Frontiers in Scientific Workflows: Pervasive Integration with HPC

Rafael Ferreira da Silva, *Oak Ridge National Laboratory, Oak Ridge, TN, 37831, USA*
Rosa M. Badia, *Barcelona Supercomputing Center, Barcelona, 08034, Spain*
Deborah Bard, *National Energy Research Scientific Computing Center, Berkeley, CA, 94720, USA*
Ian T. Foster, *Argonne National Lab, Lemont, IL, 60439, USA; University of Chicago, Chicago, IL, 60637, USA*
Shantenu Jha, *Rutgers University, Piscataway, NJ, 08854, USA*
Frédéric Suter, *Oak Ridge National Laboratory, Oak Ridge, TN, 37831, USA*

Abstract—This paper presents a forward-looking analysis of the evolution of scientific workflows in the context of High-Performance Computing (HPC). It emphasizes the pivotal role of scientific workflows in modern research, addressing their increasing complexity and the need for robust, adaptable, and flexible computational support systems. With HPC's crucial role in supporting these demands, the paper explores five key trends: the synergy of Artificial Intelligence (AI) and HPC workflows, the rise of cross-facility workflows, the dynamics of data-driven HPC workflows, the management of performance variability in heterogeneous HPC systems, and sustainable practices in HPC workflow design. This comprehensive examination provides insights into future challenges and opportunities, underscoring the intertwined advancement of scientific workflows and HPC technologies.

In recent years, scientific workflows have emerged as a cornerstone in advancing research in many scientific domains [1]. Characterized by their ability to streamline and automate complex data processing and analysis sequences, they have been pivotal in enhancing scientific research's efficiency, repeatability, and scalability. From decoding genomic sequences to modeling climate change, scientific workflows have facilitated groundbreaking discoveries and innovations. Their importance cannot be overstated, as they serve as the backbone for modern scientific inquiry, enabling researchers to tackle increasingly complex problems with greater precision and speed.

In this groundbreaking era of exascale computation for scientific exploration, the evolution of scientific workflows is increasingly pivotal, marking a significant shift in the research paradigm and the realization of innovations. Workflows are no longer just tools; they have become the new applications driving forward the frontiers of science [2]. For example, scientific workflows like IMPECCABLE [3] have dramatically accelerated the pace of scientific research, particularly in urgent situations like the COVID-19 pandemic. These workflows synergize high-throughput computational methods and Artificial Intelligence (AI) to rapidly screen billions of molecules against multiple drug targets, delivering a more efficient and scalable approach to drug discovery. The IMPECCABLE framework has leveraged this integrated methodology to achieve up to 1,000 times increase in efficiency over traditional methods, enabling the identification of over 1,000 compounds and progressing to the discovery of promising drug candidates.

In this context, their development and optimization have become crucial for advancing research methodologies. The intricate nature of contemporary scientific challenges demands robust, reliable, flexible, and

XXXX-XXX © 2024 IEEE
Digital Object Identifier 10.1109/XXX.0000.0000000

adaptable workflows to rapidly changing research landscapes [4]. This evolution marks a significant shift in scientific research, underscoring the need for advanced computational support systems.

In response to this paradigm shift, preparing high-performance computing (HPC) for these emerging applications is essential. With its unparalleled processing power and capacity, HPC is uniquely positioned to support the complex demands of "modern scientific workflows" [5]. As these workflows evolve to become more data-intensive and computationally demanding, the role of HPC in enabling and accelerating scientific discovery becomes increasingly critical. The future of scientific workflows is inextricably linked to the advancement of HPC technologies, necessitating a concerted effort to align HPC capabilities with the evolving needs of scientific workflows [4].

This paper, informed by a thorough examination of recent literature and expert discussions, aims to provide a forward-looking perspective on scientific workflows, particularly in the context of HPC, outlining potential challenges and opportunities that will shape research and practice in this dynamic and critical field. This paper delves into five key trends of future scientific workflow research (Figure 1):

- ***Increased Synergy between AI and HPC Workflows:*** The integration of AI with HPC is expected to lead to major advancements in computational research. AI's capacity to learn and adapt, combined with HPC's raw processing power, promises to enable more efficient, accurate, and complex scientific inquiries.
- ***Cross-Facility Workflows in Science:*** The next decade will likely witness an increased emphasis on workflows that span multiple facilities (e.g., scientific instruments and HPC systems), requiring advanced coordination, data sharing, and interoperability across diverse environments. Cross-facility collaboration will be critical in tackling large-scale, complex scientific problems.
- ***Data-Driven HPC Workflow Dynamics:*** The surge in data volume, velocity, and variety will drive the evolution of next-generation workflows. Emphasis will be on near real-time data processing, integration of heterogeneous data sources, and exploiting data to drive workflow dynamics.
- ***Heterogeneity and Performance in HPC:*** The evolution of HPC systems brings the challenge of managing performance variability due to various form of hardware heterogeneity. This will require optimizing workflows for diverse hardware architectures, addressing variability in computational resources, and ensuring efficient task execution in a heterogeneous hardware landscape.
- ***Sustainable Practices in HPC Workflow Design:*** The increasing demand for HPC resources calls for more sustainable practices. Future research will likely focus on designing energy-efficient workflows and minimizing the environmental impact of HPC operations while meeting continued demands for increased computational performance.

