Integrating Machine Learning Standards in Disseminating Machine Learning Research

RESEARCH PAPER

SCOTT C. EDMUNDS
NICOLE NOGOY
QING LAN
HONGFANG ZHANG
YANNAN FAN
HONGLING ZHOU
CHRIS ARMIT

*Author affiliations can be found in the back matter of this article

ABSTRACT

The increasing use of AI-based approaches such as machine learning (ML) across diverse scientific fields presents challenges for reproducibly disseminating and assessing research. As ML becomes integral to a growing range of computationally intensive applications (e.g., clinical research), there is a critical need for transparent reporting methods to ensure both comprehensibility and the reproducibility of the supporting studies. There are a growing number of standards, checklists, and guidelines enabling more standardised reporting of ML research, but the proliferation and complexity of these make them challenging to use—particularly in assessment and peer review, which has, to date, been an *ad hoc* process that has struggled to throw light on increasingly complicated computational supporting methods that are otherwise unintelligible to other researchers. Taking the publication process beyond these black boxes, GigaScience Press has experimented with integrating many of these ML standards into the publication process. Having a broad scope necessitated looking at more generalist and automated approaches. Here, we map the current landscape of artificial intelligence (AI) standards and outline our adoption of the Data, Optimization, Model, Evaluation (DOME) recommendations for ML in biology. We developed a publishing workflow that integrates the DOME Data Stewardship Wizard (DOME-DSW) and DOME Registry tools into the peer review and publication process. From this publisher's case study, we provide journal authors, reviewers, and editors with examples of approaches, workflows, and strategies to more logically disseminate and review ML research. This demonstrates the need for continued dialogue and collaboration among various ML communities to create unified, comprehensive standards and to enhance the credibility, sustainability, and impact of ML-based scientific research.

CORRESPONDING AUTHOR:
Scott C. Edmunds
GigaScience Press, BGI Hong Kong Tech Co Ltd., HK SAR
s.c.edmunds@gmail.com

KEYWORDS:
reproducibility; machine learning; metadata standards; information standards

TO CITE THIS ARTICLE:
Edmunds, S.C., Nogoy, N., Lan, Q., Zhang, H., Fan, Y., Zhou, H. and Armit, C. 2026 Integrating Machine Learning Standards in Disseminating Machine Learning Research. *Data Science Journal*, 25: 1, pp. 1–12. DOI: https://doi.org/10.5334/dsj-2026-001

INTRODUCTION

In recent years, there has been a substantial increase in scientific publications utilising artificial intelligence (AI) techniques such as machine learning (ML) (Figure 1). This represents a significant challenge for disseminating and assessing scientific research, as ML is increasingly a component in scientific publications on diverse research areas such as genomics, drug discovery, remote sensing, medical imaging, and health informatics. As ML is increasingly applied to studies with clinical importance, there is a need for reporting methods to enable a researcher to understand the ML approach used in a research study, and also improve the quality of evidence coming out of these studies for eventual practical application. Furthermore, there has been acknowledgement in the literature that there is a lack of rigorous clinical trials and data validation for the many approved AI medical devices now entering the market (Lenharo, 2024).

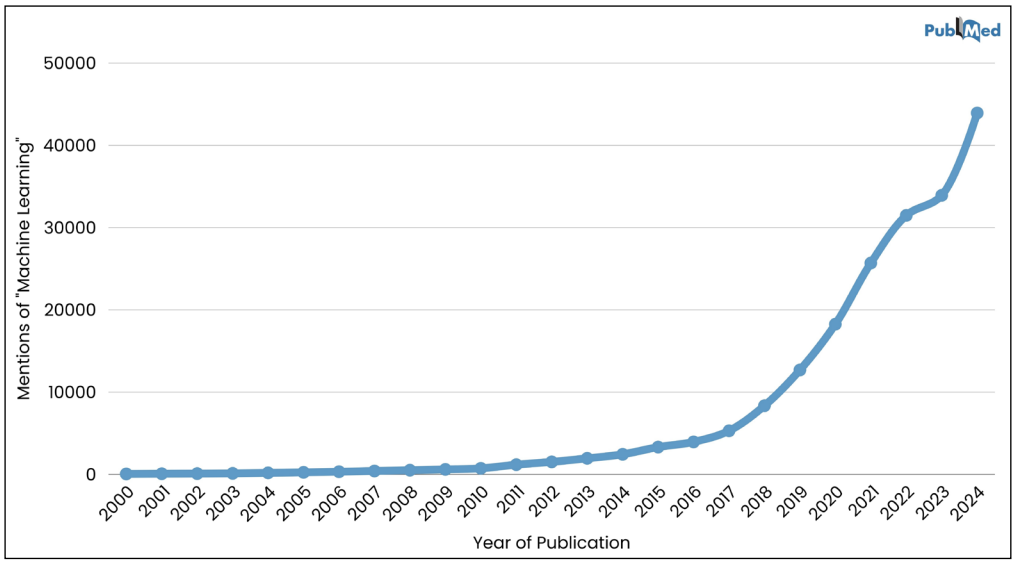

Figure 1 Timeline of publications in PubMed that include the phrase **'machine learning'** since 2000, demonstrating the rapid rise in these publications across the globe. In 2024 alone, PubMed published a total of 43,931 articles covering **'machine learning'**.

As a publisher of data-intensive computational research, the GigaScience Press journal *GigaScience* has seen an equally rapid rise in the number of submissions utilising ML since the journal's launch in 2012 (Figure 2).

