Digital Discovery

ROYAL SOCIETY OF CHEMISTRY

PERSPECTIVE

View Article Online
View Journal | View Issue

Cite this: *Digital Discovery*, 2023, **2**, 1251

A rigorous uncertainty-aware quantification framework is essential for reproducible and replicable machine learning workflows

Line Pouchard,[a] Kristofer G. Reyes,[ab] Francis J. Alexander[c] and Byung-Jun Yoon *[ad]

The capability to replicate the predictions by machine learning (ML) or artificial intelligence (AI) models and the results in scientific workflows that incorporate such ML/AI predictions is driven by a variety of factors. An uncertainty-aware metric that can quantitatively assess the reproducibility of the quantities of interest (QoI) would contribute to the trustworthiness of the results obtained from scientific workflows involving ML/AI models. In this article, we discuss how uncertainty quantification (UQ) in a Bayesian paradigm can provide a general and rigorous framework for quantifying the reproducibility of complex scientific workflows. Such frameworks have the potential to fill a critical gap that currently exists in ML/AI for scientific workflows, as they will enable researchers to determine the impact of ML/AI model prediction variability on the predictive outcomes of ML/AI-powered workflows. We expect that the envisioned framework will contribute to the design of more reproducible and trustworthy workflows for diverse scientific applications, and ultimately, accelerate scientific discoveries.

Received 22nd May 2023
Accepted 23rd August 2023

DOI: 10.1039/d3dd00094j

rsc.li/digitaldiscovery

The ability to successfully replicate predictions by machine learning (ML) or artificial intelligence (AI) models and results in scientific workflows that incorporate such ML/AI predictions is driven by numerous factors, such as the availability of training/testing datasets, the choice of model architectures and parameters, and initial conditions.[1,2] In applications relying on deep learning models, *e.g.*, in image recognition, reproducibility depends upon initializing random seeds that can be silently set by underlying libraries, among other factors.[69] Even when the same data input and initial scripts are re-used, predictions by ML/AI models can exhibit large variability, including outliers that make these results appear unreliable.[3,4] Changing the underlying ML platforms, even new versions of the same, can alter results in a significant way.[5,6] For example, a recent reproducibility study[70] reported that a simple transcription of the same model that was originally implemented in the Java-based Magpie/Weka framework[71] to the Python-based Matminer/scikit-learn framework resulted in a significant unexpected discrepancy in the predictions made by the two platforms. Published results for ML experiments often privilege accuracy obtained with much tuning, and the publications reporting these results may not necessarily provide the ranges of input conditions that produce the reported accuracy, resulting in irreproducible results.[7,8] Varying input ranges for key physical variables in physical experiments and computational studies can be crucial to the applicability of ML algorithms to various classes of experiments. The systems-level view that encourages users to ignore low-level details and focuses instead on modeling the aggregate input–output of a particular process has generated progress in automating experimental and computational scientific workflows. In this perspective, complicated sub-systems are replaced by black-box ML/AI built from data. However, the probabilistic viewpoint that makes ML powerful at general-purpose modeling can also make its calculations opaque.[9–12,42] This is particularly important when ML models behave in unpredictable ways or when models are used to predict quantities for which there is no ground truth, as in the cases of models developed for scientific discovery. Instead of a verified result based on trusted calculations, scientists may be faced with varying predictions and no rationale to determine the best course of action.

One of the major challenges scientists will face in the coming years is the integration of ML/AI models and predictions in scientific computational and experimental workflows, whether these predictions replace expensive computational calculations, aid in predicting calculation results, help search through high-dimensional spaces to obtain preliminary candidates for analysis, and numerous new, emerging or yet unforeseen applications. We consider ML/AI predictions for scientific experiments that include numerical simulation campaigns and machine learning tasks, typically orchestrated in computational workflows. Traditionally, scientific workflows rely on building blocks carefully composed with high quality, curated data and first-principles scientific calculations often executed on High

[a]*Brookhaven National Laboratory, Upton, NY 11973, USA*
[b]*University at Buffalo, Buffalo, NY 14260, USA*
[c]*Argonne National Laboratory, Lemont, IL 60439, USA*
[d]*Texas A&M University, College Station, TX 77843, USA. E-mail: bjyoon@tamu.edu*

Performance Computing (HPC) systems.[13] Examples of promising use of ML/AI in scientific workflows include replacing some computationally expensive modules with cheaper ML-based ones, mitigating challenges that arise from limited and possibly noisy observational data, efficiently sorting through potentially billions of combinations of input candidates in discovery, and aiding just-in-time analysis of high-volume sensor data. To maximize benefits, scientists must be able to replicate the results obtained with these methods, and the ability to measure reproducibility and replicability of computational experiments, scientific workflows, and their outcomes is paramount to establishing trust in the predictions produced by ML/AI models and workflows that incorporate such models.