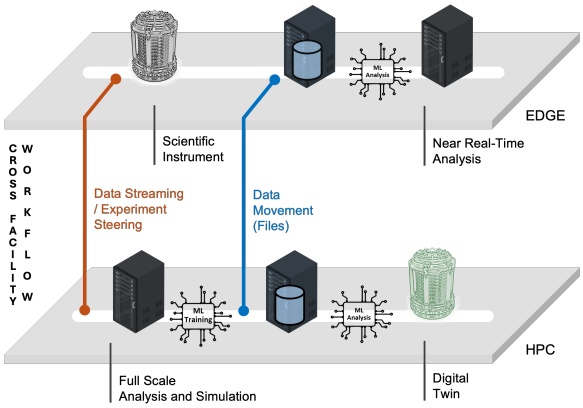

FIGURE 1. Illustration of an integrated cross-facility workflow: Data streams from a scientific instrument to HPC for ML training, enabling real-time experiment steering and data-driven discovery. ML analysis at the edge allows data to be pre-processed. Data movement via files enables full-scale analysis and the development of digital twins.

CONVERGENCE OF AI AND HPC WORKFLOWS

AI-coupled HPC workflows represent an important class of workflows, wherein the ability to couple AI methods with traditional simulation-modeling HPC workflows presents the promise of enhancing the "effective performance" of HPC workflows, where enhancement is measured by "science for a given amount of core hours or energy." Multiple recent publications have demonstrated how AI/ML coupled to traditional HPC simulations can provide practical performance enhancements of three orders of magnitude or more.

AI-coupled HPC involves the concurrent, coupled execution of AI and HPC tasks in ways that allow the AI systems to influence the HPC simulations and vice versa. Thus, an entire AI system or workflow, from data ingestion to inference through training, may have to run in parallel with the HPC simulation.

An AI system may be coupled with an HPC workflow for various reasons, from generating surrogates for

a selected component in a classical HPC simulation to guiding a complex simulation campaign towards some defined goal. An AI system can also be coupled with an HPC workflow to optimize its execution in other ways, for example, by providing Active/Reinforcement Learning capabilities, including Pareto-optimal or resource-optimal computational campaigns.

The past five years have produced promising examples of workflows that couple AI methods with traditional HPC workflows [6], typically by extending HPC workflows with additional capabilities to support the concurrent execution of AI modules [7]. For example, the Colmena framework [8] guides the execution of dynamic ensembles of simulation and learning tasks to address problems in molecular design and other domains. Such solutions have successfully coupled AI systems to be either “about” the HPC workflows or “outside” (typically in the outer loop or even possibly remote from) the main HPC workflow [6]. However, the ability to substitute an AI system “inside” of an HPC simulation remains challenging, with most solutions being customized to specific simulation software or platforms.

Greater consolidation and consistency are needed in creating “surrogates" to replace an entire HPC simulation or part thereof. The advances must encompass improved methods for substituting the HPC component/code and for training the surrogate to achieve optimality and efficiency. The ability to optimally and efficiently train a surrogate before substituting a part of the HPC component/code, and to do so for multiple concurrent, and possibly heterogeneous, HPC simulations requires advanced workflow coordination and capabilities.

Furthermore, integrating data – observational and experimental – to continuously train AI systems is powerful but revolutionary when used with simulations, as illustrated by recent successes in machine learning weather forecast models trained on reanalysis data. Creating software systems and middleware that facilitate the integration of experimental and observational data into the continuous learning and training loop is another promising avenue of HPC systems research.

Not only will workflows continue to overcome the limitations of traditional forward simulations in increasingly sophisticated and pervasive ways, but AI-coupled HPC workflows will overcome traditional bottlenecks that prevent greater scale – physical and temporal or higher resolutions. Together, AI-coupled HPC workflows present a promising paradigm, which leverages the pervasive interest and wide capabilities being developed for AI but employs them to overcome performance bottlenecks due to unsustainable approaches, such as solving PDEs at excessive granularity.