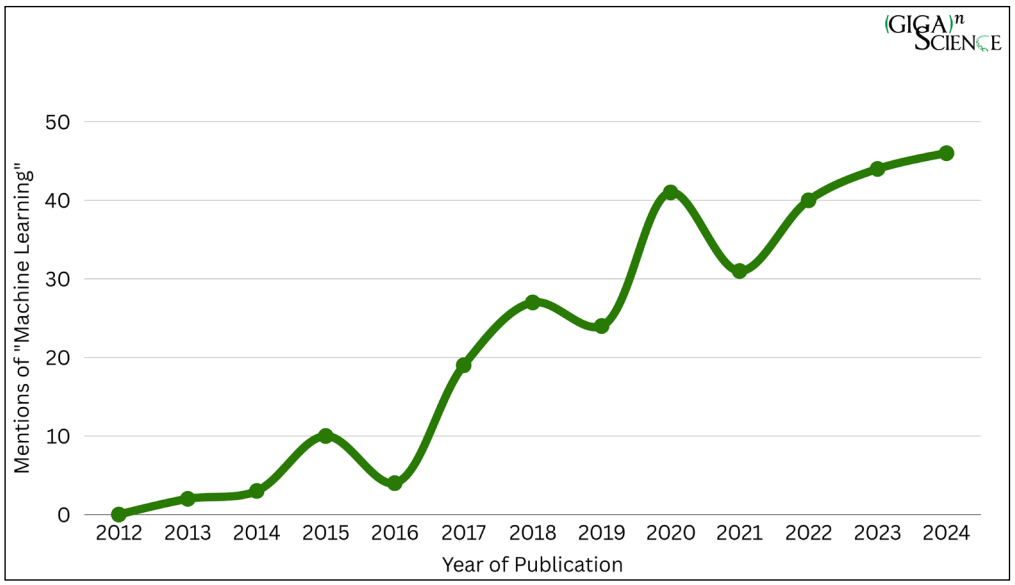

Figure 2 Timeline of publications in *GigaScience* that include the phrase 'machine learning' since the launch of the journal in 2012.

There are myriad reasons for this deluge of ML publications, alongside a wider range of researchers becoming more advanced in computer literacy and usage, summed up in a commentary by David Jones on 'Setting the standards for machine learning in biology' (Jones, 2019), where he stated:

> First, the technology itself has become trivial to use by any reasonably competent programmer, owing to the availability of software that has made it possible for anyone to carry out deep learning experiments, which would have been difficult even for experienced computer scientists just a few years ago.

With further usage beyond the traditional computer science user base, this ease of accessing ML software has necessitated the need for reporting standards to ensure that ML methods are intelligible to other researchers. The black box and often proprietary (non-open source) nature of ML products might potentially make intelligibility even more challenging than traditional computational research. Due to this complexity, in academic publishing, the practical considerations for supporting these ML products have, to date, been an afterthought—if they are addressed at all.

Beyond the academic and the clinical context, commercial internet and AI companies are being pushed to be more transparent in how AI technology works. With this in mind, Google has proposed Model Cards for model reporting, detailing the performance characteristics of ML models that provide benchmarked evaluation in a variety of conditions that are relevant to the intended application domains (Mitchell *et al.*, 2019). In addition, these Model Cards disclose the context in which models are intended to be used and detail performance evaluation procedures alongside other relevant information, providing a summarised and more digestible version of detailed, academic-style technical reports. The Model Cards checklist has been adopted by many users, including the HuggingFace model hub and community for ML (Hugging Face Model Cards), and has an accompanying Dataset Cards checklist to help users understand the contents of the equally essential training data (Hugging Face Dataset Cards). Another community-driven approach integrated into ML dataset repositories like HuggingFace and MLCommons is the Croissant community-driven metadata vocabulary for describing ML datasets that builds on Schema.org (Akhtar *et al.*, 2024). One criticism is that these general standards do not meet the requirements of clinical or basic and translational academic research. On top of benchmarking, researchers in the life sciences and other more applied fields are also interested in the provenance, reproducibility, and reusability of the models and data (Mitchell *et al.*, 2019).

MAPPING THE LANDSCAPE OF ML STANDARDS

In this section, we review the literature and examine the landscape of field-specific ML standards, and then move to look at more general multidisciplinary approaches. To address the challenges of transparency and standardisation, there have been a number of efforts for creating checklists, reporting standards and submission guidelines in ML—particularly in medical AI, where at least 26 reporting guidelines have been published between 2009 and 2023 (Kolbinger *et al.*, 2024). These have been very field-specific, such as the Checklist for Artificial Intelligence in Medical Imaging (CLAIM), which is only relevant to address applications of AI in medical imaging that include classification, image reconstruction, text analysis, and workflow optimisation (Mongan, Moy and Kahn, 2020). Or there is the R-AI-DIOLOGY checklist for evaluation of AI tools specifically in clinical neuroradiology (Haller *et al.*, 2022), and the Proposed Requirements for Cardiovascular Imaging-Related Machine Learning Evaluation (PRIME) checklist for Cardiovascular Imaging (Sengupta *et al.*, 2020). There are Minimal Information checklists for Clinical Artificial Intelligence Modeling (MI-CLAIM) (Norgeot *et al.*, 2020), and reporting standards for AI in health-care (MINIMAR) (Hernandez-Boussard *et al.*, 2020). Both of these follow the style of the Minimum Information for Biological and Biomedical Investigation (MIBBI) initiative, but with the performance parts focusing on the clinical utility of the models. There have been efforts to define standards for ML conference submissions, with the Neural Information Processing Systems (NeurIPS) 2019 reproducibility programme introducing a reproducibility checklist for submissions to the NeurIPS conference (Pineau *et al.*, 2022).

The increasing use of ML in clinical applications has led to updating a number of the reporting guidelines available in the global EQUATOR (Enhancing the QUAlity and Transparency Of health Research) Network that maintains these guidelines to improve the reliability and value of published health research literature (see: https://www.equator-network.org). These include new extensions for clinical trial protocols (SPIRIT-AI) (Rivera *et al.*, 2020) and clinical trials (CONSORT-AI) (Liu *et al.*, 2020), and development of tools used for assistance in decision-

making, such as DECIDE-AI (Vasey *et al.*, 2022). The TRIPOD (Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis) statement was another early set of recommendations for the reporting of studies developing, validating, or updating a prediction model, whether for diagnostic or prognostic purposes (Collins *et al.*, 2015). This has been recently updated as the TRIPOD+AI statement to now work for reporting clinical prediction models that use ML (Collins *et al.*, 2024), with an even newer additional extension addressing the unique challenges of Large Language Models (LLMs) in biomedical applications (Gallifant *et al.*, 2025).