The attempts to mitigate the problems related to the lack of reproducibility of scientific and computational workflows have generally focused on increasing transparency and settling on a taxonomy of reproducibility. Numerous papers point to the need to increase transparency by providing access to data, code, adequate documentation for methods and execution;[14–16] these papers often propose rules and rubrics for measuring the degree to which scientific papers rate on various reproducibility scales.[17,18] To increase transparency, publishers and major conferences have adopted reproducibility requirements for submitted papers.[19–21] Containers and software package managers, tools that help build software by capturing dependencies, are often used to satisfy these requirements and provide mitigating solutions.[22] Related to these efforts, the Findable, Accessible, Interoperable, Re-useable (FAIR) principles[23,24] have provided structured guiding concepts and stimulated tool development for FAIR data and software,[25–27] including metrics of FAIR compliance.[28,29]

Trustworthy computing has been an active area of research for several decades notably within the National Science Foundation and multiple other federal agencies.[30,31] Trustworthy AI is dealing with complex systems and analytical processing pipelines that raise the bar for trust in computing results.[32] Trustworthy AI requires additional properties to achieve the goal of trustworthy computing: in particular, probabilistic accuracy under uncertainty, fairness, robustness, accountability, explainability, and formal methods are needed.[33] While trust can be subjective, it can be built from such objectified properties, the ability to reproduce results is the property of interest here. The definitions of concepts related to reproducing scientific experiments have been the object of debate among scientists and practitioners, and the preferred taxonomies have not been uniformly adopted across communities.[34,35] The taxonomies proposed by Claerbout,[36] Donoho[37] and Peng[38] have informed the definitions proposed in a 2019 report from the National Academies of Sciences, Engineering, and Medicine (NASEM).[39] NASEM defines *Reproducibility* as "Obtaining consistent computational results using the same input data, computational steps, methods, code, and conditions of analysis". NASEM assigns *Replicability* a broader definition: "Obtaining consistent results across studies aimed at answering the same scientific question, each of which has obtained its own data". In a reversal from its earlier position, the Association for Computing Machinery (ACM) now follows a similar taxonomy to NASEM in its submission policies.† In addition, ACM introduces measurements obtained with stated precision and the measuring system. In medical studies and clinical research, the consequences of overfitting for the purpose of raising the statistical significance of a study have been described early.[40] In machine learning for health care "technical, statistical, and conceptual replicability" that are required for full reproducibility of a study involve internal and external validity.[41] While we note that replicability and reproducibility are reversed from the NASEM definitions in (ref. 41), the introduction of statistical measures to assess replicability is without doubt a critical initial step in the direction we are proposing for computational workflows.

While the term "reproducibility" is commonly used across diverse science and engineering fields, its meaning is often complex and multi-faceted. An underlying reason is that, while there may be various factors affecting the reproducibility of an experimental outcome or inference result of a scientific workflow, "reproducibility" is frequently used as an umbrella term referring to the net effect of multiple factors affecting the outcome without necessarily differentiating the main sources or factors that contribute to reproducibility, or the lack thereof. It is critical to have the computational means to rigorously quantify the reproducibility of the quantities of interest, as well as assess the respective impact of diverse factors on reproducibility, *e.g.*, stochasticity of the data generation process, potential data corruption (noise or missing values) issues in scientific measurements, model uncertainty, randomness in the model training process. Even when holding fixed these quantities, additional sources of variability exist linked to software, hardware and algorithms[68] - these sources of variability are not addressed here but can be with our approach in a broader perspective.

Key to successfully employing current and future ML/AI methods is quantifying and understanding the uncertainty inherent in their recommendations and predictions. For example, when ML/AI models are employed for decision-making in scientific applications to guide future experiments – where experiments might be costly and time-consuming, experimental resources are limited, or decisions are irreversible – care must be taken in every choice made.

In this article, we focus on reproducibility (as defined in the NASEM and ACM definitions) as a necessary components of Trustworthy AI. We propose that Uncertainty Quantification (UQ) metrics[45,46] can be defined within a generalizable, objective-driven, and uncertainty-aware framework to enhance reproducibility for ML/AI. Of specific interest are scientific workflows that involve ML/AI models, whose predictions directly guide or indirectly inform decision-making in the workflow to achieve scientific goals – *e.g.* discovery, operation, verification.

For example, consider a relatively simple workflow illustrated in Fig. 1, which involves processing X-ray scattering data using a convolutional neural network (CNN) for data

† https://www.acm.org/publications/policies/artifact-review-and-badging-current.

classification. The inherent measurement noise and heterogeneity of the training data introduces significant variability in the trained ML/AI model. This gets exacerbated when the sample size of the training data is small with respect to the model complexity, which is common in scientific applications due to the high cost of data acquisition. For example, the design and training of deep network models adds additional variability, where the model architecture, choice of hyper parameters, and the use of popular optimizers based on stochastic gradient descent (SGD) with adaptive learning rates.[55–57] The interactions between these sources of variability can be highly complex, partly owing to the non-linear transformations inherent to deep learning models, which contribute to the uncertainty of the ML/AI predictions as well as the final outcomes of the overall workflow. This gives rise to several practical questions. Knowing that there will be inherent uncertainty in the predictions, to what extent can they be trusted? How will the various sources of variabilities and uncertainties affect the reproducibility of the outcomes of a given scientific workflow? How can scientists determine under what conditions they can accept and trust the results obtained from complex workflows that comprise ML/AI models? Clearly, the capability to accurately quantify the reproducibility of the outcomes of a given scientific workflow in the presence of variabilities and uncertainties would be crucial for answering these questions.