In this sense, since AI and HPC applications are typically developed using very different programming stacks, there is a need for tools that support the development of these AI-coupled-HPC workflows. The eFlows4HPC project has contributed with extensions to the PyCOMPSs workflow environment [9] to better support the integration of HPC and AI in a single workflow.

Predictions:

- Increasingly, powerful AI systems for science will be trained using data generated from HPC workflows.
- AI-coupled HPC workflows will address many of the limitations of traditional simulation-based exploration.
- Integration of experimental and observational data will become increasingly important in the steering of AI-coupled HPC workflows.

THE RISE OF CROSS-FACILITY WORKFLOWS IN NEXT-GENERATION SCIENTIFIC COLLABORATION

The increasing complexity and scale of scientific data, generated by a wide range of experimental and observational instruments (e.g., light sources, telescopes, and electron microscopes) are producing data at unprecedented rates and volumes, thereby surpassing the capabilities of individual experiment site computing facilities and necessitating advanced HPC-scale resources for intricate analysis [10]. To address these challenges, there is an increasing recognition of the need for cross-facility (or federated) workflows. These workflows, enriched with AI, enable the integration of capabilities across multiple HPC and experimental facilities. We see three main benefits to this tighter integration: (1) access to resources for near real-time data analysis; (2) better resilience for experiments; and (3) easier matching of workflows to the specific architectural strengths of different HPC systems [11], [12], [13].

Several pioneering applications exemplify the emerging landscape of cross-facility workflows in scientific research. The DECAT-DDF pipeline in astronomy processes astronomical images from the DECam imager to identify transient objects such as supernovae. The workflow leverages AI algorithms to efficiently manage data across different supercomputers

and manage access to remote databases [12]. In proteomics, the AutoSFX Workflow at the Linac Coherent Light Source (LCLS) for protein structure determination based on X-ray experiments represents a significant advancement in cross-facility coordination, integrating complex data analysis between LCLS and National Energy Research Scientific Computing Center (NERSC) facilities. We note that these workflows are not restricted to small-scale analysis, as exemplified by the ExaFEL project which aims to bring LCLS workflows to the exascale.

The Superfacility concept at Lawrence Berkeley National Laboratory (LBNL) epitomizes the integration of experimental and computational resources, supporting diverse scientific domains like cosmology and electron microscopy and facilitating near real-time data analysis [14]. Similarly, the Materials Data Facility (MDF) demonstrates the intricate process of data publication involving cross-facility workflows, encompassing data upload, quality control, and metadata management [15]. The broad deployment and powerful capabilities of Globus resource federation services [15] have facilitated the development of a wide variety of cross-facility workflows involving instruments [16] and distributed computing. However, executing complex data analysis workflows across multiple HPC facilities presents many challenges [4].

Heterogeneous hardware and software across HPC systems necessitate sophisticated middleware solutions, now increasingly AI-enhanced, for seamless workflow execution. We anticipate that efficient data management across distributed environments will become paramount, particularly for large-scale data, requiring robust data transfer, storage, and security strategies. Network constraints, such as bandwidth limitations and latency, pose additional hurdles in data transfer processes, emphasizing the need for AI-optimized network resources and effective data protocols. Scheduling and resource allocation challenges will also arise due to the need to coordinate computational resources across multiple independent facilities.

We foresee that these challenges will drive the development of advanced algorithms for efficient task allocation. Furthermore, establishing interoperability and standardization across diverse workflow management systems is crucial, involving the development of common interface standards and protocols. Scientists will also need to navigate the complexities of authentication and authorization in varied security landscapes. Addressing these challenges will be essential to successfully implement cross-facility workflows, and we see an urgent need for a new collaborative and innovative approach among researchers, developers, and facility administrators, across multiple physical sites.

Predictions:

- Widespread adoption of AI in cross-facility workflows, significantly improving data processing, analysis, and near real-time decision-making capabilities.
- Establishment of global interoperability standards and/or specifications, facilitation of smoother collaboration, specification of interfaces, and sharing of data among scientific facilities.
- Development of standardized data streaming frameworks for enabling more efficient handling of increasing volume and complexity of data across multiple facilities.

SHAPING NEXT-GENERATION WORKFLOWS WITH NEAR REAL-TIME DATA INTEGRATION

Many scientific workflows have predominantly adhered to a “move a little data, compute for a long time" model for an extended period. This was especially true in HPC, with large-scale simulations as the core components of workflows. Computationally intensive codes were run from a small set of input parameters to produce a wealth of output data that is then analyzed and visualized locally. Moreover, moving data to the computation was common as computing resources constitute the bottleneck, and communications are affordable or can be overlapped. Consequently, most of the workflow management systems developed over the last two decades focused on optimizing the compute part of the workflows and considered data movement as an effect of this orchestration [5]. However, recent evolutions in the HPC landscape call for a profound reevaluation of how workflow should be managed and how workflow systems should be designed.