The use of reporting guidelines has been long-established in clinical research, and in non-clinical fields, less of these types of ML guidelines have been developed, but more general best practice recommendations are still being published in some fields—for example, for ML in chemistry (Artrith *et al.*, 2021) and antibody discovery and development (Wossnig *et al.*, 2024). Beyond these field-specific efforts and more general guidelines, there have been a few attempts to create more general checklists and standards, and these are of more practical use to authors, editors, and reviewers. The broader bioimaging community has created a specific checklist for ML workflows amongst their suite of community-developed checklists for publishing images and image analyses (Schmied *et al.*, 2024). Individual journals such as *NPJ Digital Medicine* have also published journal-specific guidance, and the wider Nature portfolio has started to use an internally developed checklist for relevant submissions. With this growing interest, to avoid silos and having to reinvent the wheel, there is an obvious need to have wider and broader guidelines that work across journals and fields (Kakarmath *et al.*, 2020). Funders have also been involved in laying the ground rules, with the NIH Common Fund's Bridge to Artificial Intelligence (Bridge2AI) Consortium presenting their own best practices for ML analysis (Clark *et al.*, 2024). More general multidisciplinary guidelines providing a list of reporting items to be included in a research article for developing predictive models were developed nearly a decade ago (Luo *et al.*, 2016), but in light of how much the field has progressed, these guidelines are in need of updating, especially as usage of ML also goes much beyond predictive models and biomedical research.

Attempting to distil all of these many efforts into an overarching set of general guidelines for publishing is extremely challenging, especially as recent efforts to map the different guidelines, platforms, and tools involved in implementing AI research in the life sciences alone has listed over 300 components of the 'AI ecosystem' (see: https://dome-ml.org/ai-ecosystem) (Farrell *et al.*, 2025). Focusing on the most broad and multidisciplinary approaches, several groups have proposed more general ML standards based on publishing data, models, and code. Several of these standards have been published in a 2021 special issue of *Nature Methods* on 'Reporting standards for machine learning in biology' (Nature Methods, 2021), followed more recently by the publication of the REFORMS consensus-based recommendation (Kapoor *et al.*, 2024). Here, we go into more detail on these generalist approaches, and present the first-hand experiences of a publisher attempting to implement these ML standards into our peer review and publication workflows so that others may gain insight into this process.

REFORMS: CONSENSUS-BASED RECOMMENDATIONS FOR ML-BASED SCIENCE

The latest attempt to provide broader guidelines is the REFORMS checklist, created through obtaining a consensus of 19 researchers across the fields of computer science, data science, mathematics, social sciences, and biomedical sciences (Kapoor *et al.*, 2024). This consensus resulted in the creation of a very broad field-agnostic checklist that consists of 32 items across eight modules. The REFORMS checklist attempts to tackle some of the issues not covered in older checklists, such as data leakage—including paired release guidelines for each of the checklist points. Being consensus-based, these recommendations are designed to also inform readers of best practices and provide a stance on how certain research activities should be conducted.

REFORMS is designed to be applicable to what authors refer to as 'ML-based science', that is, scientific research that uses ML models to contribute to scientific knowledge. These types of studies include those using ML in making predictions, conducting measurements, or performing other tasks that help answer scientific questions of interest. The scope of this checklist encompasses ML techniques, such as supervised, unsupervised, and reinforcement learning, but REFORMS does not work for ML methods research and predictive analytics, which fall outside the checklist's scope. Thus, other standards need to be applied for these types of studies.

MACHINE LEARNING REPRODUCIBILITY AND THE GOLD, SILVER, AND BRONZE STANDARDS

The most general of the *Nature Methods* special issues publications was by Heil *et al*. (2021), who proposed a three-tier system based on the ease of reproducibility. The top-level tier, or 'Gold standard', represents full automation, but is difficult to achieve and may be a more theoretical and aspirational end goal to eventually work towards. The middle-tier, or 'Silver standard', utilises containerisation in the form of Docker images or something similar, which enables swift installation of the dependencies necessary to run an analysis. The lowest tier, or 'Bronze standard', is the scenario where the authors make the data, models, and code used in the analysis publicly available. The 'Bronze standard' is the minimal standard for reproducibility.

To incentivise the uptake of the Gold/Silver/Bronze standard by scientific publishers, the authors of Heil *et al*. include the following statement:

> Journals can enforce reproducibility standards as a condition of publication. The bronze standard should be the minimal standard, though some journals may wish to differentiate themselves by setting higher standards. Such journals may require the silver or gold standards for all manuscripts, or for particular classes of articles such as those focused on analysis. If journals act as the enforcing body for reproducibility standards, they can verify that the standards are met by either requiring reviewers to report which standards the work meets or by including a special reproducibility reviewer to evaluate the work.

AIMe2021 STANDARDS AND REGISTRY FOR AI IN BIOMEDICAL RESEARCH

The core focus of the Artificial Intelligence in Medicine (AIMe) Consortium is to provide a community-driven registry that allows authors to generate citable reports that enable a deeper understanding of ML models. Funded by the EU Horizon 2020 programme, the mission of the AIMe initiative is 'to promote open, transparent and reproducible biomedical AI research' (Matschinske *et al*., 2021). To this end, the AIMe Consortium has provided the community-driven AIMe2021 Standard (the specification available at https://aime-registry.org/specification/)—a report-generating tool, and the AIMe Registry where researchers can post their reports on ML models.

The AIMe2021 Standard included the following categories: Metadata, Purpose, Data, Method, and Reproducibility. In addition, there are optional sections in the AIMe2021 Standard for Privacy (information related to privacy-relevant methods) and Epistasis-related methods (genetics research looking at interactions between genes).

Following completion of a report, AIMe Registry HTML reports are assigned a unique AIMe identifier in the AIMe Registry Database. The intention of AIMe is 'to enhance the accessibility, reproducibility and usability of biomedical AI models, and allow future revisions by the community'. To this end, a notable feature of the AIMe Registry is the ability to comment on existing reports.

TRUST AND TRANSPARENCY AND THE DOME RECOMMENDATIONS

An alternative focus for ML standards is 'Trust and Transparency' rather than *de facto* reproducibility. In this scenario, researchers are invited to provide sufficient detail to enable a computational biologist to understand the supervised ML approach used in an *in silico* analysis. This has been the focus of the DOME (Data, Optimisation, Model, Evaluation) Consortium, which has followed a similar approach to the AIMe Consortium by developing a standard and registry (Walsh *et al*., 2021). The DOME recommendations were initially formulated through the ELIXIR-funded Machine Learning Focus Group (ELIXIR ML) after the publication of the aforementioned commentary by Jones (2019), which called for the establishment of standards for ML in biology. Like AIMe, this community was supported by the EU Horizon 2020 programme, and these standards were designed for supervised ML-based analyses specifically applied to biological studies. ELIXIR ML community members predominantly use ML to analyse large omics datasets—for example, the use of these tools by the ELIXIR Intrinsically Disordered Proteins (IDP) Community to advance the understanding of protein science.