In fact, an accurate "uncertainty-aware" metric that can quantitatively assess the reproducibility (or lack thereof) of quantities of interest (QoI) would meaningfully contribute to the trustworthiness of results obtained from scientific workflows involving ML/AI models. Moreover, such a UQ metric will allow us to prioritize the various sources of uncertainties and attribute the (lack of) reproducibility to its primary source, thereby suggesting potential ways to enhance the design and training of the models and the workflow to enhance reproducibility. In addition, it may be used to assess the trade-offs between reproducibility and performance (*e.g.*, prediction accuracy), which will inform researchers to optimize the design and training of the ML models and the overall workflow.

As we elaborate in what follows, UQ in a Bayesian paradigm can provide a general and rigorous quantification framework for reproducibility for complex scientific workflows. It has the potential to fill a critical gap that currently exists in ML/AI for scientific workflows, as it will enable researchers to determine the impact of ML/AI model prediction variability as well as other sources of variabilities on the predictive outcomes of ML/AI-powered workflows. The envisioned framework will contribute to the design of more reproducible and trustworthy workflows for diverse scientific applications, and as a result, ultimately, accelerating scientific discoveries.

Uncertainty quantification metrics for reproducibility and replicability in ML models

Availability of code and datasets are insufficient metrics to assess reproducibility and replicability for ensemble models and composable workflows as they do not account for the stochasticity of training deep learning models and do not guarantee the reproducibility of the QoI in workflows involving such models.

One promising approach to help design more effective metrics is to consider the distribution of ML/AI prediction results iterated over different training sets and quantify the uncertainty of the predictions (reflected in these distributions). Uncertainty quantification based on a Bayesian paradigm is helpful here – not only because its efficacy in taking these uncertainties into account but also because many factors, including physics-informed quantities/relations and expert knowledge, can contribute to the construction of priors that mathematically represent such uncertainties.[58] Our key

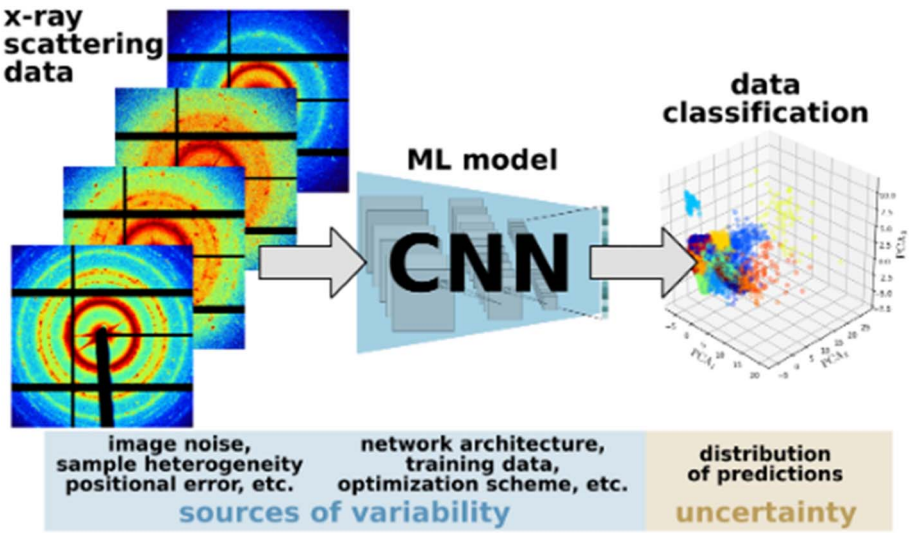

Fig. 1 The complex interactions between sources of variability result in a distribution of predictions that are difficult to interpret (credit: Kevin Yager, Center for Functional Nanomaterials, Brookhaven National Laboratory).

hypothesis is that designing metrics that introduce UQ based on a rigorous Bayesian paradigm, will help evaluate the variability of ML predictions when using such algorithms into operations. The process of training and optimizing ML/AI models typically involve various random components, which draw on several sources of variabilities – including the random splitting of available data into training, validation, and test datasets as well as the utilization of stochastic gradient descent (SGD) techniques for model training. This results in randomness of the model predictions, and as a result, when the process is repeated, we obtain an entire probability distribution of model predictions (Fig. 1). In this context, *UQ primarily involves understanding and characterizing this distribution as well as quantifying its impact on the predictions of interest*. We can assess existing ML/AI models for the respective impact of diverse factors – stochasticity of the data generation process (*e.g.*, in hydrology[62] and microgrid applications[63]), potential data corruption issues (noise or missing values), model uncertainty, randomness in the model training process, and so forth – on the reproducibility of the results obtained from compositional workflows in operational settings.