The end of Moore’s law led to an evolution of processors and an increase in the number of available cores. This increase was even greater with the generalization of hardware accelerators such as GPUs in modern supercomputers. Compute resources are now the most affordable ones. Conversely, we observed a relative stagnation of network and I/O bandwidths, making data movement the new bottleneck in many scientific experiments. On the application side, the recent AI revolution in scientific computing [17] is fundamentally changing the traditional model of workflows. AI models must ingest large volumes of data

in their training phase and then produce small data in their inference phase. This new capacity to exploit multiple data sources to get more scientific insights also led scientists to add more probes to or export more information from their experiments and then add more analysis components to their workflows. Data is becoming the core component of modern and future workflows and the dynamics of entire workflows must be data-driven to cope with the surge in data volume, velocity, and variety [18].

We identify two main challenges related to this change in workflow model and the greater importance of data movement and storage in modern workflows. First, the need for near real-time analysis capabilities corresponds to more command and control of cross-facility workflows. Scientists desire to react to what they can observe in the produced data to prevent a catastrophic event on the instrument, detect a misconfiguration, or correct a diverging behavior. These new capabilities also have the potential to improve and automate the steering of computational workflows by cutting off uninteresting branches, spawning new components, generating new input parameters, or changing the accuracy and speed of the physical models. Finally, improved control of workflow resilience is needed. Science and data analysis must progress despite component failures or severe performance degradation. To address these needs, the challenge is thus to rely on a highly dynamic and flexible *data plane* controlling data streams from experimental to HPC facilities and enabling advanced interactions, such as dynamic publication of and subscription to new data streams, dynamic data replication, and advanced querying of quantities of interests.

The second challenge comes from the tighter integration of AI and HPC components within workflows. While both types of components manipulate large amounts of data, they greatly differ regarding access patterns and granularity. HPC simulations usually produce large files and data sets written by multiple processes in coordinated ways. Conversely, AI models usually consume many small files read by multiple independent processes in their training phase. The online coupling and concurrent execution of AI and HPC thus increase the competition over shared storage resources and force file systems to handle adversarial access patterns. Consequently, data management frameworks independently optimized for AI and HPC see their respective performance degrade when executed concurrently. Additional coordination between AI and HPC components is thus needed in the data plane to mitigate such impactful perturbations.

Predictions:

- Development of a more dynamic, near real-time workflow control plane in which new services and data streams can be triggered or stopped according to data-related events.
- Generalization of adaptive data reduction at the core of workflow data management to cope with the surge in velocity and volume.
- Adoption of a service-oriented approach of the data plane to abstract advanced data movement and storage methods away from the execution flow of workflow components.

NAVIGATING PERFORMANCE VARIABILITY IN EVOLVING HETEROGENEOUS HPC ENVIRONMENTS

The HPC landscape is undergoing a significant transformation, characterized by increasing heterogeneity in computing systems. As evidenced by the latest Top500 list[1], where 8 out of the top 10 systems extensively use GPUs for peak performance and approximately 35% of all Top500 machines incorporate hardware accelerators, the trend towards diverse and specialized hardware is clear. This evolution is not just limited to traditional computing centers but extends to the realm of edge computing, where resources are strategically positioned close to scientific instruments. This scenario presents a unique computing and data continuum that spans from edge computing to HPC, and potentially to cloud environments. Furthermore, the emergence of specialized AI hardware, like tensor processing units (TPUs), and the integration of dedicated AI partitions in large supercomputers, underscore the dynamic nature of modern HPC environments.

The primary challenge in this evolving heterogeneous landscape is the issue of performance portability. Maintaining consistent performance across various systems, each with distinct hardware configurations, is a daunting task. This challenge is further compounded when dealing with workflows that blend traditional HPC simulations with AI models and span across multiple facilities. To navigate this complexity, there is a pressing need for sophisticated software solutions

[1] https://www.top500.org/lists/top500/2023/11/

like the Kokkos performance portability programming ecosystem [19], and innovative approaches to orchestrate workflows. Such solutions must leverage “local services" that are tailored to optimize performance on specific hardware configurations [15], [20].

Looking ahead, we anticipate significant advancements in addressing the performance variability in heterogeneous HPC environments. As hardware continues to diversify and specialize, particularly with the rise of AI-dedicated components, the development of more adaptive and intelligent workflow management systems solutions will be paramount. Furthermore, the integration of edge computing with traditional HPC resources is expected to evolve, creating a more fluid and efficient computing continuum. This will necessitate the development of workflow orchestration tools designed to optimize the use of varying compute resources across this continuum. These solutions will enable seamless transitions between different hardware architectures within complex workflows by relying on advanced and performant local services.

