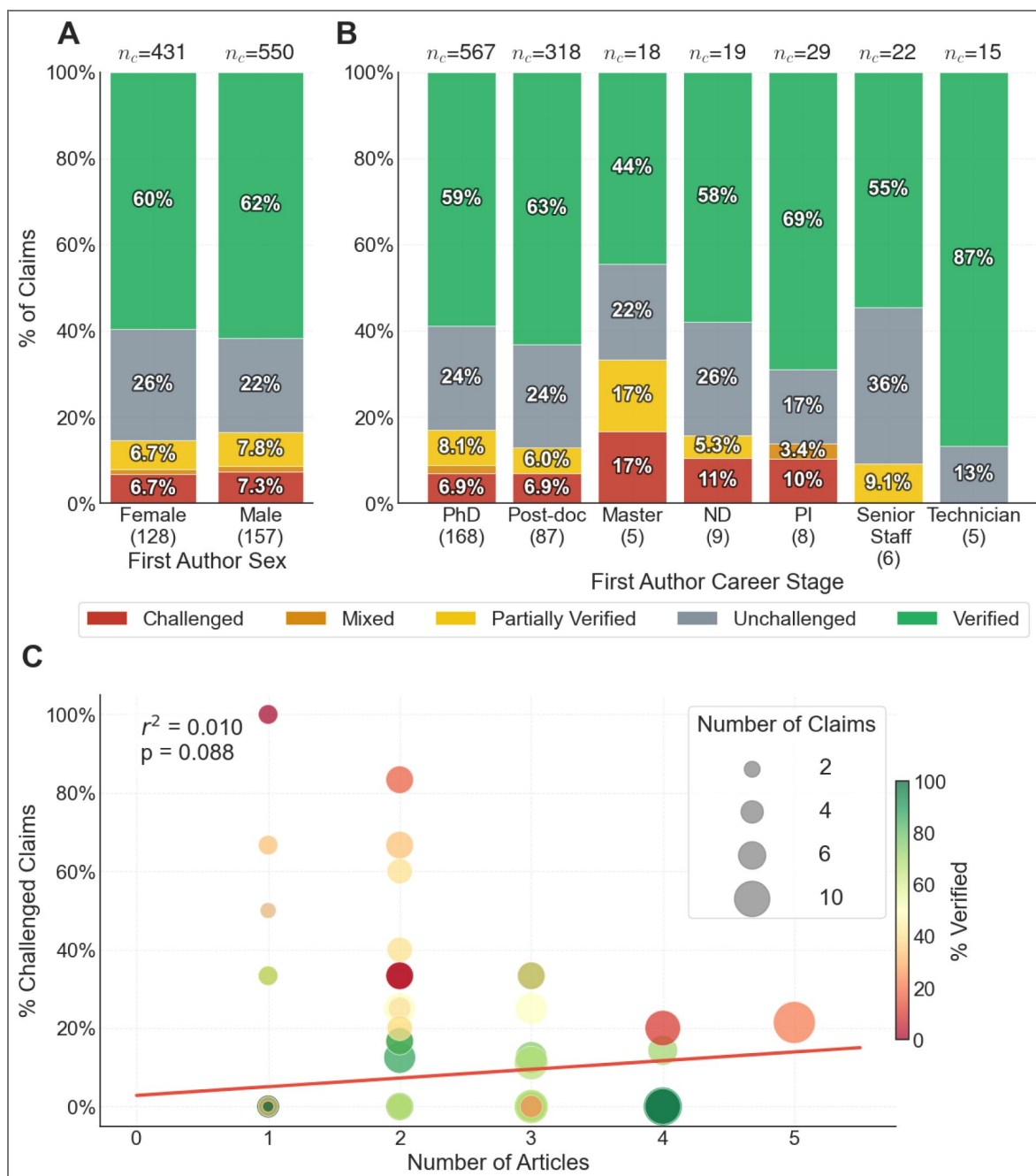
## Lead authors pattern of irreproducibility

We now turn to the question of whether irreproducibility varies according to laboratory. As it is common practice in biology, we assigned the last author (referred to as PI for principal investigator) to represent the laboratory. The 400 articles analyzed were co-authored by 156 leading authors from various institutions, predominantly from France, Sweden, Japan, South Korea, and the USA. For each last author, we estimated the number of major claims in each of the five categories. The variation in irreproducibility across these PIs is shown in [Figure 6A](#). A



**Figure 4. Distribution of irreproducibility according to first authors.**

A) Distribution of irreproducible claims among first author. We note that all irreproducible claims are produced by less than 20% of the first authors. B) Distribution of irreproducibility for all first authors. Each dot represents an author, ranked according to the proportion of challenged claims. The size of the circle indicates the number of claims and the color the percentage of verified claims.



**Figure 5. Influence of parameters associated with first author on reproducibility.**

A) Repartition of claims in each assessment category according to the gender of the first authors. The gender of four authors could not be inferred based on first name ambiguity explaining why the total number of authors is 285. B) Repartition of claims in each assessment category according to status of the first authors. The stages of career of 279 first authors were obtained by interviewing last authors, querying LinkedIn public profiles or extrapolated according to the publication list. Senior staff refers to permanent position scientist working in a lab. ND stands for Not Determined. C) Proportion of challenged claims as a function of the number of published articles for first-authors. The color indicates the percentage of claims that are verified.

statistical dispersion analysis in [Figure 6B](#) revealed a Gini coefficient of 0.856: the top 10% of all last authors accounted for 71.0% of all challenged claims, while the top 20% accounted for 94.2%, suggesting a highly unequal distribution of irreproducibility according to laboratory. We next explored whether there were differences in irreproducibility between male and female last authors. The percentage of challenged claims was slightly higher for male last authors (7.0%, 95% CI: 7.0%–7.1%) compared to female last authors (5.8%, 95% CI: 5.7%–6.0%), but this difference was not statistically significant (OR = 1.22, 95% CI: 0.59–2.51; Fisher’s exact  $p = 0.73$ ; [Figure 7A](#)).

Pressure to publish is typically stronger among junior PIs who often lack tenure and must establish their laboratory. To examine whether this pressure might be associated with irreproducibility, we classified PIs as junior or senior based on whether they had published a last-author article at least five years prior to the considered publication. Although [Figure 7B](#) indicates a slightly higher proportion of challenged claims among junior PIs (7.2%, 95% CI: 7.2%–7.3%) compared to senior PIs (6.5%, 95% CI: 6.5%–6.6%), this difference was not statistically significant (OR = 1.12, 95% CI: 0.68–1.83, Fisher’s exact  $p = 0.70$ ). This modest difference was further attenuated when considering mixed and partially verified claims.

We then explored whether the percentage of irreproducibility tended to decline according to the number of published articles as last author ([Figure 7C](#)). We could hypothesize that PIs that have transiently work on *Drosophila* immunity would have more irreproducible claims, or on the contrary that the most productive laboratories compromise speed over rigor. However, no significant trend was observed between the percentage of challenged claims and the number of articles published by an author (linear regression;  $r^2=0.004$ ,  $p=0.67$ ).

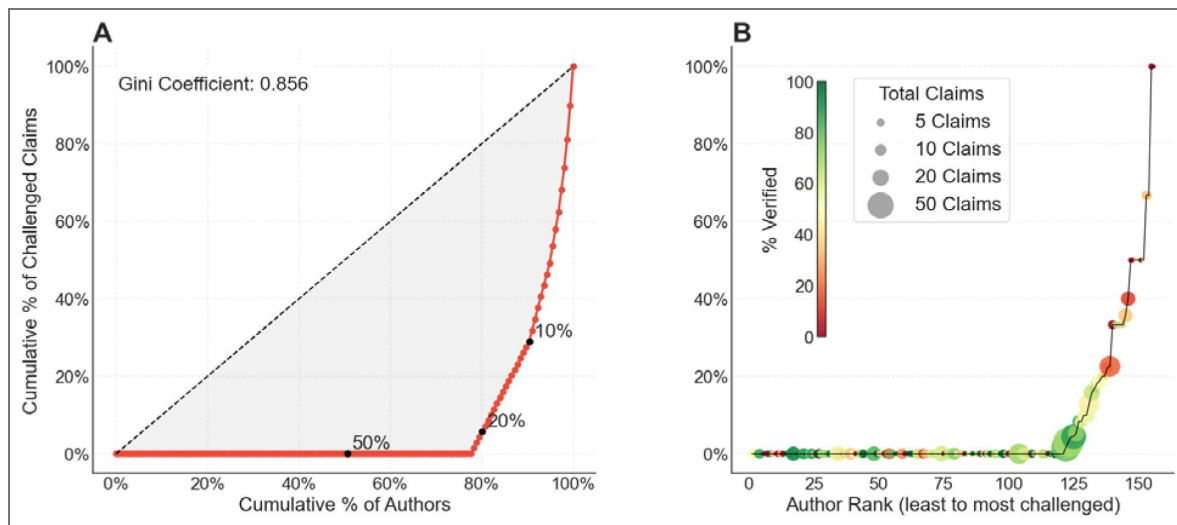
We found that the percentage of irreproducibility has increased over time as the field became more popular. We therefore explored if this increase in irreproducibility affects all PIs or only new PIs that joined the field at a later time. We show in [Figure 8A](#) the proportion of challenged claims by last authors according to the publication date of their first article on *Drosophila* immunity, either as last or as first authors. We observed a significant association between irreproducibility and the period when principal investigators (PIs) established their laboratories. Specifically, PIs, who started their laboratories after 1995, had a significantly higher percentage of challenged claims (8.1%, 95% CI: 8.0%–8.1%) compared to those who began their labs before 1995 (2.6%, 95% CI: 2.6%–2.7%; OR = 3.24, 95% CI: 1.38–7.59, Fisher’s exact  $p < 0.001$ ).