Of specific interest is addressing the question of robustness[47,48] in ill- or poorly-posed training of ML/AI models with many parameters, such as deep neural networks. While traditional methods address this through regularization methods[67] through specific terms in a loss function or certain network architectural components such as drop-out regularization, the fundamental issue concerning robustness still remains. Training such models occurs through the optimization of a cost function over a high dimensional space of model parameters, and the optimization methods employed (such as SGD) perform local optimizations without global optimization guarantees. Thus, to mitigate this localness, such optimizations employ heuristics to improve global exploration of the space, such as starting such local optimizations at various, randomly sampled initial points in the parameter space or introducing stochastic perturbations to a search. Such random elements do not guarantee convergence to a global optimum, but stochastic estimates of the optimal parameters with sufficient performance. Studying the stochastic nature of training robustness through a UQ lens does not reduce the uncertainty, or lack of robustness that occurs during training through better optimization techniques or improved regularization. Instead, we quantify such a lack of robustness as a source of uncertainty and understand how such uncertainty impacts a ML/AI model's effectiveness in achieving any experimental objectives for which such a model would be used.

Another aspect of this approach allows understanding reproducibility in the context of the development of models trained on synthetic data: to what degree does model performance transfer when applied to real, experimental data. This workflow – which entails (a) the use of physics-based models and simulators to generate synthetic data and the corresponding labels, (b) training the ML/AI models on such synthetic data, and (c) applying the synthetically-trained model to real world situations – is of special importance to experimental sciences in which obtaining real-world examples and labels is difficult or impossible. Learning how different types of models and their architectures generalize and transfer rules learned from synthetic data to real world experimental data (observations of ground truth) is another benefit of this approach. When a model predicts an observable quantity, this study is relatively straightforward and employing typical metrics to quantify the difference between ML-model predictions and experimental observations may suffice. However, in many cases, models are trained to predict intermediate quantities or quantities that are only partially observable in real-word experiments. For example, we may build a ML/AI model that predicts structure given tomographic data (or otherwise solves an inverse problem). Generating synthetic data to train such a model could involve forward simulations from structure to tomography. However, when applied to physical tomographic data, while a ML/AI model trained on such data may be able to make predictions, it could be hard or impossible to assess to what extent such structural predictions can be trusted. A UQ metric for reproducibility can provide critical insights into the reproducibility of predictions by ML/AI models trained on synthetic data, where variabilities and uncertainties in the design and training process are inevitable. Unless their impact on reproducibility – both of the model predictions and also of the various QoI in the experimental workflow incorporating such predictions – can be rigorously quantified, the role of ML/AI in accelerating scientific discoveries would be significantly hampered.

Reproducibility and replicability from the perspective of decision-making

Many factors come into play for quantifying uncertainties in ML/AI outcomes for the purpose of designing reproducibility metrics. While ML/AI models may facilitate scientific discoveries in various ways, their ability to effectively assist – and ultimately, automate – decision-making in complex scientific experiments and workflows has an especially strong potential to accelerate scientific advances. For example, ML/AI models can remove the guesswork from experimental design, thereby substantially improving the efficacy of the designed experiments. Furthermore, they may minimize (or eliminate) the need for human intervention in experimental design – an area that still heavily relies on expert intuition – ultimately, enabling autonomous experimental design loops.[59,60] Without doubt, the reproducibility of the ML/AI predictions that inform or guide "decision-making" in such autonomous experimental loops would be even more crucial. In the context of decision-making, it makes sense to assess reproducibility in terms of how the uncertainties and variabilities in the ML/AI predictions affect the expected efficacy of decision-making.

As we discussed earlier, a UQ metric for reproducibility based on a Bayesian paradigm can provide effective means of quantifying reproducibility (or lack thereof) of various QoI under uncertainty. When the decision-making aspect of ML/AI in scientific workflows is of primary interest, one may define the reproducibility metric based on an "objective-based" UQ framework that quantifies uncertainty based on its impact on

decision-making, which will likely suffer due to its presence. For example, the objective-based UQ framework *via* MOCU (mean objective cost of uncertainty)[43,44] characterizes the model uncertainty by integrating them into a decision-making framework. More specifically, MOCU quantifies the differential cost of decision-making that is expected to increase due to uncertainty, thereby solely focusing on what "actually matters" instead of various other QoI that may be of secondary importance. Due to this focus on optimal and robust decision making under uncertainty, MOCU has been actively applied to optimal experimental design (OED)[49–51] and active learning (AL)[52–54] in recent years, where the resulting OED and AL strategies have been demonstrated to outperform traditional approaches in goal-driven scientific discoveries. By defining an objective-based UQ metric for reproducibility based on a framework, we can characterize the impact of uncertainties in ML/AI predictions as well as various other sources of variabilities on reproducible decision-making in complex scientific workflows and experiments. Furthermore, such an "objective-driven" reproducibility metric will enable: (i) the identification of sources of uncertainties/variabilities that matter to users; (ii) the measurement of impact on decision-making and its scientific outcomes; and (iii) the design optimization of the model/workflow to enhance reproducibility.