The DOME recommendations are presented in a format consisting of questions, and are defined as the minimal requirements to ML implementers, in order to ensure reliability and reproducibility of the reported methods. The DOME recommendations are one output from the ELIXIR ML Focus Group, and are an actionable outcome usable by the wider scientific community, including researchers, publishers, funders, and policy makers. Beyond the ELIXIR ML group, the DOME recommendations have been further interpreted at the domain level, such as the use and reporting of ML in proteomics and metabolomics (Palmblad *et al.*, 2022). The DOME recommendations for supervised ML in biology were published in the 2021 *Nature Methods* special issue (Walsh *et al.*, 2021), and an updated report on the DOME Registry has recently been published, demonstrating that this work is ongoing (Attafi *et al.*, 2024). A brief summary of the DOME recommendations are below.

Data

- There is a focus in this section on data provenance, data splits (test, training, validation), redundancy between data splits, and the open availability of the underlying data.

Optimization

- This section focuses on the ML algorithm that is used. In addition, meta-predictions wherein the model utilises data from other ML algorithms as input, are also to be detailed in this section. Optimization additionally includes details of how data were encoded and processed by the ML algorithm, parameters used in the model, details of fitting and how overfitting/underfitting were ruled out, and details of availability of configuration.

Model

- This section details interpretability. Specifically, authors are invited to state whether the model is a black box or interpretable, and whether the output represents a classification model or a regression model. In addition, details of execution time and software availability, including details of where the source code is released, are included in this section.

Evaluation

- This section includes details of the evaluation method—for example, cross-validation, independent dataset, or novel experiments. Evaluation additionally includes details of performance metrics, comparison to publicly available methods or simpler baselines, and confidence intervals for performance metrics.

The DOME Consortium has also integrated a DOME Wizard (https://dome.dsw.elixir-europe.org) into the ELIXIR Data Stewardship Wizard (DSW) tool (Pergl *et al.*, 2019) that enables researchers to submit their DOME annotations to the central DOME repository, known as the DOME Registry (https://registry.dome-ml.org/). Particularly relevant for the scientific publishing community, a key feature of the DOME Wizard is that it allows authors to share draft DOME annotations with reviewers. And by assigning identifiers and DOME scores to publications, the registry fosters a standardised evaluation of ML methods that can be useful for assessment and peer review (Attafi *et al.*, 2024).

Table 1 summarises the different general, non-field-specific, ML standards that have been published, highlighting the scope and strengths and weaknesses of the different approaches. The engaging community behind it that was carrying out outreach to publishers, the broad but not overly complicated scope, the ease of use of its automated wizard, and active and growing registry are the reason for *GigaScience* testing DOME-ML further.

THE ML PUBLISHING USE CASE: INTEGRATING THE DOME RECOMMENDATIONS INTO A JOURNAL WORKFLOW

As a case study to show how ML standards could be incorporated into scientific publishing, the following example highlights the integration of DOME recommendations into the data and software publishing workflow of GigaScience Press. This has specifically been carried out by both *GigaScience* and *GigaByte* journals to help deal with their growing number of submissions using

Standard	DOME-ML (Walsh *et al.*, 2021)	REFORMS (Kapoor *et al.*, 2024)	AIMe (Matschinske *et al.*, 2021)	Bronze-Silver-Gold (Heil *et al.*, 2021)
Complexity (No. of fields)	17 recommendations (10 required)	32 questions (8 modules)	63 questions (6 optional)	3 standards (7 criteria)
Extensions	No	No	Yes	No
Tools	Recommendations + Wizard + Registry	Checklist	Standard + Reporting tool + Registry	Standard
Focus	Supervised ML	For ML-based science, not ML methods research	Biomedical AI	Machine-learning analyses in the life sciences

Table 1 Summary of the more general ML standards that go beyond medicine to biological science and beyond, comparing the DOME-ML, REFORMS, AIMe and Heil *et al.*'s Bronze-Silver-Gold approaches.

ML-methods, the first example being in March 2023 (Guo *et al.*, 2023). At the time of writing, there have been 55 submissions that have used this approach for assessing ML research, with the 53 from *GigaScience* making up 19% of the content in the DOME Registry.

The provision of a 'Big Data' archiving solution to support omics and other large-scale biological and biomedical data is a core tenet of GigaScience Press, and it is the role of the GigaScience DataBase (GigaDB) Curation team to ensure that the Data Submission process runs as smoothly as possible (Armit, Tuli and Hunter, 2022). In addition to Data Archiving, Data Curators liaise with authors and advise on the appropriate Data Standards to use for what can be quite complex studies (e.g., genomic, transcriptomic, digital pathology, neuroimaging) and that increasingly utilise ML.

When an ML manuscript is submitted, the editors check the manuscript for suitability and consider the following factors: whether the work presented fits with the journal's scope, whether the test data is openly available, whether the source code is open with an Open Source Initiative license, and whether the ML models are also open and available. If these criteria are met, the editors then check whether the data, code, and models are easily deployable in a containerised format. If not, authors are asked to containerise the work and add the details to the manuscript. This step makes it comply with the 'Silver' standard of Heil *et al.* (2021) and also the DOME-ML reproducibility aspect.

The GigaScience Editors have found these guidelines useful to help screen which ML publications are suitable to send out to peer review. As the journal has a strict policy promoting reproducible research, if submissions have no potential to meet even the 'Bronze' levels of reproducibility, they are then rejected. If the submission has the potential to meet these standards, the paper is then passed to the GigaDB Curation team, who then scans these manuscripts for ML content, and performs more detailed reproducibility checks to ensure that DOME annotations, in support of *GigaScience* and *GigaByte* manuscripts, are sufficiently complete. A workflow for using DOME recommendations in the context of peer review has been developed by GigaScience Press in partnership with the DOME Consortium, and this involves the following:

1. When authors submit a manuscript detailing supervised ML approaches, the editors take the DOME-ML Standards into consideration in their screening processes, only considering the manuscript for peer review if the work meets minimal standards.
2. If a manuscript passes these and other editorial pre-peer review checks, GigaDB curators then invite the authors to log into the DOME Wizard (DOME-DSW) and complete a report.
3. The DOME recommendation report is checked for completeness by curators. If sufficient, the handling editor then sends the DOME report to peer reviewers alongside links to the supporting data and code.
4. Following peer review, the annotation becomes public. The author is handed control of their entry, and it becomes public in the DOME Registry with a unique persistent identifier (DOME Registry ID).
5. The DOME Registry ID is highlighted and cited in the final published paper to allow readers to read it and more easily tease apart and access the different ML components (data, models, methods, etc.) making up the study.