Among the PIs working in the *Drosophila* immunity field, who established their lab after 1995, we identified 22 who previously published at least one first-author article in this field during their PhD or postdoctoral training in another laboratory. We hypothesized that these authors, having gained prior hands-on experience on *Drosophila* immunity, would be less prone to publishing irreproducible claims. Consistent with this hypothesis, we observed significantly lower levels of irreproducibility among these experienced authors (3.5%, 95% CI: 3.4%–3.5%) compared to authors who had not been trained in laboratories working on *Drosophila* immunity (8.2%, 95% CI: 8.1%–8.2%; OR = 2.48, 95% CI: 1.19–5.18, Fisher’s exact  $p = 0.01$ ; [Figure 8B](#)).

We conclude that irreproducibility varies significantly across laboratories with a limited number of laboratories accounting for most of the irreproducible claims. As for first authors, author gender is not significantly associated with reproducibility. Our studies point to a significant role of experience: PIs that had previously worked in another *Drosophila* immunity laboratory to the extent that they published a first author article in that laboratory were less prone to publish irreproducible claims.

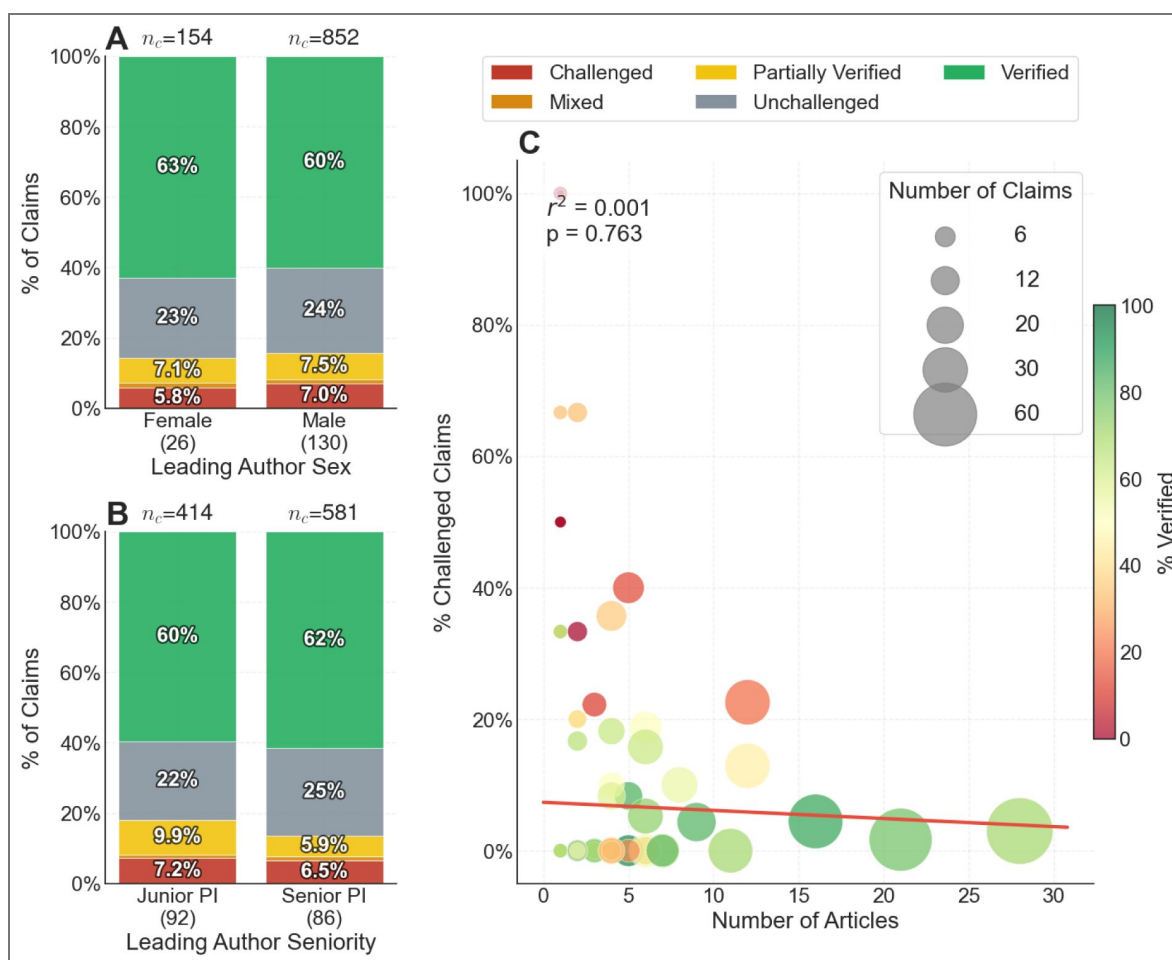
## Irreproducibility according to research styles

Among the various career progression strategies, one involves deepening expertise and specialization within the same field, while another entails transitioning to different fields or exploring different career paths. This may reflect the personality of the scientist, either opting for a more detailed, focused approach or seeking continual new challenges. Scientists from the past were often spending their life on a specific topic within a field. In the last decades, changing field has become the norm ([Zeng et al., 2019](#)). Accordingly, scientists are not funded by institutions to



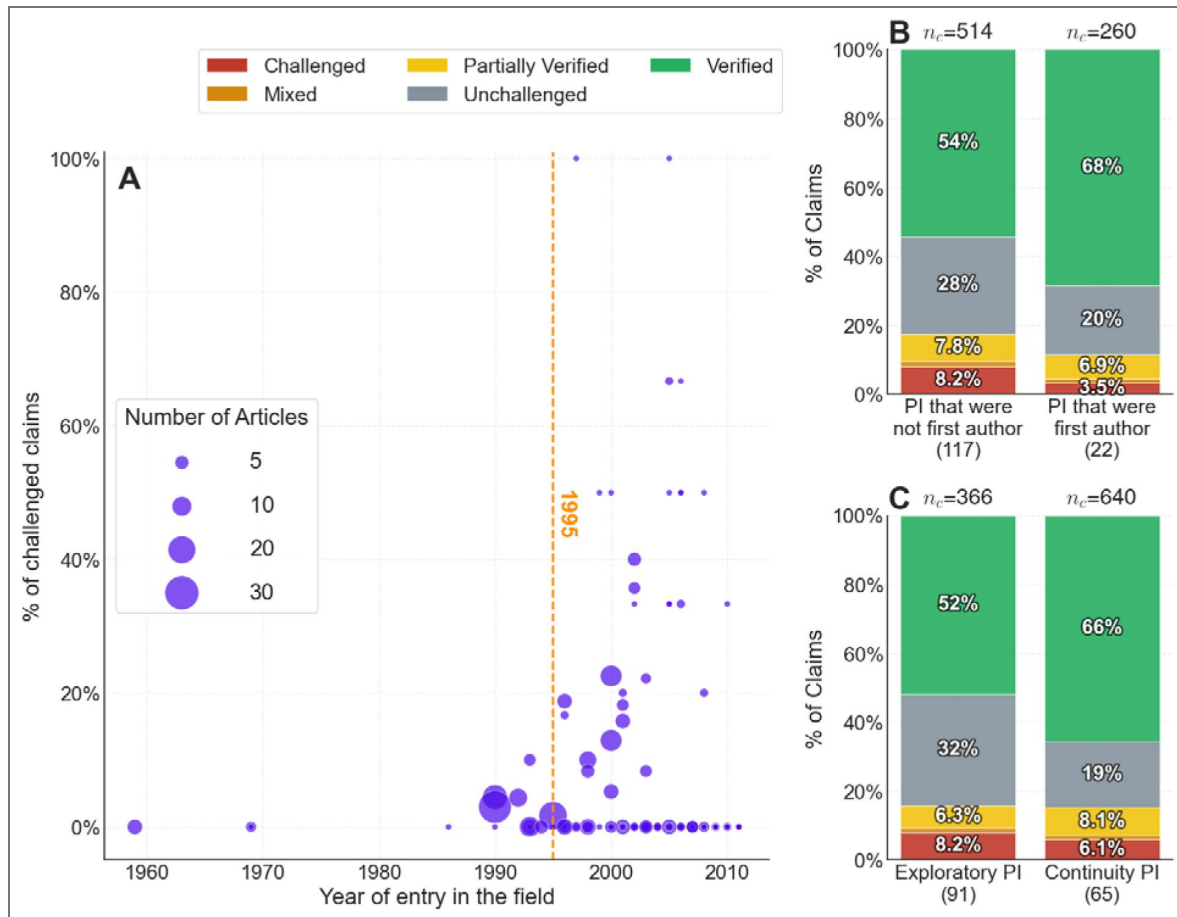
**Figure 6. Distribution of irreproducibility according to last authors (PI).**