Despite our primary focus on ML/AI-based scientific workflows, the aforementioned approaches can be generally applied to assess reproducibility of scientific workflows and experiments that involve non-AI based modules. These rigorous and general uncertainty-aware frameworks for quantifying reproducibility metrics will be essential in enabling trustworthy ML/AI scientific workflows that produce reproducible outcomes. Such UQ metric for reproducibility can be critical in the development phase of, for instance, climate models that integrate heterogeneous modules developed by a large scientific team of experts and run on different leadership computing facilities (LCF).[61] In addition, these metrics, when applied with granularity can help identify missing or low-quality data, as well as other potential sources of low reproducibility. Leveraging the decision-making power of frameworks such as MOCU[43,44] is an effective way of aggregating the impact of uncertainties present in ML/AI models on the reproducibility of end results.[5] Currently, it is practically challenging to accurately identify and attribute the major factors that cause ML/AI model outputs to be highly variable and therefore unreliable. A generic and objective-driven UQ metric for quantifying the reproducibility of ML/AI in the presence of diverse uncertainties in scientific applications can provide the formalism that scientists need to make informed decisions about their choice of models, and parameters given a desired level of reproducibility.

Potential applications of uncertainty-aware reproducibility metrics

The Bayesian UQ paradigm for uncertainty-aware and objective-based reproducibility quantification as well as the resulting UQ metrics for reproducibility discussed in this article, can play critical roles in enhancing the overall trustworthiness of ML/AI models in scientific workflows. The presented framework provides the means to measure potential trade-offs between accuracy and reproducibility in designing and training ML/AI models, to identify the major sources of variability and uncertainty affecting reproducibility, and to propose potential ways to make the ML/AI predictions and the end results of the workflows incorporating their predictions more reproducible. While this article did not focus on controls and processes in monitoring the training related to hardware, software, and algorithms, the preliminary inquiries in (ref. 6) indicate that our proposed research direction is critical for enhancing reproducibility on future ML/AI platforms. In a long-term, we envisage various potential applications building on such uncertainty-aware reproducibility metrics, which include: adaptive learning with measuring the reproducibility of predictions when models incorporate pieces of data not seen in training; quantitative evaluation and comparison of ML/AI surrogates in terms of reproducibility; mitigating critical gaps in input data; and automatic model calibration to maximize reproducibility or to optimize the trade-off between accuracy and reproducibility. Finally, we would like to note that there is increasing research efforts to develop UQ methods for ML/AI models, which would play important roles in enabling Bayesian UQ paradigm for uncertainty-aware quantification of reproducibility. While detailed presentation of such methods would be outside the scope of this article, we refer interested readers to relevant papers on the uncertainty quantification of ML/AI models,[64–66,72–74] as well as the references therein.

Data availability statement

As this is a "Perspective" article, no primary research results, data, software or code have been included.

Author contributions

All authors helped to perform the research by conceptualizing ideas and investigating the process and methods to be used; LP, BJY and KR wrote the original manuscript; LP, BJY, and FJA revised the manuscript; FJA supervised the research activity. All authors have reviewed the submission and agreed to the published version of the manuscript.

Conflicts of interest

There are no conflicts to declare.

Acknowledgements

The authors of this manuscript have been supported in part by Brookhaven Science Associates, LLC operator of Brookhaven National Laboratory, a U.S Department of Energy Office of Science laboratory operated under Contract No. DESC0012704.