The steps in this workflow that utilise the DOME Wizard and Registry are shown in Figure 3. This workflow can be adapted for other journals and has been presented at conferences and webinars to encourage uptake.

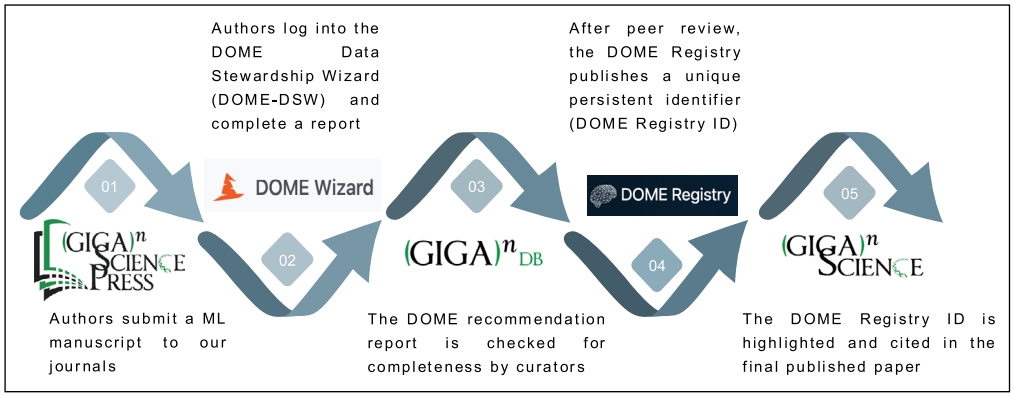

Figure 3 Workflow of the steps taken that utilise the DOME guidelines, DOME-DSW, and DOME Registry to aid the peer review of ML research. Step 1 is using the guidelines during pre-review editorial assessment; Step 2 is the creation of a DOME-DSW entry so authors can begin inputting annotations; in Step 3, the annotations are checked by the in-house curation team; Step 4 is when after review, the DSW annotations are transferred, curated, and published in the DOME registry; and in Step 5, the final DOME Registry ID is cited in the paper, and is then published.

ML annotation, such as DOME, is extremely useful for peer review as it provides a more detailed understanding of the underlying data and methodology. Whereas the manuscripts under review provide a high-level overview of the data and analysis, DOME annotation provides an intermediate-level of detail between that of the manuscript and the supporting code. This is particularly useful for understanding how the data used in an ML analysis was organised, and for detailing whether any preprocessing steps were required.

The collective opinion of Editors and Biocurators at GigaScience Press is that DOME annotations are a great asset to the peer review process, providing the necessary high-level overview to properly understand an ML study. The DOME Registry persistent identifier is included in the *GigaScience* or *GigaByte* manuscript and also in the supporting GigaDB dataset, which increases the visibility and discoverability of the DOME annotations in the research community. Through DOME Registry linking entries to ORCID identifiers, and also the APICURON database (Hatos *et al.*, 2021) of biocurators, these efforts can be further credited and encouraged (Attafi *et al.*, 2024).

Following the best practice demonstrated by the Data Citation Principles (Data Citation Synthesis Group, 2014), it is preferable to cite the DOME Registry ID in the references of the paper, both to signal that these are an important object of research, and to enable tracking of use and reuse through citation registries. To do this by including the author, year, and title, state the publisher in DOME Registry, as well as the URL of the identifier in the citation details. For example, for Atkins *et al*. (2025), GigaScience Press journals cited the DOME Registry entry following this format:

> Atkins K, Garzón-Martínez GA, Lloyd L, Doonan JH and Lu C. (2024) Unlocking the Power of AI for Phenotyping Fruit Morphology in Arabidopsis. [DOME-ML Annotations]. DOME-ML Registry, https://registry.dome-ml.org/review/a8q3rb7qrv

On top of providing this guidance to the Editorial Production team and their third-party vendors that do the typesetting, it is also useful to provide this example and guidance in the journal's instructions for authors.

USER EXPERIENCES USING ML STANDARDS IN PUBLISHING

This work presents the first-hand perspectives of a publisher implementing ML standards, but we also attempted to gather some initial feedback and data from submitting authors and users of this work. We sent a survey to the corresponding authors of all of our submissions that used the DOME Wizard, and of the 54 ML studies that tested this publishing process, we received responses from 13 authors. None had any previous experience or knowledge of ML standards, but all the respondents said they were either generally positive or neutral about their experiences annotating their experimental details into DOME. According to the curators of the DOME Registry, they estimate it takes them 1–3 hours to curate an entry, and our polling of corresponding authors found it took them an average of one hour to provide the necessary information via the DOME Wizard. Adding the time our curators needed to check, transfer, and publish information to the DOME-registry, this is in line with the 1–3 hour estimate of the DOME

team. Our feedback confirms it takes longer than the 15–30 minute estimates to annotate just the data aspects of ML studies reported by the Croissant ML metadata format (Akhtar *et al.*, 2024). Although from our experience, using DOME is quicker and easier than the REFORMS and AIMe checklists, which have nearly double and quadruple the number of fields, respectively. In particular, large and complex projects can take much longer than this average, and the DOME-Registry curators told us the submission of AlphaFold2 took approximately eight hours of work to annotate and curate given the scale of the project.

We and others have reported on the difficulty of reproducing independently collected research (González-Beltrán *et al.*, 2015), but working directly with a cohort of authors of future ML studies would be an interesting area of work to properly benchmark and assess the costs and benefits of using these different standards. As would be a more detailed assessment of the time and cost savings for the database from combining their curation processes with journal peer review. Longer-term longitudinal studies and larger sample sizes will be required to quantitatively measure any potential benefits, such as increased reuse and citations. And by continuing to contribute studies to the DOME-registry, these efforts can hopefully facilitate this future work.