Predictions:

- The HPC landscape will evolve with more specialized hardware, like GPUs and AI-specific components, enhancing specific computational performance needs.
- New software solutions will be needed to ensure performance consistency across hardware configurations in integrated HPC and AI workflows.
- Edge computing will increasingly integrate with traditional HPC, leading to more cohesive and efficient, but also more heterogeneous, computing systems that have to be exploited by cross-facility workflows.

ENERGY-EFFICIENT INNOVATIONS IN HPC WORKFLOWS

In the post-exascale era, we are witnessing increasing demands for processing scientific data and a steady enhancement in large-scale HPC capabilities, driven by a deepening commitment to sustainable, energy-efficient, eco-friendly practices in HPC. This evolution transcends mere technological achievements, marking a shift towards a future where the energy and power consumption of HPC systems, while remaining on an increasing trend, is more efficiently used by scientific applications.

In the evolving landscape of HPC and AI-driven scientific workflows, we also observe an increasing need for energy-aware workflow management that strikes a balance between time-to-solution and energy-to-solution. The integration of sustainability into these workflows is imperative to foster a comprehensive understanding of their impact across environmental, economic, societal footprint, and technical dimensions. However, the path to environmental sustainability in HPC faces several challenges.

A first challenge is to obtain certified provenance data to accurately quantify the impact on the environment and natural resources, such as carbon emissions, and the consumption of energy and water. The lack of such traceability metadata hinders the differentiation between renewable and non-renewable energy sources in HPC operations, underscoring the necessity for benchmarks and standards that promote best practices. Additionally, social sustainability is challenged by the lack of incentives for researchers, service providers, and funding agencies to prioritize sustainability, often leading to choices that are not economically or environmentally sustainable. These multifaceted challenges, spanning the entire lifecycle of the workflow, call for a holistic approach to sustainability in HPC workflows.

The envisioned sustainable future for HPC workflows is one where the principles of eco-efficiency are deeply integrated into every aspect of computing. This future is characterized by the strategic use of AI and large language models (LLMs) to optimize workflows for maximum energy efficiency and minimum environmental impact. In this scenario, workflow provenance plays a crucial role, providing essential metadata that allows for the precise tracking of energy consumption and environmental impact at every stage of the workflow. This data is pivotal in distinguishing between renewable and non-renewable energy sources, aiding in the development of benchmarks and standards for sustainable practices.

Predictions:

- Accelerated use of AI and LLMs to optimize HPC workflows, focusing on energy efficiency and environmental impact reduction, for more intelligent and adaptive scientific computing.
- Creation of new sustainability benchmarks and standards in HPC workflows for accountable and traceable energy use, differentiating renewable from non-renewable sources, and encouraging eco-friendly computing.
- Advancements in workflow provenance systems to provide for detailed tracking of energy consumption, carbon emissions, and water usage – key to minimizing environmental impact and enhancing sustainability.

CONCLUSION

This paper has underscored the crucial and evolving interplay between scientific workflows and HPC, particularly as we venture into new frontiers of research. It emphasized the increasing complexity of scientific workflows and the urgent necessity for advancements in HPC technology to meet these emerging challenges. Delving into the areas of AI integration, cross-facility workflows, data-driven dynamics, and hardware heterogeneity, we presented a comprehensive picture of the future joint trajectory of HPC systems and scientific workflows.

Simultaneously, this paper predicts a dynamic evolution in these domain areas. It foresees a future where the integration of AI with HPC will significantly enhance computational research capabilities, while new cross-facility workflows will address complex, large-scale scientific challenges. Additionally, it highlights the growing impact of voluminous data on workflow dynamics and the challenges posed by hardware heterogeneity. Finally, emphasizing the need for sustainable practices in HPC operations, the paper spotlights the pressing need for innovative solutions to adapt to the rapidly evolving landscape of scientific research and computing technologies.

Although not directly addressed in this paper, we stress that as scientific workflow systems become increasingly integral to research, addressing the user experience (UX) challenges they present is critical. Challenges include the rapid evolution of user needs, the diverse user base, and the complexity of workflow systems. To enhance UX, potential approaches include engaging in iterative design processes, adopting mixed-method research for in-depth user behavior analysis, and designing with a focus on usability and accessibility. Emphasizing these strategies can help bridge the gap between the sophistication of scientific workflows and the practical ease with which researchers interact with them, fostering broader and more effective adoption.

ACKNOWLEDGMENTS

This research used resources of the Oak Ridge Leadership Computing Facility at the Oak Ridge National Laboratory, supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC05-00OR22725. This work was supported in part by the U.S. Department of Energy under Contract DE-AC02-06CH11357.