A) Distribution of irreproducible claims among leading authors. B) Distribution of irreproducibility for all leading authors. Each dot represents a leading author, ranked according to the proportion of challenged claims. The size of the circle indicates the number of claims and the color the percentage of verified claims.



**Figure 7. Influence of various parameters associated with the last author on reproducibility.**

A) Distribution of claims by category according to the gender of the last authors. B) Distribution of claims by category according to the seniority of the last authors. Senior PI were defined as having published a last author article at least 5 years before the considered articles. The total of junior and senior authors is higher than the total of last authors as some PIs published both as junior and senior. C) Proportion of challenged claims by last author as a function of number of articles.



**Figure 8. Patterns of irreproducibility by last author according to time of starting their laboratory and research style.**

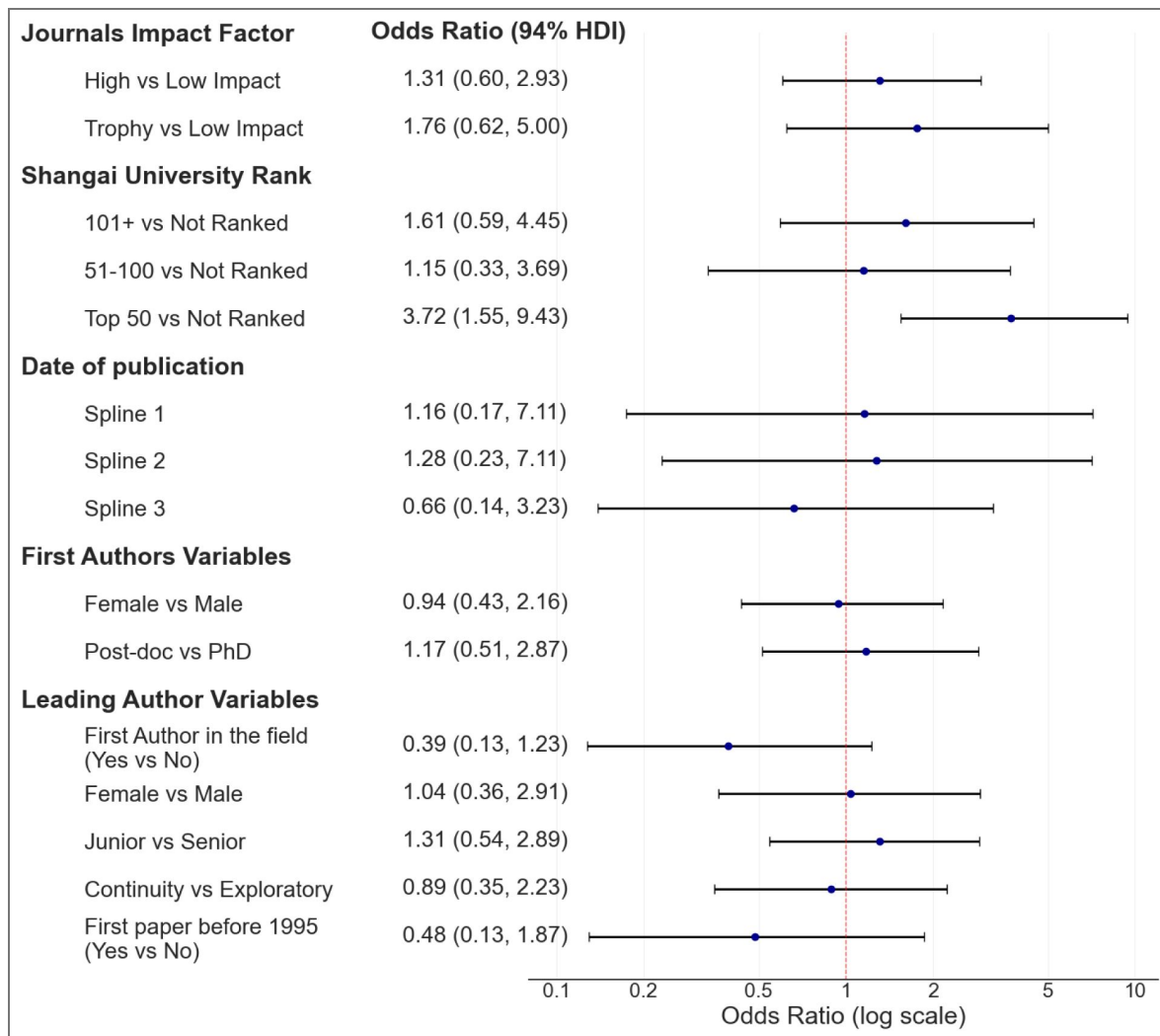
A. The proportion of challenged claims by last author is shown according to the time at which the author published their first paper as a first or last authors. The increase in irreproducibility is associated with PIs who started their lab after 1995. Circle size indicates the number of published articles by each author. B. Distribution of claims by last author according to whether they have published at least one first author article on *Drosophila* immunity in another laboratory. C. Distribution of claims according to research style. Proportion of challenged claims by PIs who have continuously work on the *Drosophila* immunity ('continuity PIs') or made a transient incursion into the field ('exploratory PIs'). Only PIs, who published as last authors after 1995, were taken into consideration.

work on a topic long-term anymore, but rather funding is granted by agencies for specific short-term projects. Project grants define research with the primary goal to do ‘breakthrough discoveries’, not accumulative research, which is sometimes treated as the ‘pedestrian’ or the ‘stamp collector’ way of doing research. This leads scientists to move quickly into new research fields in search for opportunities of funding and recognition, a nuanced hypothesis supported by the recent description of a “pivot penalty” for scientists that change career tracks (Hill et al., 2025 [↗](#)). We can define these two strategies as either “continuity PIs,” which spend their career on one core topic, and “exploratory PIs,” which regularly venture into new topics. To further explore this question, we classified all last authors that set up their laboratory after 1995 as ‘continuity’ PIs or ‘exploratory’ PIs. Exploratory PIs are defined as those who made a transient incursion in the *Drosophila* immunity field before moving on to another field or leaving academia. In contrast continuity PIs worked continuously in the *Drosophila* immunity field throughout their careers. We observed that continuity PIs had a lower proportion of challenged claims (6.1%, 95% CI: 6.0%–6.2%) compared to exploratory PIs (8.2%, 95% CI: 8.1%–8.3%). However, this difference was not statistically significant (OR = 1.38, 95% CI: 0.84–2.26, Fisher’s exact  $p = 0.24$ ; Figure 8C [↗](#)). We also noted a markedly higher share of unchallenged claims among exploratory PIs than among continuity PIs, 32.2 % (95 % CI 32.1–32.4 %) versus 19.1 % (95 % CI 19.0–19.2 %), respectively. This difference was statistically significant (OR = 2.02, 95 % CI 1.50–2.71; Fisher’s exact  $p < 0.001$ ), reinforcing the idea that researchers, who move on to new topics, tend to leave a larger fraction of their earlier findings without follow-up or validation by the community. Since unchallenged claims are more likely to be irreproducible, this suggests that irreproducibility is higher in the work of PIs with an exploratory style, as opposed to those engaged in a continuity-driven research style.