CONCLUSIONS

The proliferation of ML standards across all fields of data-driven research demonstrates that there is a clear need for these in the assessment and publication of scientific research. The generalist DOME recommendations (Walsh *et al.*, 2021), REFORMS (Kapoor *et al.*, 2024), and the three-tier (Gold, Silver, Bronze) reproducibility framework (Heil *et al.*, 2021), are three possible solutions. There is potential overlap between these approaches. Indeed, the focus of the three-tier reproducibility framework is the ease of reproducibility, whereas the focus of REFORMS and DOME is on a deeper understanding of the supervised ML approach. A continued dialogue between these communities will be most fruitful in developing a common standard for ML researchers. This will be an invaluable aid to scientific publishers by providing confidence in ML reporting. Currently, this approach has been trialled in the GigaScience Press journals *GigaScience* and *GigaByte*. However, it would be useful to test this approach more widely with the inclusion of journals with different scopes and publication volumes to ascertain whether it remains practical and can become a common standard for sharing of ML research in a rigorous, reproducible, and FAIR manner, as well as collect more entries over a longer time-period to enable more quantitative data to be collected on the compliance costs and outcomes. Encouraging researchers to use these standards earlier in the research cycle will be of immense value to the entire research community by enabling a deeper understanding of the research approach, but also because of the reproducibility and provenance aspects. This effort future-proofs their work for however they may wish to disseminate and reuse it in the future—for when they and others may wish to come back to it many years in the future. Users and producers of ML research are also greatly assisted by these efforts, making it easier for them to discover and reuse novel open-source algorithms, models, and training data. The recent publication of the open-source DeepSeq-R1 model with open peer reviews and detailed supplemental materials on the training and evaluation details demonstrates that even large-scale general foundational LLMs can be opened up and transparently published in this manner (Guo *et al.*, 2025). The 83 pages of supplementary information provided with the paper are much longer and more detailed than any of these checklists, but being buried in supplemental PDFs is less visible and more difficult for humans and machines to parse and read, demonstrating the utility of more structured and standardised approaches such as DOME.

Similar to the way others have written about how data-producing teams should focus on Documentation, Automation, Traceability, and Autonomy (DATA) as priorities to make their research FAIR and reusable (Quilez *et al.*, 2017), these checklists can likewise help focus what AI-driven research should capture for future reuse. For potential users of ML-based research, as text-based descriptions have, to date, taken priority over a comprehensive sharing of the methods and training data, this new approach of capturing and making it much easier to find the different digital components making up these experiments should greatly reduce the barriers of reusability. With concerns over the sustainability and environmental impact of AI, greater reuse and reimplementation of models and data rather than the constant development of new ones should help improve efficiency and reduce this waste.

ABBREVIATIONS

AI: Artificial Intelligence; AIMe: Artificial Intelligence in Medicine; DOME: Data, Optimization, Model, Evaluation; DSW: Data Stewardship Wizard; FAIR: Findable, Accessible, Interoperable and Reusable; GigaDB: GigaScience Database; LLMs: Large Language Models; ML: Machine Learning; TRIPOD: The Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis.

ETHICS AND CONSENT

This manuscript does not involve human or animal subjects; author satisfaction surveys were anonymous, and upon discussion with the BGI IRB, formal ethical approval was not required.

ACKNOWLEDGEMENTS

The authors would like to thank the ELIXIR ML Focus Group and DOME Consortium for their support and assistance, Gavin Farrell from the DOME-registry for providing user insight and statistics, as well as the many authors and reviewers who have patiently worked with the GigaScience Press Curation and Editorial teams in integrating the resulting DOME-DSW and Registry tools into the review and publication process. We would also like to thank Chima Okafor and Chris Hunter for their help with the figures.

FUNDING INFORMATION

HLZ acknowledges the support from Guangdong Province Science and Technology Journal Excellent Talents Projects [No. 2025B1212100003] and [No. 2025B1212070003].

COMPETING INTERESTS

The authors are currently, or have been, employees of GigaScience Press. Chris Armit's contribution to this work was accomplished whilst at GigaScience Press, and prior to affiliation with Elsevier.

AUTHOR AFFILIATIONS

Scott C. Edmunds orcid.org/0000-0001-6444-1436
GigaScience Press, BGI Hong Kong Tech Co Ltd., HK SAR

Nicole Nogoy orcid.org/0000-0002-5192-9835
GigaScience Press, BGI Hong Kong Tech Co Ltd., HK SAR

Qing Lan orcid.org/0009-0009-4963-4876
GigaScience Press, BGI Center, Shenzhen 518081, Guangdong, CN

Hongfang Zhang orcid.org/0000-0002-8368-1555
GigaScience Press, BGI Center, Shenzhen 518081, Guangdong, CN

Yannan Fan orcid.org/0000-0003-3308-6878
GigaScience Press, BGI Center, Shenzhen 518081, Guangdong, CN

Hongling Zhou orcid.org/0000-0002-7295-8176
GigaScience Press, BGI Center, Shenzhen 518081, Guangdong, CN

Chris Armit orcid.org/0000-0002-9952-8141
Elsevier, 125 London Wall, Barbican, London EC2Y 5AS, UK

REFERENCES

Akhtar, M., Benjelloun, O., Conforti, C., Foschini, L. *et al.* (2024) 'Croissant: A metadata format for ML-ready datasets', *Advances in Neural Information Processing Systems*, 37, pp. 82133–82148. Available at: https://doi.org/10.52202/079017-2610

Armit, C., Tuli, M.A. and **Hunter, C.I.** (2022) 'A decade of GigaScience: GigaDB and the Open Data movement', *Gigascience*, 11, p. giac053. Available at: https://doi.org/10.1093/gigascience/giac053

Artrith, N., Butler, K.T., Coudert, F.X. *et al.* (2021) 'Best practices in machine learning for chemistry', *Nature Chemistry*, 13, pp. 505–508. Available at: https://doi.org/10.1038/s41557-021-00716-z

Atkins, K., Garzón-Martínez, G.A., Lloyd, A., Doonan, J.H. and **Lu, C.** (2025) 'Unlocking the power of AI for phenotyping fruit morphology in Arabidopsis', *Gigascience*, 14, p. giae123. Available at: https://doi.org/10.1093/gigascience/giae123

Attafi, O.A., Clementel, D., Kyritsis, K., Capriotti, E. *et al.* (2024) 'DOME Registry: Implementing community-wide recommendations for reporting supervised machine learning in biology', *GigaScience*, 13, p. giae094. Available at: https://doi.org/10.1093/gigascience/giae094

Clark, T., Caufield, H., Parker, J.A., Al Manir, S. *et al.* (2024) 'AI-readiness for biomedical data: Bridge2AI recommendations', *bioRxiv* [Preprint], 2024.10.23.619844. Available at: https://doi.org/10.1101/2024.10.23.619844