REFERENCES

1. R. M. Badia Sala, E. Ayguadé Parra, and J. J. Labarta Mancho, "Workflows for science: A challenge when facing the convergence of HPC and big data," *Supercomputing Frontiers and Innovations*, vol. 4, no. 1, pp. 27–47, 2017.
2. T. Ben-Nun, T. Gamblin, D. S. Hollman, H. Krishnan, and C. J. Newburn, "Workflows are the new applications: Challenges in performance, portability, and productivity," in *Proceedings of the IEEE/ACM International Workshop on Performance, Portability and Productivity in HPC*. IEEE, 2020, pp. 57–69.
3. A. A. Saadi, D. Alfe, Y. Babuji, A. Bhati, B. Blaiszik, A. Brace, T. Brettin, K. Chard, R. Chard, A. Clyde *et al.*, "Impeccable: Integrated modeling pipeline for covid cure by assessing better leads," in *Proceedings of the 50th International Conference on Parallel Processing*, 2021, pp. 1–12.
4. R. Ferreira da Silva, R. M. Badia, V. Bala, D. Bard, T. Bremer, I. Buckley, S. Caino-Lores, K. Chard, C. Goble, S. Jha, D. S. Katz, D. Laney, M. Parashar, F. Suter, N. Tyler, T. Uram, I. Altintas *et al.*, "Workflows Community Summit 2022: A Roadmap Revolution," Oak Ridge National Laboratory, Tech. Rep. ORNL/TM-2023/2885, Mar. 2023.
5. C. S. Liew, M. Atkinson, M. Galea, T. F. Ang, P. Martin, and J. I. V. Hemert, "Scientific workflows: Moving across paradigms," *ACM Computing Surveys*, vol. 49, no. 4, pp. 1–39, 2016.
6. S. Jha, V. R. Pascuzzi, and M. Turilli, "Ai-coupled hpc workflows," *Chapter 28, Artificial Intelligence for Science, pp. 515-534 (2023)*, p. arXiv:2208.11745, Aug. 2022.

7. A. Brace, I. Yakushin, H. Ma, A. Trifan, T. Munson, I. Foster, A. Ramanathan, H. Lee, M. Turilli, and S. Jha, "Coupling streaming ai and hpc ensembles to achieve 100–1000× faster biomolecular simulations," in *2022 IEEE International Parallel and Distributed Processing Symposium (IPDPS)*. IEEE, 2022, pp. 806–816.
8. L. Ward, G. Sivaraman, J. G. Pauloski, Y. Babuji, R. Chard, N. Dandu, P. C. Redfern, R. S. Assary, K. Chard, L. A. Curtiss, R. Thakur, and I. T. Foster, "Colmena: Scalable machine-learning-based steering of ensemble simulations for high performance computing," in *Proceedings of the IEEE/ACM Workshop on Machine Learning in High Performance Computing Environments*. IEEE, 2021, pp. 9–20.
9. E. Tejedor, Y. Becerra, G. Alomar, A. Queralt, R. M. Badia, J. Torres, T. Cortes, and J. Labarta, "PyCOMPSs: Parallel computational workflows in Python," *The International Journal of High Performance Computing Applications*, vol. 31, no. 1, pp. 66–82, 2017.
10. B. L. Brown, W. L. Miller, D. Bard, A. Boehnlein, K. Fagnan, C. Guok, E. Lançon, S. J. Ramprakash, M. Shankar, and N. Schwarz, "Integrated Research Infrastructure architecture blueprint activity (final report 2023)," US Department of Energy (USDOE), Tech. Rep., 2023, https://doi.org/10.2172/1984466.
11. K. B. Antypas, D. Bard, J. P. Blaschke, R. S. Canon, B. Enders, M. A. Shankar, S. Somnath, D. Stansberry, T. D. Uram, and S. R. Wilkinson, "Enabling discovery data science through cross-facility workflows," in *Proceedings of the IEEE International Conference on Big Data*. IEEE, 2021, pp. 3671–3680.
12. N. Tyler, R. Knop, D. Bard, and P. Nugent, "Cross-facility workflows: Case studies with active experiments," in *Proceedings of the IEEE/ACM Workshop on Workflows in Support of Large-Scale Science*. IEEE, 2022, pp. 68–75.
13. Z. Liu, A. Ali, P. Kenesei, A. Miceli, H. Sharma, N. Schwarz, D. Trujillo, H. Yoo, R. Coffee, N. Layad, J. Thayer, R. Herbst, C. H. Yoon, and I. Foster, "Bridging data center AI systems with edge computing for actionable information retrieval," in *Proceedings of the 3rd Annual Workshop on Extreme-scale Experiment-in-the-Loop Computing*. IEEE, 2021, pp. 15–23.
14. B. Enders, D. Bard, C. Snavely, L. Gerhardt, J. Lee, B. Totzke, K. Antypas, S. Byna, R. Cheema, S. Cholia *et al.*, "Cross-facility science with the Superfacility Project at LBNL," in *Proceedings of the IEEE/ACM 2nd Annual Workshop on Extreme-scale Experiment-in-the-Loop Computing*. IEEE, 2020, pp. 1–7.
15. R. Chard, J. Pruyne, K. McKee, J. Bryan, B. Raumann, R. Ananthakrishnan, K. Chard, and I. T. Foster, "Globus Automation Services: Research process automation across the space–time continuum," *Future Generation Computer Systems*, vol. 142, pp. 393–409, 2023.
16. R. Vescovi, R. Chard, N. D. Saint, B. Blaiszik, J. Pruyne, T. Bicer, A. Lavens, Z. Liu, M. E. Papka, S. Narayanan, N. Schwarz, K. Chard, and I. T. Foster, "Linking scientific instruments and computation: Patterns, technologies, and experiences," *Patterns*, vol. 3, no. 10, p. 100606, 2022.
17. R. Stevens, V. Taylor, J. Nichols, A. B. Maccabe, K. Yelick, and D. Brown, "AI for science: Report on the Department of Energy (DOE) town halls on artificial intelligence (AI) for science," Argonne National Lab (ANL), Argonne, IL (United States), Tech. Rep., 2020.
18. F. Suter, R. Ferreira da Silva, A. Gainaru, and S. Klasky, "Driving next-generation workflows from the data plane," in *Proceedings of the 19th IEEE Conference on eScience*, 2023.
19. C. R. Trott, D. Lebrun-Grandié, D. Arndt, J. Ciesko, V. Dang, N. Ellingwood, R. Gayatri, E. Harvey, D. S. Hollman, D. Ibanez *et al.*, "Kokkos 3: Programming model extensions for the exascale era," *IEEE Transactions on Parallel and Distributed Systems*, vol. 33, no. 4, pp. 805–817, 2021.
20. R. Souza, T. J. Skluzacek, S. R. Wilkinson, M. Ziatdinov, and R. F. da Silva, "Towards lightweight data integration using multi-workflow provenance and data observability," in *Proceedings of the IEEE 19th International Conference on e-Science*. IEEE, 2023.