## Multivariable analysis of predictors of claim irreproducibility

We have presented above a comprehensive exploratory data analysis, examining one variable at a time. However, drawing robust conclusions necessitates the joint analysis of all covariates, while appropriately accounting for the non-independence of claims made by the same authors. Thus, we fitted a multivariate Bayesian mixed-effects logistic model to uncover the effect of each covariate, while taking into account all other confounding variables at our disposal (Abril-Pla et al., 2023 [↗](#); Capretto et al., 2022 [↗](#))(Figure 10 [↗](#)). The model simultaneously adjusts for author demographics, laboratory attributes and journal or institutional prestige and predict a binary outcome for each claim: challenged or not challenged (including Verified, Partially Verified, Mixed and Unchallenged) (See in Supplementary figure S3 [↗](#) a similar model to predict unchallenged or not unchallenged claims). As the claims from the same first or last author are not independent, we included a mixed effect for each first author (289) and each last author (156). These effects capture the influence of a particular author in driving a claim reproducibility. We conducted extensive test on model formulation and convergence (see methods, and Supplementary information for posterior predictive and diagnostics plot). Note that as this analysis is Bayesian, we report the Highest Density Interval (HDI) instead of the confidence interval. We observe that, none of the first-author factors (gender: OR= 0.94, 94% HDI: 0.42–2.16; trainee stage: OR = 1.17, 94% HDI: 0.51–2.87) or leading-author factors (gender, seniority, continuity, or having entered the field before 1995) showed a decisive association with irreproducibility once every other covariate was taken into account (all HDIs span 1). We still observe a trend toward reduction of challenged claims for lead authors having being trained in *Drosophila* immunity or have started their laboratory before 1995. Figure S3 [↗](#) shows that lead author with an exploratory life style tends to have significantly more unchallenged claims (OR: 2.22, 1.18, 4.55). In contrast, university ranking and journal level covariates dominated the model: articles from top-50 universities were almost four times more likely to contain a challenged claim than those from un-ranked institutions (OR 3.72, 1.55–9.43), and papers in trophy journals (Science, Nature, Cell) carried an elevated uncertain risk (OR 1.76, 0.62–5.00) compared to low-impact journals.

Intermediate effects for high-impact journals and mid-ranked universities were positive but imprecise. Publication year was modelled with a three-knot B-spline to allow for non-linear time trends. None of the individual spline coefficients differed credibly from zero, indicating that once



**Figure 10.** Multivariable analysis of predictors of claim irreproducibility.

journal tier and the other covariates were included the residual calendar-year effect on irreproducibility was negligible, indicating that time of publication itself is not a major factor.

The hierarchical component comprised 289 random intercepts for first authors and 156 for PIs. The posterior distribution of these intercepts (see Supplementary Information) is broad, confirming that some individuals or laboratories have a systematically higher or lower propensity for challenged claims. Nevertheless, only the three most extreme first-author effects had 95 % highest-density intervals that excluded zero (and all with positive irreproducibility) meaning that statistically robust clustering of irreproducibility is confined to a handful of contributors; for the vast majority of authors the collected evidence does not support significant differences in reproducibility from what is expected from a random distribution.

Taken together, the model confirms our exploratory data analysis, namely that it is the prestige of institution and publication in trophy journals rather than the personal characteristics of the scientists that best predicts which claims will later prove to be irreproducible.

The forest plot reports the odds ratio and the 94% highest density interval of each covariate effect on a claim becoming challenged.

## Discussion

The *ReproSci* project presents an in-depth analysis of reproducibility in an experimental science field with a 14-year retrospective period. This initiative focuses specifically on *conceptual replicability*—the ability to independently reproduce key findings based on the information provided—rather than direct reproducibility using the exact same data and conditions (Plesser, 2018). Replicability is of critical importance to the scientific community who seeks to build upon published claims. One of our project's key strengths lies in its comprehensive scope: it spans over fifty years of research, from 1959 to 2011, within a specific scientific community, resulting in a robust database that synthesizes the main findings across decades. This time frame encompasses the formative years of the field, characterized by a small number of pioneering laboratories, followed by a period of rapid expansion driven by increased interest, heightened competition, and publication in high-impact journals. This growth phase was marked by a proliferation of knowledge and conceptual innovation, with novel approaches to studying the immune system emerging. With the awarding of the Nobel Prize in 2011 and subsequently the shift toward more translational research, the growth pace of the *Drosophila* immunity field has since stabilized, and the field remains highly active.

Although our study centers on a specific experimental domain at a particular moment in time with a specific geographical development, it reveals broader patterns of irreproducibility that are likely generalizable. These findings contribute meaningfully to the ongoing conversation about the reproducibility crisis in the experimental life sciences (Fanelli, 2018; França and Monserrat, 2018; Hunter, 2017; Peng, 2015).

Contrary to the more dramatic narratives surrounding the reproducibility crisis (Amaral et al., 2025; Baker, 2015; Errington et al., 2021), our study shows that a substantial proportion of claims—61%—can be considered as verified. The proportion of claims that have been directly challenged remains relatively low (7%), though likely underestimated. This indicates that research in the *Drosophila* immunity field is rather solid, possibly due to the power of *Drosophila* genetics, the collective ethos of sharing within the community, a less competitive environment, and a lack of influence associated with translational research.

A novel and important contribution of our work is the identification of a large fraction of *unchallenged* claims—findings that, despite being published, have never been independently tested. The presence of these unchallenged claims could result from the rapid expansion of the field limiting overlapping research that would validate them. However, the existence of unchallenged claims could also be explained if they are not reproducible. For instance, the absence of follow-up studies could reflect authors that are active in the field being aware of the fragility of specific unchallenged claims, and so not pursuing them. The existence of unchallenged claims could also stem from the reluctance of other researchers to publish contradictory statement

due to the stigma of publishing negative results, the so-called “file drawer” effect (Rosenthal, 1979). Indeed, many scientists report abandoning verification efforts when unable to reproduce a colleague’s findings, fearing personal conflict or professional retaliation (DeCoursey, 2006; Kannan and Gowri, 2014; Nissen et al., 2016). Our experimental validation demonstrates that a significant proportion of unchallenged claims are, in fact, non-reproducible. However, we should note that we did not randomly select the unchallenged claims that were to be experimentally tested, and we favored suspicious claims and claims that were easy to experimentally test. Thus, the true rate of reproducibility of unchallenged claims should not be directly extrapolated from our experimental validation attempts. However, we found that *none* of unchallenged, suspicious claims could be reproduced. While confirmation bias driven by our initial suspicion cannot be entirely excluded, the absence of follow-up in the published literature for over a decade later supports our interpretation that these findings were not robust. Of note, we considered many of these unchallenged claims were considered suspicious based on informal exchanges within research communities, that reported difficulties in reproducing the findings. Most of these non-reproducible suspicious claims were published in high-impact or trophy journals (26/31) with great visibility, perhaps contributing to this privately-shared stigma within the community. Taken together, our findings support the notion that “big stories” published without subsequent follow-up may be indicative of irreproducibility.

Another takeaway from our experimental validation is that approximately 50% of the unchallenged claims were not previously flagged as problematic failed the experimental validation test regardless. While a claim could remain unchallenged for years because it is valid and no one has anything novel to add to it, our results suggest the status of a claim being “unchallenged” is not a reliable proxy for its validity. If we reclassify the dataset by counting logically consistent unchallenged claims as “verified” and logically inconsistent ones as challenged, and conservatively assume that 50% of the remaining unchallenged claims fall into the latter, we estimate that the true proportion of challenged claims may approach 14%. When we include mixed and partially verified claims, a rough upper-bound estimate suggests that at least 80% of all published claims on *Drosophila* immunity between 1959-2011 are verified or likely to be true leaving 20% of claims either challenged or difficult to interpret. We will take this moment to remind that challenged claims can include work that was done well, but where the field has advanced in its understanding and can better interpret the original findings. We believe the percentage of challenged claims likely varies across scientific fields, depending on intrinsic challenges such as the complexity of biological processes, methodological variability, and the field’s attractiveness, which may increase competitive pressure.

In a recent editorial, Fiala and Diamandis, suggested that reproducibility efforts should focus on high-impact papers as these articles are most deleterious to the scientific community (Fiala and Diamandis, 2018). In our database, we indeed found that “trophy journals”, considered to contain the most impactful science, are more likely to publish articles with challenged claims than high or low-impact, although the difference was not significant. Three intertwined factors can explain why high-impact factor journals have higher rate of non-reproducibility. First, the uncertainty that comes with exploring the frontiers of knowledge may lead to investigations that are founded on an incomplete understanding of the major factors influencing the results. Second, the unconventional nature of some high-impact discoveries, and the fact that claims reported in these journals generally attract more attention, increases the likelihood that other researchers will attempt to replicate them — thereby raising the number of challenged claims. And third, systemic biases that pressure scientists to publish in these journals for prestige might encourage scientists to take shortcuts in finalizing their studies, increasing the chance of challenged claims. In our study, we observed a decreased proportion of unchallenged claims in articles published in high-impact and trophy journals indicating that claims published in these journals are more reproduced. However, this higher rate of follow-up work cannot fully explain by itself the higher proportion of challenged claims in trophy journals.