Collins, G.S., Moons, K.G.M., Dhiman, P., Riley, R.D. *et al.* (2024) 'TRIPOD+AI statement: Updated guidance for reporting clinical prediction models that use regression or machine learning methods', *BMJ*, 385, p. e078378. Available at: https://doi.org/10.1136/bmj-2023-078378

Collins, G.S., Reitsma, J.B., Altman, D.G. and **Moons, K.G.** (2015) 'Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): The TRIPOD statement', *BMJ*, 350, p. g7594. Available at: https://doi.org/10.1136/bmj.g7594

Data Citation Synthesis Group. (2014) 'Joint Declaration of Data Citation Principles', in M. Martone (ed.) *San Diego CA: FORCE11*. https://doi.org/10.25490/a97f-egyk

Farrell, G., Adamidi, E., Buono, R.A., Anton, M. *et al.* (2025) 'Open and sustainable AI: Challenges, opportunities and the road ahead in the life sciences', *arXiv*, 2505.16619. Available at: https://doi.org/10.48550/arXiv.2505.16619

Gallifant, J., Afshar, M., Ameen, S., Aphinyanaphongs, Y., Chen, S., Cacciamani, G., Demner-Fushman, D., Dligach, D., Daneshjou, R., Fernandes, C., Hansen, L.H., Landman, A., Lehmann, L., McCoy, L.G., Miller, T., Moreno, A., Munch, N., Restrepo, D., Savova, G., Umeton, R., Gichoya, J.W., Collins, G.S., Moons, K.G.M., Celi, L.A. and **Bitterman, D.S.** (2025) 'The TRIPOD-LLM reporting guideline for studies using large language models', *Nature Medicine*, 31(1), pp. 60–69. Available at: https://doi.org/10.1038/s41591-024-03425-5

González-Beltrán, A., Li, P., Zhao, J., Avila-Garcia, M.S. *et al.* (2015) 'From peer-reviewed to peer-reproduced in scholarly publishing: The complementary roles of data models and workflows in bioinformatics', *PLoS One*, 10(7), p. e0127612. Available at: https://doi.org/10.1371/journal.pone.0127612

Guo, A., Chen, Z., Li, F. and **Luo, Q.** (2023) 'Supporting data for "delineating Regions-of-Interest for mass spectrometry imaging by multimodally corroborated spatial segmentation"', *GigaScience Database*. Available at: https://doi.org/10.5524/102374

Guo, D., Yang, D., Zhang, H., Song, J. *et al.* (2025) 'DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning', *Nature*, 645(8081), pp. 633–638. Available at: https://doi.org/10.1038/s41586-025-09422-z

Haller, S., Van Cauter, S., Federau, C., Hedderich, D.M. and **Edjlali, M.** (2022) 'The R-AI-DIOLOGY checklist: A practical checklist for evaluation of artificial intelligence tools in clinical neuroradiology', *Neuroradiology*, 64(5), pp. 851–864. Available at: https://doi.org/10.1007/s00234-021-02890-w

Hatos, A., Quaglia, F., Piovesan, D. and **Tosatto S.C.E.** (2021) 'APICURON: A database to credit and acknowledge the work of biocurators', *Database (Oxford)*, 2021, p. baab019. Available at: https://doi.org/10.1093/database/baab019

Heil, B.J., Hoffman, M.M., Markowetz, F., Lee, S.I., Greene, C.S. and **Hicks, S.C.** (2021) 'Reproducibility standards for machine learning in the life sciences', *Nature Methods*, 18(10), pp. 1132–1135. Available at: https://doi.org/10.1038/s41592-021-01256-7

Hernandez-Boussard, T., Bozkurt, S., Ioannidis, J.P.A. and **Shah, N.H.** (2020) 'MINIMAR (MINimum Information for Medical AI Reporting): Developing reporting standards for artificial intelligence in health care', *Journal of the American Medical Informatics Association*, 27(12), pp. 2011–2015. Available at: https://doi.org/10.1093/jamia/ocaa088

Hugging Face. Dataset Cards. https://huggingface.co/docs/hub/en/datasets-cards.

Hugging Face. Model Cards. https://huggingface.co/docs/hub/en/model-cards.

Jones, D.T. (2019) 'Setting the standards for machine learning in biology', *Nature Reviews Molecular Cell Biology*, 20, pp. 659–660. Available at: https://doi.org/10.1038/s41580-019-0176-5

Kakarmath, S., Esteva, A., Arnaout, R., Harvey, H., Kumar, S., Muse, E., Dong, F., Wedlund, L. and **Kvedar, J.** (2020) 'Best practices for authors of healthcare-related artificial intelligence manuscripts', *NPJ Digital Medicine*, 3(134). Available at: https://doi.org/10.1038/s41746-020-00336-w

Kapoor, S., Cantrell, E.M., Peng, K., Pham, T.H., Bail, C.A., Gundersen, O.E., Hofman, J.M., Hullman, J., Lones, M.A., Malik, M.M., Nanayakkara, P., Poldrack, R.A., Raji, I.D., Roberts, M., Salganik, M.J., Serra-Garcia, M., Stewart, B.M., Vandewiele, G. and **Narayanan, A.** (2024) 'REFORMS: Consensus-based recommendations for machine-learning-based science', *Science Advances*, 10(18), p. eadk3452. Available at: https://doi.org/10.1126/sciadv.adk3452

Kolbinger, F.R., Veldhuizen, G.P., Zhu, J., Truhn, D. and **Kather, J.N.** (2024) 'Reporting guidelines in medical artificial intelligence: A systematic review and meta-analysis', *Communications Medicine (Lond)*, 4(1), p. 71. Available at: https://doi.org/10.1038/s43856-024-00492-0

Lenharo, M. (2024) 'The testing of AI in medicine is a mess. Here's how it should be done', *Nature*, 632, pp. 722–724. Available at: https://doi.org/10.1038/d41586-024-02675-0

Liu, X., Rivera, S.C., Moher, D., Calvert, M.J., Denniston, A.K., SPIRIT-AI and **CONSORT-AI Working Group.** (2020) 'Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: The CONSORT-AI extension', *BMJ*, 370, p. m3164. Available at: https://doi.org/10.1136/bmj.m3164