References

1 L. A. Barba, Trustworthy computational evidence through transparency and reproducibility, *Comput. Sci. Eng.*, 2021, **23**, 58–64.
2 O. E. Gundersen and S. Kjenso, State of the Art: Reproducibility in Artificial Intelligence, *AAAI*, 2018, **32**, 1, https://ojs.aaai.org/index.php/AAAI/article/view/11503.
3 S. S. Alahmari, D. B. Goldgof, P. R. Mouton and L. O. Hall, Challenges for the Repeatability of Deep Learning Models, *IEEE Access*, 2020, **8**, 211860–211868, DOI: 10.1109/ACCESS.2020.3039833.
4 L. Pouchard (ORCID:0000000221206521), Y. Lin, and H. van Dam (ORCID:0000000208763294), *Replicating Machine Learning Experiments in Materials Science*, IOS Press, 2020, DOI: 10.3233/APC200105.
5 R. Isdahl and O. E. Gundersen, "Out-of-the-Box Reproducibility: A Survey of Machine Learning Platforms," in *2019 15th International Conference on eScience (eScience)*, 2019, pp. 86–95, DOI: 10.1109/eScience.2019.00017.
6 O. E. Gundersen, S. Shamsaliei, and R. J. Isdahl, "Do machine learning platforms provide out-of-the-box reproducibility?," *Future Generation Computer Systems*, vol. 126, pp. 34–47, 2022, DOI: 10.1016/j.future.2021.06.014.
7 B. Haibe-Kains, *et al.*, Transparency and reproducibility in artificial intelligence, *Nature*, 2020, **586**(7829), 7829, DOI: 10.1038/s41586-020-2766-y.
8 M. Hutson, Artificial intelligence faces reproducibility crisis, *Science*, 2018, **359**(6377), 725–726, DOI: 10.1126/science.359.6377.725.
9 A. B. Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado and F. Herrera, Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI, *Inform. Fusion*, 2020, **58**, 82–115.
10 V. Arya, R. K. Bellamy, P. Y. Chen, A. Dhurandhar, M. Hind, S. C. Hoffman, and Y. Zhang, One explanation does not fit all: A toolkit and taxonomy of ai explainability techniques, *arXiv*, 2019, preprint arXiv:1909.03012.
11 L. Longo, R. Goebel, F. Lecue, P. Kieseberg, and A. Holzinger, Explainable artificial intelligence: Concepts, applications, research challenges and visions, In *International Cross-Domain Conference for Machine Learning and Knowledge Extraction*, Springer, Cham, 2020, pp. 1–16.
12 W. J. Murdoch, C. Singh, K. Kumbier, R. Abbasi-Asl and B. Yu, Definitions, methods, and applications in interpretable machine learning, *Proc. Natl. Acad. Sci. U. S. A.*, 2019, **116**(44), 22071–22080.
13 L. Pouchard, S. Baldwin, T. Elsethagen, S. Jha, B. Raju, E. Stephan, L. Tang and K. K. V. Dam, Computational reproducibility of scientific workflows at extreme scales, *Int. J. High Perform. Comput. Appl.*, 2019, **33**(5), 763–776, DOI: 10.1177/1094342019839124.
14 D. A. Brown, K. Vahi, M. Taufer, V. Welch and E. Deelman, Reproducing GW150914: The First Observation of Gravitational Waves From a Binary Black Hole Merger, *Comput. Sci. Eng.*, 2021, **23**, 73–82, DOI: 10.1109/MCSE.2021.3059232.
15 O. E. Gundersen, Y. Gil and D. W. Aha, On Reproducible AI: Towards Reproducible Research, Open Science, and Digital Scholarship in AI Publications, *AIMag*, 2018, **39**, 56–68, DOI: 10.1609/aimag.v39i3.2816.
16 R. Peng and S. Hicks, Reproducible Research: A Retrospective, *Annu. Rev. Public Health*, 2021, **42**, 79–93, DOI: 10.1146/annurev-publhealth-012420-105110.
17 M. Taschuk and G. Wilson, Ten simple rules for making research software more robust, *PLoS Comput. Biol.*, 2017, **13**, e1005412, DOI: 10.1371/journal.pcbi.1005412.
18 M. S. Krafczyk, A. Shi, A. Bhaskar, D. Marinov and V. Stodden, Learning from reproducing computational results: introducing three principles and the Reproduction Package, *Philos. Trans. R. Soc., A*, 2021, **379**(2197), 20200069, DOI: 10.1098/rsta.2020.0069.
19 B. A. Plale, T. Malik and L. C. Pouchard, Reproducibility Practice in High-Performance Computing: Community Survey Results, *Comput. Sci. Eng.*, 2021, **23**, 55–60, DOI: 10.1109/MCSE.2021.3096678.
20 B. Plale and S. L. Harrell, Transparency and Reproducibility Practice in Large-scale Computational Science: A Preface to the Special Section, *IEEE Trans. Parallel Distrib. Syst.*, 2021, **32**(11), 2607–2608.
21 K. Sinha, J. Pineau, J. Forde, R. N. Ke, and H. Larochelle, *NeurIPS 2019 Reproducibility Challenge*, 2020, DOI: 10.5281/ZENODO.3818627.
22 P. Olaya, J. Lofstead, and M. Taufer, Building Containerized Environments for Reproducibility and Traceability of Scientific Workflows, 2020, arXiv:2009.08495 [cs].
23 M. D. Wilkinson, M. Dumontier, Ij. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L. B. da Silva Santos, P. E. Bourne and J. Bouwman, The FAIR Guiding Principles for scientific data management and stewardship, *Sci. Data*, 2016, **3**(1), 1–9, DOI: 10.1038/sdata.2016.18.
24 M. D. Wilkinson, M. Dumontier, I. Jan Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L. B. da Silva Santos, P. E. Bourne, *et al.*). Addendum: The FAIR Guiding Principles for scientific data management and stewardship, *Sci. Data*, 2019, **6**, 1–2, DOI: 10.1038/s41597-019-0009-6.
25 D. S. Katz, M. Gruenpeter and T. Honeyman, Taking a fresh look at FAIR for research software, *Patterns*, 2021, **2**, 100222, DOI: 10.1016/j.patter.2021.100222.
26 D. S. Katz, T. Pollard, F. Psomopoulos, E. Huerta, C. Erdmann, and B. Blaiszik, *FAIR principles for Machine Learning models*, 2020, DOI: 10.5281/zenodo.4271996.
27 H. Koers, D. Bangert, E. Hermans, R. van Horik, M. de Jong and M. Mokrane, Recommendations for Services in a FAIR Data Ecosystem, *Patterns*, 2020, **1**(5), 100058, DOI: 10.1016/j.patter.2020.100058.
28 A. Devaraju and R. Huber, An automated solution for measuring the progress toward FAIR research data, *Patterns*, 2021, **2**, 100370, DOI: 10.1016/j.patter.2021.100370.