Rafael Ferreira da Silva is a Senior Research Scientist at the Oak Ridge National Laboratory. His current research interests include parallel and distributed computing systems, with a primary focus on scientific workflows. Rafael received his Ph.D. degree in Computer Science from Institut National des Sciences Appliquées de Lyon. He is a Senior Member of the ACM and IEEE. Contact him at silvarf@ornl.gov.

Rosa M. Badia is the manager of the Workflows and Distributed Computing group at the Barcelona Supercomputing Center (BSC). Her research interest include workflow development, HPC-AI convergence and software development for edge-to-cloud. Rosa received her Ph.D. degree in Computer Science from the Technical University of Catalonia. She is an ACM Distinguished member and an IEEE member. Contact her at rosa.m.badia@bsc.es.

Deborah Bard is the Group Lead for Data Science Engagement at the National Energy Research Scien-

tific Computing center (NERSC). Her research interests include complex workflows and the intersection of HPC and experimental science. She holds a PhD in experimental particle physics from the University of Edinburgh. Contact her at djbard@lbl.gov.

Ian T. Foster is Director of the Data Science and Learning Division at Argonne National Laboratory, and Professor of Computer Science at the University of Chicago. His research interests include high-performance and distributed computing. He holds a PhD in Computer Science from Imperial College, UK. He is a Fellow of the ACM and IEEE. Contact him at foster@anl.gov.

Shantenu Jha is a Professor of Computer Engineering at Rutgers University and the Division Director for Data-Driven Discovery at Brookhaven National Laboratory. His research interests are at the intersection of high-performance distributed computing and computational and data-driven science. He received his Ph.D. in Physics from Syracuse University. Contact him at shantenu.jha@rutgers.edu.

Frédéric Suter is a Senior Research Scientist at the Oak Ridge National Laboratory. His research interests include scheduling, scientific workflows, and simulation of parallel and distributed systems and applications. He received his Ph.D. in Computer Science from the Ecole Normale Supérieure de Lyon, France. Contact him at suterf@ornl.gov.

COMSI-2024-01-0038 – Frontiers in Scientific Workflows: Pervasive Integration with HPC

ANSWER TO REVIEWERS

We provide answers below to all specific criticisms, suggestions, and/or questions (we do not quote reviewer comments that were purely positive nor comments that summarized the content of the paper).