Compared to low rank and trophy journals, high-ranking journals have the best ratio of verified to challenged claims, as low-impact journals have a high percentage of unchallenged claims. Strikingly, we observed higher rate of non-reproducible claims for scientists working in universities in the top 50 Shanghai ranking. Previous work has suggested journal rank as a major associate of irreproducibility (Brembs et al., 2013 [↗](#)), and that high-impact journals favour authors from leading institutions (Kulal et al., 2025 [↗](#)). Our study highlights how university ranking is among the most striking correlates of irreproducibility. Taken together, the cause and effect relationship at play here may be driven most by factors associated with institutional prestige, such as trust among networks of colleagues that act as editors and reviewers, leading to greater acceptance rates at high-impact journals. Importantly, our results suggest this greater acceptance does not strictly reflect a high quality of the work being published, but rather a willingness to trust irreproducible work more frequently if it comes from a prestigious institute. Our study spans a period during which most research was done in North America, Europe, Japan and South Korea and does not capture the rise of science in China, or rapid growth of scientific output across the global north and south, where scientists in some countries face strong incentive for ‘high impact’ publications (Delgado-López-Cózar et al., 2021 [↗](#); Quan et al., 2017 [↗](#)). The fact that the *Drosophila* immunity field monitored in this study developed initially in Sweden and France, which have unique geographic and research community characteristics, might affect some of our conclusions or limit them to a specific period.

Future studies in other areas of research might help to confirm if working in top ranked universities correlates with higher rate of non-reproducible science.

Our data align with the broader narrative of increasing rates of non-reproducible science. One possible explanation is a general decline in the standards of rigor in recent years in the life sciences; however, a non-mutually exclusive hypothesis is that the rise in irreproducibility may reflect the increasing popularity of certain fields, creating greater variance by attracting more researchers, who themselves may not be fully acquainted with the field’s established norms and methodological rigor, or who move opportunistically from field to field to publish striking papers. When research on the gut microbiota gained momentum in the 1990s, it was rapidly positioned as a promising avenue to explain a wide array of diseases. Substantial funding and media attention enabled the swift development of the field. However, many early studies lacked methodological rigor and could not be reproduced (Wesolowska-Andersen et al., 2014 [↗](#)). The lower scientific standards tolerated in this emerging area —compared to more established fields— facilitated the publication of high-profile articles, encouraging a rapid influx of new research groups. In this context, scientific novelty and perceived importance often took precedence over precision and reproducibility. A similar but more acute pattern was observed during the SARS-CoV-2 pandemic. Alternatively, this reproducibility crisis might be more linked to specific features of our time. Although this notion is disputed (Lemaitre, 2016 [↗](#); Twenge and Campbell, 2009 [↗](#)), an increase of narcissism in society might increase the chance of distorting science to favor our own success.

Importantly, our analysis shows that irreproducibility is not evenly distributed across all researchers. While irreproducibility can stem from the inherent difficulty or novelty of certain research questions, our findings point to the significant influence of both the first author (who performs the work) and the laboratory environment. Surprisingly, we did not observe a major effect of experience level: irreproducibility rates were similar between PhD students and postdoctoral researchers and did not markedly decrease with the number of publications. Similarly, we only observe a slightly but not significant higher irreproducibility rate among junior faculty compared to senior professors. This suggests that irreproducibility is not inherent to early career stage but rather to certain research practices and environments. That said, we did observe a higher irreproducibility rate among PIs that entered the field from a non-*Drosophila* immunity background. This observation supports the value of traditional scientific training pathways, where early-career researchers are mentored in well-established environments before gaining independence.

Another noteworthy finding is our suggested link between irreproducibility and “continuous” or “exploratory” research style. We observed that authors with a more exploratory research style have slightly higher proportions of challenged claims and a significantly higher proportion of unchallenged claims, the latter being more likely to be challenged as shown by our experimental replication. Exploratory researchers have worked only transiently in the field, often leaving it after its peak period of popularity. In recent years funding agencies and universities favored this exploratory research style focusing on breakthroughs, notably by funding projects that may be rapidly discontinued. As a counterpoint of this, our study highlights the value and reliability of laboratories that tend to maintain long-term commitment to a specific field.

An important aspiration in the field of metascience study would be the ability to predict claims or articles that are likely irreproducible. Our retrospective analysis identifies several criteria that correlate with irreproducibility, although not all of them reach significance. These include i) articles being published by authors working in top-ranked institutions, ii) articles being published in trophy journals, iii) the absence of follow-up studies (i.e. being ‘unchallenged’), and iv) articles being published by authors with an exploratory research style. In contrast, having published as first author in another laboratory within the field, and publishing in high-impact journals but not trophy journals correlate with higher reproducibility. The absence of follow-up studies regarding claims published in high-impact journals is likely the most useful proxy to question the reproducibility of these claims.

## Limitations

This work has several limitations. First, the definition of the claim, the validity, the determination of the validation was influenced by the annotating authors (HW and BL). While the *ReproSci* website was opened to the community, many PIs were reluctant or not interested in the project of testing old claims that represent “cold scientific cases”. Second the leading author is stakeholder of the project having published several articles, that were included in the analysis, including 44 articles with B.L. as co-author with 4 as first and 22 as last authors. Thus, this may not represent the general view of the community. Finally, the findings might be specific to this field and strongly influenced by specific factors idiosyncratic to its development. However, we believe that despite these limitations the systematic and retrospective analysis of a research field provides insights to better apprehend how science progresses despite errors and the impact of irreproducibility.

## Methods

### Selected articles, annotation, experimental validation of the *ReproSci* project

We provide here a succinct method of the *ReproSci* reproducibility project (see the companion article (Westlake et al., 2025 [DOI](#)) for more details). In brief, a list of 400 publications published before 2011 was generated using a curated search string on the publicly available PubMed database. This limitation was imposed to allow a time frame in which follow-up studies may be published that may verify, expand on or contradict the content of the publications selected for the study. Selected primary articles were annotated by a single researcher (HW) and reviewed by a second researcher (BL) to ensure consistency and accuracy across the project. An online database, *ReproSci* (<https://ReproSci.epfl.ch/> [DOI](#)), was constructed to organize the annotations and provide an interactive and searchable interface. Claims were extracted through careful reading of the selected primary articles and organized according to their importance within the paper: a single **Main claim** representing the title claim of the paper or overarching finding; 3-4 **Major claims** emphasized in the abstract or those contributing to the overarching finding; and a variable number of **Minor claims** representing other findings of note not emphasized in the abstract. Claims were cross-checked with evidence from previous, contemporary and subsequent publications and assigned a verification category (see Table 1 [DOI](#)). A short statement explaining the evidence and citing sources was written to justify the categorization and provide additional context for the claim; this assessment was linked to each claim and can be viewed on the *ReproSci*

website. A subset of unchallenged claims (45 major and 11 minor) was selected for direct experimental reproduction by various authors from different laboratories with relevant expertise. This project is concerned with the accuracy of the conclusions themselves rather than the reliability of methods, so we employed **indirect** or **inferential reproducibility**, where experimental procedures that may be different from those originally employed are used to independently verify the claim. The *ReproSci* website was made available to the *Drosophila* scientific community on July 11, 2023, with an email to encourage to comment on the annotations and verifications and contribute evidence.

## First authors and last author classification

Information regarding the gender and status of first authors were either obtained by interviewing PIs (for 191 first authors) or found on publicly available websites (LinkedIn and university websites) or were implied based on their first name usage (gender) and their PubMed publication lists (for the other 100 first authors). When some of these information could not be retrieved, these authors were not taken in consideration.

PIs were manually classified as senior or junior whether they have or have not published a last author article at least five years before the considered publication. Continuity and exploratory PIs were manually classified according to their PubMed publication list. Continuity PIs are PIs that have continued to work in the *Drosophila* immunity field until now or to the end of their scientific career (deceased, retired, left academia). Exploratory PIs are PIs that have done a transient incursion in the *Drosophila* immunity field moving to other topics of research.

## Statistical analysis

The analysis was done on the 1006 major claims of the 400 articles. The cleaned database contains one row per major claim with linked article-, author- and laboratory-level covariates, which was subsequently analyzed. Every output, counts, proportions and graphics, exploratory summaries and Bayesian mixed-effects models are defined under a reproducible Python workflow, which enable the full replication of the main results and every figure in the manuscript.

The ninety-five-percent confidence intervals (CI) for single proportions were determined using Wilson's score method. Whenever a binary predictor (e.g. author gender) was cross-classified with the binary outcome "claim challenged", we reported the odds ratio with its 95% CI and the two-sided *p*-value from Fisher's exact test; The raw contingency tables for every binary predictor appear in [Supplementary Document](#).