29 M. D. Wilkinson, M. Dumontier, S.-A. Sansone, L. O. B. da S. Santos, M. Prieto, P. McQuilton, J. Gautier, D. Murphy, M. Crosas, and E. Schultes, *Evaluating FAIR-Compliance Through an Objective, Automated, Community-Governed Framework*, 2018, p. 418376, DOI: **10.1101/418376**.

30 Trust in Cyberspace. National Research Council, F. B. Schneider, ed., *Trust in Cyberspace*, National Academies Press, 1999.

31 I. Linkov, S. Galaitsi, B. D. Trump, J. M. Keisler and A. Kott, Cybertrust: From Explainable to Actionable and Interpretable Artificial Intelligence, *Computer*, 2020, **53**, 91–96, DOI: **10.1109/MC.2020.2993623**.

32 P. V. Coveney and R. R. Highfield, When we can trust computers (and when we can't), *Philos. Trans. R. Soc., A*, 2021, **379**, 20200067, DOI: **10.1098/rsta.2020.0067**.

33 J. M. Wing, Trustworthy AI, *Commun. ACM*, 2021, **64**, 64–71, DOI: **10.1145/3448248**.

34 M. A. Heroux, L. Barba, M. Parashar, V. Stodden, and M. Taufer, *Toward a Compatible Reproducibility Taxonomy for Computational and Computing Sciences*, Sandia National Lab.(SNL-NM), Albuquerque, NM (United States), 2018.

35 H. E. Plesser, Reproducibility *vs.* Replicability: A Brief History of a Confused Terminology, *Front. Neuroinform.*, 2018, **11**, 76, DOI: **10.3389/fninf.2017.00076**.

36 J. F. Claerbout, and M. Karrenbach, Electronic documents give reproducible research a new meaning, In *SEG Technical Program Expanded Abstracts 1992 SEG Technical Program Expanded Abstracts*, Society of Exploration Geophysicists, 1992, pp. 601–604, DOI: **10.1190/1.1822162**.

37 D. L. Donoho, A. Maleki, I. U. Rahman, M. Shahram and V. Stodden, Reproducible Research in Computational Harmonic Analysis, *Comput. Sci. Eng.*, 2009, **11**, 8–18, DOI: **10.1109/MCSE.2009.15**.

38 R. D. Peng, Reproducible research in computational science, *Science*, 2011, **334**, 1226–1227, DOI: **10.1126/science.1213847**.

39 National Academies of Sciences, Engineering and Medicine, *Reproducibility and Replicability in Science*, 2019, DOI: **10.17226/25303**.

40 J. P. A. Ioannidis, Why Most Published Research Findings Are False, *PLoS Med.*, 2005, **2**, e124, DOI: **10.1371/journal.pmed.0020124**.

41 M. B. A. McDermott, S. Wang, N. Marinsek, R. Ranganath, L. Foschini and M. Ghassemi, Reproducibility in machine learning for health research: Still a ways to go, *Sci. Transl. Med.*, 2021, **13**(586), eabb1655, DOI: **10.1126/scitranslmed.abb1655**.

42 C. Rudin, Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead, *Nat. Machine Intell.*, 2019, **1**, 206–215, DOI: **10.1038/s42256-019-0048-x**.

43 B.-J. Yoon, X. Qian and E. R. Dougherty, Quantifying the Objective Cost of Uncertainty in Complex Dynamical Systems, *IEEE Trans. Acoust., Speech, Signal Process.*, 2013, **61**(9), 2256–2266, DOI: **10.1109/TSP.2013.2251336**.

44 B.-J. Yoon, X. Qian and E. R. Dougherty, Quantifying the multi-objective cost of uncertainty, *IEEE Access*, 2021, **9**, 80351–80359, DOI: **10.1109/ACCESS.2021.3085486**.

45 R. Ghanem, D. Higdon, and H. Owhadi, eds., *Handbook of uncertainty quantification*, New York, Springer, 2017, vol. 6.

46 M. Abdar, F. Pourpanah, S. Hussain, D. Rezazadegan, L. Liu, M. Ghavamzadeh, P. Fieguth, X. Cao, A. Khosravi, U. R. Acharya and V. Makarenkov, A review of uncertainty quantification in deep learning: Techniques, applications and challenges, *Inform. Fusion*, 2021, **76**, 243–297.

47 L. Li, M. Weber, X. Xu, L. Rimanic, B. Kailkhura, T. Xie, C. Zhang and B. Li, November. Tss: Transformation-specific smoothing for robustness certification, In *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security*, 2021, pp. 535–557.

48 L. Li, X. Qi, T. Xie and B. Li. Sok: Certified robustness for deep neural networks, *arXiv*, 2020, preprint arXiv:2009.04131.

49 Y. Hong, B. Kwon and B.-J. Yoon, Optimal experimental design for uncertain systems based on coupled differential equations, *IEEE Access*, 2021, **9**, 53804–53810, DOI: **10.1109/ACCESS.2021.3071038**.

50 G. Zhao, X. Qian, B.-J. Yoon, F. J. Alexander and E. R. Dougherty, Model-based robust filtering and experimental design for stochastic differential equation systems, *IEEE Trans. Acoust., Speech, Signal Process.*, 2020, **68**, 3849–3859, DOI: **10.1109/TSP.2020.3001384**.