Reviewer #1

> *I'm having a hard time seeing how the workflow management systems themselves can address performance portability. There are several other strategies to address this (Kokkos was a good one as are some container strategies), but the statement, "These solutions will likely focus on enhancing performance portability," seems specious. The best the WMS can do is help with the higher level ability to migrate elements of a workflow, but that's not performance portability.*

We agree that the stating that WMSs will enhance performance portability is incorrect. We rewrote the end of this paragraph to express that WMSs will enable and ease the composability of advanced local services deployed across heterogeneous resources. These local services will be in charge of getting the best performance from the underlying resources, while their composability brings the required performance portability.

> *I also don't see that power consumption is really going to be substantially reduced in absolute terms. Hopefully we can see lower power increases for more performance benefit but that's as far as I'm willing to go.*

This is right. The power consumption of future supercomputers is not likely to decrease. Speaking of substantial reduction of energy consumption is not realistic. However, we can hope for a more efficient usage of the available power by scientific applications and workflows. We rephrased this paragraph accordingly.

> *There are statements that could use the backing of references. [. . .] Important references are missing; more references are needed*

The guidelines for the submission of articles to IEEE Computer Magazine limit the number of references to 20 (https://www.computer.org/csdl/magazine/co/write-for-us/15913?title=Author%20Information&periodical=Computer). We thus had to carefully select articles to cite.

> *There are various grammar issues.*

We thank the reviewer for the annotated PDF. We have carefully revised the grammar in the manuscript.

Reviewer #2

> *This manuscript can be enhanced for at least three aspects:*
> 1) *Including at least one critical scientific workflow use case to demonstrate how a scientific workflow system would accelerate scientific research.*
> 2) *Exploring the interaction between the workflow systems and its end users and not simplifying focusing on scientific workflows as a technical system.*
> 3) *Adding more figures to illustrate some of the main concepts. The current manuscript reads like a proposal and is quite dry.*

1) In the introduction, we added a brief paragraph that highlights an exemplary workflow application that was instrumental in accelerating the drug discovery process amid the pandemic. This example underscores the critical role of scientific workflows in swiftly navigating the complex landscape of drug development, demonstrating their capacity to streamline the integration of vast datasets and computational methods to significantly accelerate the journey from initial screening to potential therapeutic candidates.
2) We acknowledge the significance of user experience (UX) in scientific workflow systems—a topic not directly covered within the scope of this paper but undoubtedly of critical importance. To address this, we have incorporated a concise paragraph in the conclusion that sheds light on the prevalent UX challenges and suggests potential avenues for future research.
3) We have included a diagram that encapsulates the key trends highlighted in this paper, delineating their interconnectivity and mutual influence.

> *The weakest aspect of the manuscript is its lack of convincing use cases to demonstrate how the proposed flagship scientific workflows will impact the scientific community and reshape the practice of scientific discovery.*
>
> *The scientific community needs innovative, easy-to-deploy, easy-to-use scientific workflows to accelerate their research. If future scientific workflows have so many challenges to be resolved, as foreseen by the authors, there may be more efficient ways to develop and deploy the workflows.*
>
> *After all, scientific workflows enable science rather than the other way around.*

We acknowledge the need for concrete examples demonstrating the impact of our proposed workflows. Thus, we have enriched our manuscript with an exemplar use case, the IMPECCABLE workflow, which significantly accelerated COVID-19 drug discovery by integrating AI with high-throughput methods. Additionally, we discuss cross-facility workflows that manage complex data and resources across multiple scientific facilities, illustrating their transformative potential. We also emphasize the importance of UX in the successful deployment of these systems, as highlighted in our conclusion. Addressing UX challenges ensures that these powerful tools are not only effective but also accessible and user-friendly for researchers.

Reviewer #3

> *The last paragraph of cross-facility section is too long, and can be split in several paragraphs. The second paragraph of energy-efficient is also long;*

We broke down these paragraphs to improve the reading experience.

> *I would put AI & Workflows section at the end, as somehow all the others already motivate the use of AI.*

Thank you for your suggestion regarding the placement of the AI & Workflows section. We appreciate your perspective on integrating this section towards the end of the manuscript, given its relevance to the other sections. However, we believe that positioning AI as one of the initial motivators for the paradigm shift in scientific workflows emphasizes its foundational role in driving advancements across all discussed topics.

> *Concerning the references, I would try whenever possible to avoid self-references. I would also provide more references on applications on AI and Workflows (Mainly on the first paragraph, page 2, column 2, rows 8 to 16). For instance, the cross-facility section has a good number of refs. Missing reference for AI and LLMs.*

Again, we are limited by the magazine guidelines to 20 references. We can thus not have the same number of references as in a journal article.