## Multivariable modelling

Two Bayesian regressions were fitted with weakly-informative priors. Posterior summaries are presented as odds ratios (OR) with 94 % highest-density intervals (HDI).

1. Hierarchical logistic regression for claim irreproducibility. The outcome was whether a major claim was later classified as challenged. Fixed-effect predictors captured article prestige, author characteristics and laboratory context: journal tier (Low, High, Trophy), university-ranking tier (Not ranked, 101 +, 51–100, Top 50), publication year modelled with a three-knot B-spline, first-author gender and training stage (PhD vs post-doc), last-author gender, laboratory seniority (junior vs senior PI), continuity style (continuous vs exploratory) and whether the laboratory had published in the field before 1995. Claims from one author not independent, so random intercepts were included for both first-author key and last-author key.
2. Another simpler hierarchical regression model was used with only authors random intercepts, university ranking and journal impact factor to help the conclusion in section [Supplementary Figure S2](#).

All models were estimated with the dynamic-HMC NUTS sampler (target  $\text{accept} = 0.9$ ); four independent chains of 2 000 post-warm-up draws were run for every model. Convergence was confirmed both visually and numerically (all  $R_{\text{hat}} \leq 1.01$ ; all effective sample sizes  $> 400$ ).

Supplementary Fig. S4 shows trace plots, posterior-predictive checks and Pareto-k diagnostics; no systematic lack of fit was detected. Exact formulas, priors, sampling statements and notebooks are available in the public repository (Supplementary Document [↗](#)).

## Acknowledgements

We thank the BioInformatics Competence Center of EPFL-UNIL for generating the *ReproSci* website. This work was supported by Swiss National Science Foundation 310030\_189085 and the ETH-Domain's Open Research Data (ORD) Program (2022). We thank Florent Masson and Samuel Rommelaere for comments on the manuscript.

## Additional files

**Supplemental information** [↗](#) Provides information on the full code (S1), the Multivariate model (S2, Figures S1), the Multivariate model using Unchallenged as outcome (S3, Figure S3), a table of the odd-ratios and values for the categorical variables shown in the main text, for first and last-author (S4) and other categorical comparisons (S5) and Table of journals with impact factor, and claim assessment (S6, Figure S4).

## Additional information

### Funding

| Funder                            | Grant reference number                     |
|-----------------------------------|--------------------------------------------|
| Swiss National Science Foundation | 310030_189085                              |
| ETH Domain                        | Open Research Data (ORD)<br>Program (2022) |

### Author ORCID iDs

**Joseph Lemaitre:** <http://orcid.org/0000-0002-2677-6574>  
**Blandine Ribotta:** <http://orcid.org/0000-0003-3366-6231>  
**Hannah Westlake:** <http://orcid.org/0009-0000-0013-8980>  
**Sveta Chakrabarti:** <http://orcid.org/0000-0002-5591-3718>  
**Mark A Hanson:** <http://orcid.org/0000-0002-6125-3672>  
**Haobo Jiang:** <http://orcid.org/0000-0003-1357-1315>  
**Francesca Di Cara:** <http://orcid.org/0000-0002-6973-3232>  
**Estee Kurant:** <http://orcid.org/0000-0003-4795-4074>  
**Fabrice David:** <http://orcid.org/0000-0001-8539-865X>  
**Bruno Lemaitre:** <http://orcid.org/0000-0001-7970-1667>

## References

- Abril-Pla O**, Andreani V, Carroll C, Dong L, Fonnesbeck CJ, Kochurov M, Kumar R, Lao J, Luhmann CC, Martin OA, *et al.* (2023) PyMC: a modern, and comprehensive probabilistic programming framework in Python. *PeerJ Comput Sci* **9**:e1516 <https://doi.org/10.7717/peerj-cs.1516>
- Amaral**, *et al.* (2025) Estimating the replicability of Brazilian biomedical science. *bioRxiv*
- Amaral OB**, Neves K, Wasilewska-Sampaio AP, Carneiro CF (2019) The Brazilian Reproducibility Initiative. *eLife* **8**:e41602 <https://doi.org/10.7554/eLife.41602>
- Baker M** (2016) 1,500 scientists lift the lid on reproducibility. *Nature* **533**:452-454 <https://doi.org/10.1038/533452a>
- Baker M** (2015) Over half of psychology studies fail reproducibility test. *Nature nature* **2015**:18248 <https://doi.org/10.1038/nature.2015.18248>

- Begley CG, Ioannidis JPA (2015) Reproducibility in Science: Improving the Standard for Basic and Preclinical Research. *Circulation Research* **116**:116-126  
<https://doi.org/10.1161/CIRCRESAHA.114.303819>
- Brembs B, Button K, Munafò M (2013) Deep impact: unintended consequences of journal rank. *Front Hum Neurosci* **7** <https://doi.org/10.3389/fnhum.2013.00291>
- Capretto T, Piho C, Kumar R, Westfall J, Yarkoni T, Martin OA. (2022) Bambi: A simple interface for fitting Bayesian linear models in Python. *arXiv* <https://doi.org/10.48550/arXiv.2012.10754>
- Chalmers I, Glasziou P (2009) Avoidable Waste in the Production and Reporting of Research Evidence. *Lancet* **374**:86-9
- Da Costa GG, Neves K, Amaral OB. (2022) Estimating the replicability of highly cited clinical research (2004-2018) (preprint). *Epidemiology* <https://doi.org/10.1101/2022.05.31.22275810>
- DeCoursey TE (2006) It's difficult to publish contradictory findings. *Nature* **439**:784-784  
<https://doi.org/10.1038/439784b>
- Delgado-López-Cózar E, Ràfols I, Abadal E (2021) Letter: A call for a radical change in research evaluation in Spain. *Epi* <https://doi.org/10.3145/epi.2021.may.09>
- Devezer B, Buzbas EO (2023) Rigorous exploration in a model-centric science via epistemic iteration. *Journal of Applied Research in Memory and Cognition* **12**:189-194 <https://doi.org/10.1037/mac0000121>
- Eisner DA (2018) Reproducibility of science: Fraud, impact factors and carelessness. *Journal of Molecular and Cellular Cardiology* **114**:364-368 <https://doi.org/10.1016/j.yjmcc.2017.10.009>
- Errington TM, Denis A, Perfito N, Iorns E, Nosek BA (2021) Challenges for assessing replicability in preclinical cancer biology. *eLife* **10**:e67995 <https://doi.org/10.7554/eLife.67995>
- Errington TM, Iorns E, Gunn W, Tan FE, Lomax J, Nosek BA (2014) An open investigation of the reproducibility of cancer biology research. *eLife* **3** <https://doi.org/10.7554/eLife.04333>
- Fanelli D (2018) Opinion: Is science really facing a reproducibility crisis, and do we need it to?. *Proceedings of the National Academy of Sciences* **115**:2628-2631  
<https://doi.org/10.1073/pnas.1708272114>
- Fiala C, Diamandis EP (2018) Benign and malignant scientific irreproducibility. *Clinical Biochemistry* **55**:1-2 <https://doi.org/10.1016/j.clinbiochem.2018.03.015>
- Fidler F, Chee YE, Wintle BC, Burgman MA, McCarthy MA, Gordon A (2017) Metaresearch for Evaluating Reproducibility in Ecology and Evolution. *BioScience biw* **159** <https://doi.org/10.1093/biosci/biw159>
- Fidler F, Wilcox J. (2018) Reproducibility of scientific results. In: *The Stanford Encyclopedia of Philosophy* Stanford University.
- Fortunato S, Bergstrom CT, Börner K, Evans JA, Helbing D, Milojević S, Petersen AM, Radicchi F, Sinatra R, Uzzi B, et al. (2018) Science of science. *Science* **359** <https://doi.org/10.1126/science.aao0185>
- França TF, Monserrat JM (2018) Reproducibility crisis in science or unrealistic expectations?. *EMBO reports* **19**:e46008 <https://doi.org/10.15252/embr.201846008>
- Goodman SN, Fanelli D, Ioannidis JPA (2016) What does research reproducibility mean?. *Science Translational Medicine* **8**:341ps12-341ps12 <https://doi.org/10.1126/scitranslmed.aaf5027>
- Hill R, Yin Y, Stein C, Wang X, Wang D, Jones BF (2025) The pivot penalty in research. *Nature* **1-8**  
<https://doi.org/10.1038/s41586-025-09048-1>
- Hunter P (2017) The reproducibility “crisis.”. *EMBO reports* **18**:1493-1496  
<https://doi.org/10.15252/embr.201744876>
- Ioannidis JPA, Allison DB, Ball CA, Coulibaly I, Cui X, Culhane AC, Falchi M, Furlanello C, Game L, Jurman G, et al. (2009) Repeatability of published microarray gene expression analyses. *Nature Genetics* **41**:149-155 <https://doi.org/10.1038/ng.295>
- Kannan S, Gowri S (2014) Contradicting/negative results in clinical research: Why (do we get these)? Why not (get these published)?. *Where (to publish)? Perspect Clin Res* **5**:151-153  
<https://doi.org/10.4103/2229-3485.140546>