51 R. Dehghannasiri, B.-J. Yoon and E. R. Dougherty, Optimal experimental design for gene regulatory networks in the presence of uncertainty, *IEEE/ACM Trans. Comput. Biol. Bioinf.*, 2015, **12**(4), 938–950.

52 G. Zhao, E. Dougherty, B.-J. Yoon, F. Alexander and X. Qian, Efficient Active Learning for Gaussian Process Classification by Error Reduction, *Thirty-fifth Conference on Neural Information Processing Systems (NeurIPS)*, 2021.

53 G. Zhao, E. Dougherty, B.-J. Yoon, F. Alexander and X. Qian, Bayesian Active Learning by Soft Mean Objective Cost of Uncertainty, *24th International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2021.

54 G. Zhao, E. Dougherty, B.-J. Yoon, F. Alexander and X. Qian, "Uncertainty-aware Active Learning for Optimal Bayesian Classifier," *9th International Conference on Learning Representations (ICLR)*, 2021.

55 D. P. Kingma, and J. L. Ba, Adam: a Method for Stochastic Optimization. *International Conference on Learning Representations*, 2015, pp. 1–13.

56 J. Duchi, E. Hazan and Y. Singer, Adaptive Subgradient Methods for Online Learning and Stochastic Optimization, *J. Mach. Learn. Res.*, 2011, **12**, 2121–2159.

57 M. D. Zeiler(2012). *ADADELTA: An Adaptive Learning Rate Method*. **http://arxiv.org/abs/1212.5701**.

58 S. Boluki, M. Shahrokh Esfahani, X. Qian and E. R. Dougherty, Incorporating biological prior knowledge for Bayesian learning *via* maximal knowledge-driven information priors, *BMC Bioinf.*, 2017, **18**, 552.

59 H. S. Stein and J. M. Gregoire, Progress and prospects for accelerating materials science with automated and autonomous workflows, *Chem. Sci.*, 2019, **42**, 9640–9649.
60 A. Talapatra, *et al.*, Autonomous efficient experiment design for materials discovery with Bayesian model averaging, *Phys. Rev. Mater.*, 2018, **11**, 113803.
61 **https://www.doeleadershipcomputing.org**.
62 K. Beven, Issues in generating stochastic observables for hydrological models, *Hydrol. Processes*, 2021, **35**, e14203, DOI: **10.1002/hyp.14203**.
63 À. Alonso, J. de la Hoz, H. Martín, S. Coronas, P. Salas and J. Matas, A Comprehensive Model for the Design of a Microgrid under Regulatory Constraints Using Synthetical Data Generation and Stochastic Optimization, *Energies*, 2020, **13**, 5590, DOI: **10.3390/en13215590**.
64 Y. Gal, and Z. Ghahramani, Dropout as a Bayesian approximation: Representing model uncertainty in deep learning, *International Conference on Machine Learning*, PMLR, 2016.
65 A. Kristiadi, M. Hein, and P. Hennig, Being Bayesian, even just a bit, fixes overconfidence in relu networks, *International Conference on Machine Learning*. PMLR, 2020.
66 J. Watson, *et al.*, Latent derivative Bayesian last layer networks, *International Conference on Artificial Intelligence and Statistics*, PMLR, 2021.
67 J. Kukačka, V. Golkov, and D. Cremers, Regularization for deep learning: A taxonomy, *arXiv*, 2017, preprint arXiv:1710.10686.
68 O. E. Gundersen, K. Coakley, and C. Kirkpatrick, *Sources of Irreproducibility in Machine Learning: A Review*, 2022, DOI: **10.48550/arXiv.2204.07610**.
69 A. L. Beam, A. K. Manrai and M. Ghassemi, Challenges to the reproducibility of machine learning models in health care, *JAMA*, 2020, **323**, 305–306.
70 J. Hattrick-Simpers and B. DeCost, Comment on "A simple constrained machine learning model for predicting high-pressure-hydrogen-compressor materials" by Hattrick-Simpers, *et al.*, Molecular Systems Design & Engineering, 2018, 3, 509, *Mol. Syst. Des. Eng.*, 2020, **5**, 589–591.
71 J. R. Hattrick-Simpers, K. Choudhary and C. Corgnale, A simple constrained machine learning model for predicting high-pressure-hydrogen-compressor materials, *Mol. Syst. Des. Eng.*, 2018, **3**(3), 509–517.
72 R. M. Neal, *Bayesian learning for neural networks*, Springer Science & Business Media, 2012, vol. 118.
73 C. Blundell, J. Cornebise, K. Kavukcuoglu, and D. Wierstra, Weight uncertainty in neural network, In *International Conference on Machine Learning*, PMLR, 2015, pp. 1613–1622.
74 S. Liu, T. Chen, Z. Atashgahi, X. Chen, G. Sokar, E. Mocanu, M. Pechenizkiy, Z. Wang, and D. C. Mocanu. "Deep ensembling with no overhead for either training or testing: The all-round blessings of dynamic sparsity, arXiv, 2021, preprint arXiv:2106.14568.