Luo, W., Phung, D., Tran, T., Gupta, S., Rana, S., Karmakar, C., Shilton, A., Yearwood, J., Dimitrova, N., Ho, T.B., Venkatesh, S. and **Berk, M.** (2016) 'Guidelines for developing and reporting machine learning predictive models in biomedical research: A multidisciplinary view', *Journal of Medical Internet Research*, 18(12), e323. Available at: https://doi.org/10.2196/jmir.5870

Matschinske, J., Alcaraz, N., Benis, A. *et al.* (2021) 'The AIMe registry for artificial intelligence in biomedical research', *Nat Methods*, 18, pp. 1128–1131. Available at: https://doi.org/10.1038/s41592-021-01241-0

Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I.D. and **Gebru, T.** (2019) 'Model Cards for Model Reporting', *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT* '19)*. Association for Computing Machinery, New York, NY, USA, 220–229. https://doi.org/10.1145/3287560.3287596

Mongan, J., Moy, L. and **Kahn, C.E. Jr.** (2020) 'Checklist for Artificial Intelligence in Medical Imaging (CLAIM): A guide for authors and reviewers', *Radiology: Artificial Intelligence*, 2(2), p. e200029. Available at: https://doi.org/10.1148/ryai.2020200029

Nature Methods. (2021) 'Keeping checks on machine learning', *Nature Methods*, 18(10), p. 1119. Available at: https://doi.org/10.1038/s41592-021-01300-6

Norgeot, B., Quer, G., Beaulieu-Jones, BK., Torkamani, A., Dias, R., Gianfrancesco, M., Arnaout, R., Kohane, I.S., Saria, S., Topol, E., Obermeyer, Z., Yu, B. and **Butte, A.J.** (2020) 'Minimum information about clinical artificial intelligence modeling: The MI-CLAIM checklist', *Nature Medicine*, 26(9), pp. 1320–1324. Available at: https://doi.org/10.1038/s41591-020-1041-y

Palmblad, M., Böcker, S., Degroeve, S., Kohlbacher, O., Käll, L., Noble, W.S. and **Wilhelm, M.** (2022) 'Interpretation of the DOME recommendations for machine learning in proteomics and metabolomics', *Journal of Proteome Research*, 21(4), pp. 1204–1207. Available at: https://doi.org/10.1021/acs.jproteome.1c00900

Pergl, R., Hooft, R., Suchánek, M., Knaisl, V. and **Slifka, J.** (2019) 'Data Stewardship Wizard: A tool bringing together researchers, data stewards, and data experts around data management planning', *Data Science Journal*, 18. Available at: https://doi.org/10.5334/dsj-2019-059

Pineau, J., Vincent-Lamarre, P., Sinha, K., Lariviere, V., Beygelzimer, A., d'Alche-Buc, F., Fox, E. and **Larochelle, H.** (2022) 'Improving reproducibility in machine learning research (a report from the neuriPS 2019 reproducibility program)', *Journal of Machine Learning Research*, 22, pp. 7459–7478.

Quilez, J., Vidal, E., Dily, F.L., Serra, F., Cuartero, Y., Stadhouders, R., Graf, T., Marti-Renom, M.A., Beato, M. and **Filion, G.** (2017) 'Parallel sequencing lives, or what makes large sequencing projects successful', *GigaScience*, 6(11), pp. 1–6. Available at: https://doi.org/10.1093/gigascience/gix100

Rivera, S.C., Liu, X., Chan, A.W., Denniston, A.K., Calvert, M.J., SPIRIT-AI and **CONSORT-AI Working Group.** (2020) 'Guidelines for clinical trial protocols for interventions involving artificial intelligence: The SPIRIT-AI extension', *BMJ*, 370, p. m3210. Available at: https://doi.org/10.1136/bmj.m3210

Schmied, C., Nelson, M.S., Avilov, S., Bakker, G.J. *et al.* (2024) 'Community-developed checklists for publishing images and image analyses', *Nature Methods*, 21(2), pp. 170–181. Available at: https://doi.org/10.1038/s41592-023-01987-9

Sengupta, P.P., Shrestha, S., Berthon, B., Messas, E., Donal, E., Tison, G.H., Min, J.K., D'hooge, J., Voigt, J.U., Dudley, J., Verjans, J.W., Shameer, K., Johnson, K., Lovstakken, L., Tabassian, M., Piccirilli, M., Pernot, M., Yanamala, N., Duchateau, N., Kagiyama, N., Bernard, O., Slomka, P., Deo, R. and **Arnaout, R.** (2020) 'Proposed Requirements for Cardiovascular Imaging-Related Machine Learning Evaluation (PRIME): A checklist: Reviewed by the American College of Cardiology Healthcare Innovation Council', *JACC: Cardiovascular Imaging*, 13(9), pp. 2017–2035. Available at: https://doi.org/10.1016/j.jcmg.2020.07.015

Vasey, B., Nagendran, M., Campbell, B., Clifton, D.A., Collins, G.S., Denaxas, S., Denniston, A.K., Faes, L., Geerts, B., Ibrahim, M., Liu, X., Mateen, B.A., Mathur, P., McCraddenm, M.D., Morgan, L., Ordish, J., Rogers, C., Saria, S., Ting, D.S.W., Watkinson, P., Weber, W., Wheatstone, P., McCulloch, P. and **DECIDE-AI expert group.** (2022) 'Reporting guideline for the early stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI', *BMJ*, 377, p. e070904. Available at: https://doi.org/10.1136/bmj-2022-070904

Walsh, I., Fishman, D., Garcia-Gasulla, D. *et al.* (2021) 'DOME: Recommendations for supervised machine learning validation in biology', *Nature Methods*, 18, pp. 1122–1127. Available at: https://doi.org/10.1038/s41592-021-01205-4

Wossnig, L., Furtmann, N., Buchanan, A., Kumar, S. and **Greiff, V.** (2024) 'Best practices for machine learning in antibody discovery and development', *Drug Discovery Today*, 29(7), p. 104025. Available at: https://doi.org/10.1016/j.drudis.2024.104025

TO CITE THIS ARTICLE:
Edmunds, S.C., Nogoy, N., Lan, Q., Zhang, H., Fan, Y., Zhou, H. and Armit, C 2026 Integrating Machine Learning Standards in Disseminating Machine Learning Research. *Data Science Journal*, 25: 1, pp. 1–12. DOI: https://doi.org/10.5334/dsj-2026-001

Submitted: 07 July 2025
Accepted: 23 December 2025
Published: 14 January 2026

Data Science Journal is a peer-reviewed open access journal published by Ubiquity Press.