- Kulal A N A, Shareena P, Dinesh S (2025) Unmasking Favoritism and Bias in Academic Publishing: An Empirical Study on Editorial Practices. *Public Integrity* 1-22 <https://doi.org/10.1080/10999922.2024.2448875>
- Lemaitre B. (2016) *An Essay on Science and NarcissismC: Why do high-ego personalities drive research in life sciences?* Lausanne: EPFL press.
- Lemaitre B (2004) The road to Toll. *Nature Reviews Immunology* 4:521-527 <https://doi.org/10.1038/nri1390>
- Lesperance DNA, Broderick NA. (2021) Gut Bacteria Mediate Nutrient Availability in Drosophila Diets. *Applied and Environmental Microbiology* 87
- Macleod MR, Michie S, Roberts I, Dirnagl U, Chalmers I, Ioannidis JPA, Salman RA-S, Chan A-W, Glasziou P (2014) Biomedical research: increasing value, reducing waste. *The Lancet* 383:101-104 [https://doi.org/10.1016/S0140-6736\(13\)62329-6](https://doi.org/10.1016/S0140-6736(13)62329-6)
- Meng X-L (2020) Reproducibility, Replicability, and Reliability. *Harvard Data Science Review* 2 <https://doi.org/10.1162/99608f92.dbfce7f9>
- Nissen SB, Magidson T, Gross K, Bergstrom CT (2016) Publication bias and the canonization of false facts. *eLife* 5:e21451 <https://doi.org/10.7554/eLife.21451>
- Peng R (2015) The reproducibility crisis in science: A statistical counterattack. *Significance* 12:30-32 <https://doi.org/10.1111/j.1740-9713.2015.00827.x>
- Piller C (2022) Blots on a field?. *Science* 377:358-363
- Plesser HE (2018) Reproducibility vs. Replicability: A Brief History of a Confused Terminology. *Frontiers in Neuroinformatics* 11 <https://doi.org/10.3389/fninf.2017.00076>
- Prinz F, Schlange T, Asadullah K (2011) Believe it or not: how much can we rely on published data on potential drug targets?. *Nature Reviews Drug Discovery* 10:712-712 <https://doi.org/10.1038/nrd3439-c1>
- Quan W, Chen B, Shu F (2017) Publish or impoverish: An investigation of the monetary reward system of science in China (1999-2016). *Ajim* 69:486-502 <https://doi.org/10.1108/ajim-01-2017-0014>
- Rosenthal R (1979) The file drawer problem and tolerance for null results. *Psychological Bulletin* 86:638-641 <https://doi.org/10.1037/0033-2909.86.3.638>
- Shiffrin RM, Börner K, Stigler SM (2018) Scientific progress despite irreproducibility: A seeming paradox. *Proceedings of the National Academy of Sciences* 115:2632-2639 <https://doi.org/10.1073/pnas.1711786114>
- Steneck NH (2011) The dilemma of the honest researcher. *EMBO reports* 12:745-745 <https://doi.org/10.1038/embor.2011.134>
- Stern AM, Casadevall A, Steen RG, Fang FC (2014) Financial costs and personal consequences of research misconduct resulting in retracted publications. *eLife* 3:e02956 <https://doi.org/10.7554/eLife.02956>
- Twenge JM, Campbell WK (2009) *The narcissism epidemic: Living in the age of entitlement* Simon and Schuster.
- Udesky L (2025) ‘Publish or perish’ culture blamed for reproducibility crisis. *Nature* <https://doi.org/10.1038/d41586-024-04253-w>
- van Wesel M. (2016) Evaluation by Citation: Trends in Publication Behavior, Evaluation Criteria, and the Strive for High Impact Publications. *Sci Eng Ethics* 22:199-225 <https://doi.org/10.1007/s11948-015-9638-0>
- Wesolowska-Andersen A, Bahl MI, Carvalho V, Kristiansen K, Sicheritz-Pontén T, Gupta R, Licht TR (2014) Choice of bacterial DNA extraction method from fecal material influences community structure as evaluated by metagenomic analysis. *Microbiome* 2:19 <https://doi.org/10.1186/2049-2618-2-19>

Westlake H, David FPA, Tian Y, Krakovic K, Dolgikh A, Juravlev L, Esmangart de Bournonville T, Carboni A, Melcarne C, Shan T, *et al.* (2025) Reproducibility of Scientific Claims in *Drosophila* Immunity: A Retrospective Analysis of 400 Publications. *bioRxiv*

Westlake H, Hanson MA, Lemaitre B. (2024) *The Drosophila Immunity Handbook* EPFL press.

Youyou W, Yang Y, Uzzi B (2023) A discipline-wide investigation of the replicability of Psychology papers over the past two decades. *Proc Natl Acad Sci USA* **120**:e2208863120  
<https://doi.org/10.1073/pnas.2208863120>

Zeng A, Shen Z, Zhou J, Fan Y, Di Z, Wang Y, Stanley HE, Havlin S. (2019) Increasing trend of scientists to switch between topics. *Nat Commun* **10**:3439 <https://doi.org/10.1038/s41467-019-11401-8>

## Peer reviews

### Reviewer #1 (Public review):

#### Summary:

The authors set out on the ambitious task of establishing the reproducibility of claims from the *Drosophila* immunity literature. Starting out from a corpus of 400 articles from 1959 and 2011, the authors sought to determine whether their claims were confirmed or contradicted by previous or subsequent publications. Additionally, they actively sought to replicate a subset of the claims for which no previous replications were available (although this set was not representative of the whole sample, as the authors focused on suspicious and/or easily testable claims). The focus of the article is on inferential reproducibility; thus, methods don't necessarily map exactly to the original ones.

The authors present a large-scale analysis of the individual replication findings, which are presented in a companion article (Westlake *et al.*, 2025. DOI 10.1101/2025.07.07.663442). In their retrospective analysis of reproducibility, the authors find that 61% of the original claims were verified by the literature, 7.5% were partially verified, and only 6.8% were challenged, with 23.8% having no replication available. This is in stark contrast with the result of their prospective replications, in which only 16% of claims were successfully reproduced.

The authors proceed to investigate correlates of replicability, with the most consistent finding being that findings stemming from higher-ranked universities (and possibly from very high impact journals) were more likely to be challenged.

#### Strengths:

(1) The work presents a large-scale, in-depth analysis of a particular field of science that includes authors with deep domain expertise of the field. This is a rare endeavour to establish the reproducibility of a particular subfield of science, and I'd argue that we need many more of these in different areas.

(2) The project was built on a collaborative basis (<https://ReproSci.epfl.ch/>), using an online database (<https://ReproSci.epfl.ch/>), which was used to organize the annotations and comments of the community about the claims. The website remains online and can be a valuable resource to the *Drosophila* immunity community.

(3) Data and code are shared in the authors' GitHub repository, with a Jupyter notebook available to reproduce the results.

#### Main concerns:

(1) Although the authors claim that "Drosophila immunity claims are mostly replicable", this conclusion is strictly based on the retrospective analysis - in which around 84% of the claims

for which a published verification attempt was found. This is in very stark contrast with the findings that the authors replicate prospectively, of which only 16% are verified.

Although this large discrepancy may be explained by the fact that the authors focused on unchallenged and suspicious claims (which seems to be their preferred explanation), an alternative hypothesis is that there is a large amount of confirmation bias in the *Drosophila* immunity literature, either because attempts to replicate previous findings tend to reach similar results due to researcher bias, or because results that validate previous findings are more likely to be published.

Both explanations are plausible (and, not being an expert in the field, I'd have a hard time estimating their relative probability), and in the absence of prospective replication of a systematic sample of claims - which could determine whether the replication rate for a random sample of claims is as high as that observed in the literature -, both should be considered in the manuscript.

(2) The fact that the analysis of factors correlating with reproducibility includes both prospective and retrospective replications also leads to the possibility of confusion bias in this analysis. If most of the challenged claims come from the authors' prospective replications, while most of the verified ones come from those that were replicated by the literature, it becomes unclear whether the identified factors are correlated with actual reproducibility of the claims or with the likelihood that a given claim will be tested by other authors and that this replication will be published.

(3) The methods are very brief for a project of this size, and many of the aspects in determining whether claims were conceptually replicated and how replications were set up are missing.

Some of these - such as the PubMed search string for the publications and a better description of the annotation process - are described in the companion article, but this could be more explicitly stated. Others, however, remain obscure. Statements such as "Claims were cross-checked with evidence from previous, contemporary and subsequent publications and assigned a verification category" summarize a very complex process for which more detail should be given - in particular because what constitutes inferential reproducibility is not a self-evident concept. And although I appreciate that what constitutes a replication is ultimately a case-by-case decision, a general description of the guidelines used by the authors to determine this should be provided. As these processes were done by one author and reviewed by another, it would also be useful to know the agreement rates between them to have a general sense of how reproducible the annotation process might be.

The same gap in methods descriptions holds for the prospective replications. How were labs selected, how were experimental protocols developed, and how was the validity of the experiments as a conceptual replication assessed? I understand that providing the methods for each individual replication is beyond the scope of the article, but a general description of how they were developed would be important.

(4) As far as I could tell, the large-scale analysis of the replication results was not preregistered, and many decisions seem somewhat ad hoc. In particular, the categorization of journals (e.g. low impact, high impact, "trophy") and universities (e.g. top 50, 51-100, 101+) relies on arbitrary thresholds, and it is unclear how much the results are dependent on these decisions, as no sensitivity analyses are provided.

Particularly, for analyses that correlate reproducibility with continuous variable (such as year of publication, impact factor or university ranking, I'd strongly favor using these variables as continuous variables in the analysis (e.g. using logistic regression) rather than performing pairwise comparisons between categories determined by arbitrary cutoffs. This would not only reduce the impact of arbitrary thresholds in the analysis, but would also

increase statistical power in the univariate analyses (as the whole sample can be used in at once) and reduce the number of parameters in the multivariate model (as they will be included as a single variable rather than multiple dummy variables when there are more than two categories).

(5) The multivariate model used to investigate predictors of replicability includes unchallenged claims along with verified ones in the outcome, which seems like an odd decision. If the intention is to analyze which factors are correlated with reproducibility, it would make more sense to remove the unchallenged findings, as these are likely uninformative in this sense. In fact, based on the authors' own replications of unchallenged findings, they may be more likely to belong the "challenged" category than to the "unchallenged" one if they were to be verified.

<https://doi.org/10.7554/eLife.108403.1.sa2>

## Reviewer #2 (Public review):

### Summary:

Lemaitre et al. conducted an analysis of 400 publications in the *Drosophila* immunity field (1959-2011), performing both univariable and multivariable analyses to identify factors that correlate with or influence the irreproducibility of scientific claims. Some of the findings are unexpected, for instance, neither the career stage of the PI nor that of the first author appears to matter that much, while others, such as the influence of institutional prestige or publication in "trophy journals," are more predictable. The results provide valuable insight into patterns of irreproducibility in academia and may help inform policies to improve research reproducibility in the field.

### Strengths:

This study is based on a large, manually curated dataset, complemented by a companion paper (Westlake et al., 2025. DOI 10.1101/2025.07.07.663442) that provides additional details on experimentally documented cases. The statistical methods are appropriate, and the findings are both important and informative. The results are clearly presented and supported by accessible documentation through the ReproSci project.

### Weaknesses:

The analysis is limited to a specific field (immunity) and model system (*Drosophila*). Since biological context may influence reproducibility -- for example, depending on whether mechanisms are more hardwired or variable -- and the model system itself may contribute to these effects (as the authors note), it remains unclear to what extent these findings generalize to other fields or organisms. The authors could expand the discussion to address the potential scope and limitations of the study's generalizability.

<https://doi.org/10.7554/eLife.108403.1.sa1>

## Reviewer #3 (Public review):

### Summary:

The authors of this paper were trying to identify how reproducible, or not, their subfield (*Drosophila* immunity) was since its inception over 50 years ago. This required identifying not only the papers, but the specific claims made in the paper, assessing if these claims were followed up in the literature, and if so whether the subsequent papers supported or refuted the original claim. In addition to this large manually curated effort, the authors further investigated some claims that were left unchallenged in the literature by conducting

replications themselves. This provided a rich corpus of the subfield that could be investigated into what characteristics influence reproducibility.

#### Strengths:

A major strength of this study is the focus on a subfield, the detailing of identifying the main, major, and minor claims - which is a very challenging manual task - and then cataloging not only their assessment of if these claims were followed up in the literature, but also what characteristics might be contributing to reproducibility, which also included more manual effort to supplement the data that they were able to extract from the published papers. While this provides a rich dataset for analysis, there is a major weakness with this approach, which is not unique to this study.

#### Weaknesses:

The main weakness is relying heavily on the published literature as the source for if a claim was determined to be verified or not. There are many documented issues with this stemming from every field of research - such as publication bias, selective reporting, all the way to fraud. It's understandable why the authors took this approach - it is the only way to get at a breadth of the literature - however the flaw with this approach is it takes the literature as a solid ground truth, which it is not. At the same time, it is not reasonable to expect the authors to have conducted independent replications for all of the 400 papers they identified. However, there is a big difference trying to assess the reproducibility of the literature by using the literature as the 'ground truth' vs doing this independently like other large-scale replication projects have attempted to do. This means the interpretation of the data is a bit challenging.

Below are suggestions for the authors and readers to consider:

(1) I understand why the authors prefer to mention claims as their primary means of reporting what they found, but it is nested within paper, and that makes it very hard to understand how to interpret these results at times. I also cannot understand at the high-level the relationship between claims and papers. The methods suggest there are 3-4 major claims per paper, but at 400 papers and 1,006 claims, this averages to ~2.5 claims per paper. Can the authors consider describing this relationship better (e.g., distribution of claims and papers) and/or considering presenting the data two ways (primary figures as claims and complimentary supplementary figures with papers as the unit). This will help the reader interpret the data both ways without confusion. I am also curious how the results look when presented both ways (e.g., does shifting to the paper as the unit of analysis shift the figures and interpretation?). This is especially true since the first and last author analysis shows there is varying distribution of papers and claims by authors (and thus the relationship between these is important for the reader).

(2) As mentioned above, I think the biggest weakness is that the authors are taking the literature at face value when assigning if a claim was validated or challenged vs gathering new independent evidence. This means the paper leans more on papers, making it more like a citation analysis vs an independent effort like other large-scale replication projects. I highly recommend the authors state this in their limitations section.

On top of that, I have questions that I could not figure out (though I acknowledge I did not dig super deep into the data to try). The main comment I have is How was verified (and challenged) determined? It seems from the methods it was determined by "Claims were cross-checked with evidence from previous, contemporary and subsequent publications and assigned a verification category". If this is true, and all claims were done this way - are verified claims double counted then? (e.g., an original claim is found by a future claim to be verified - and thus that future claim is also considered to be verified because of the original claim).

Related, did the authors look at the strength of validation or challenged claims? That is, if there is a relationship mapping the authors did for original claims and follow-up claims, I would imagine some claims have deeper (i.e., more) claims that followed up on them vs others. This might be interesting to look at as well.

(3) I recommend the authors add sample sizes when not present (e.g., Fig 4C). I also find that the sample sizes are a bit confusing, and I recommend the authors check them and add more explanation when not complete, like they did for Fig 4A. For example, Fig 7B equals to 178 labs (how did more than 156 labs get determined here?), and yet the total number of claims is 996 (opposed to 1,006). Another example, is why does Fig 8B not have all 156 labs accounted for? (related to Fig 8B, I caution on reporting a p value and drawing strong conclusions from this very small sample size - 22 authors). As a last example, Fig 8C has all 156 labs and 1,006 claims - is that expected? I guess it means authors who published before 1995 (as shown in Figure 8A continued to publish after 1995?) in that case, it's all authors? But the text says when they 'set up their lab' after 1995, but how can that be?

(4) Finally, I think it would help if the authors expanded on the limitations generally and potential alternative explanations and/or driving factors. For example, the line "though likely underestimated" is indicated in the discussion about the low rate of challenged claims, it might be useful to call out how publication bias is likely the driver here and thus it needs to be carefully considered in the interpretation of this. Related, I caution the authors on overinterpreting their suggestive evidence. The abstract for example, states claims of what was found in their analysis, when these are suggestive at best, which the authors acknowledge in the paper. But since most people start with the abstract, I worry this is indicating stronger evidence than what the authors actually have.

The authors should be applauded for the monumental effort they put into this project, which does a wonderful job of having experts within a subfield engage their community to understand the connectiveness of the literature and attempt to understand how reliable specific results are and what factors might contribute to them. This project provides a nice blueprint for others to build from as well as leverage the data generated from this subfield, and thus should have an impact in the broader discussion on reproducibility and reliability of research evidence.

<https://doi.org/10.7554/eLife.108403.1.sa0>