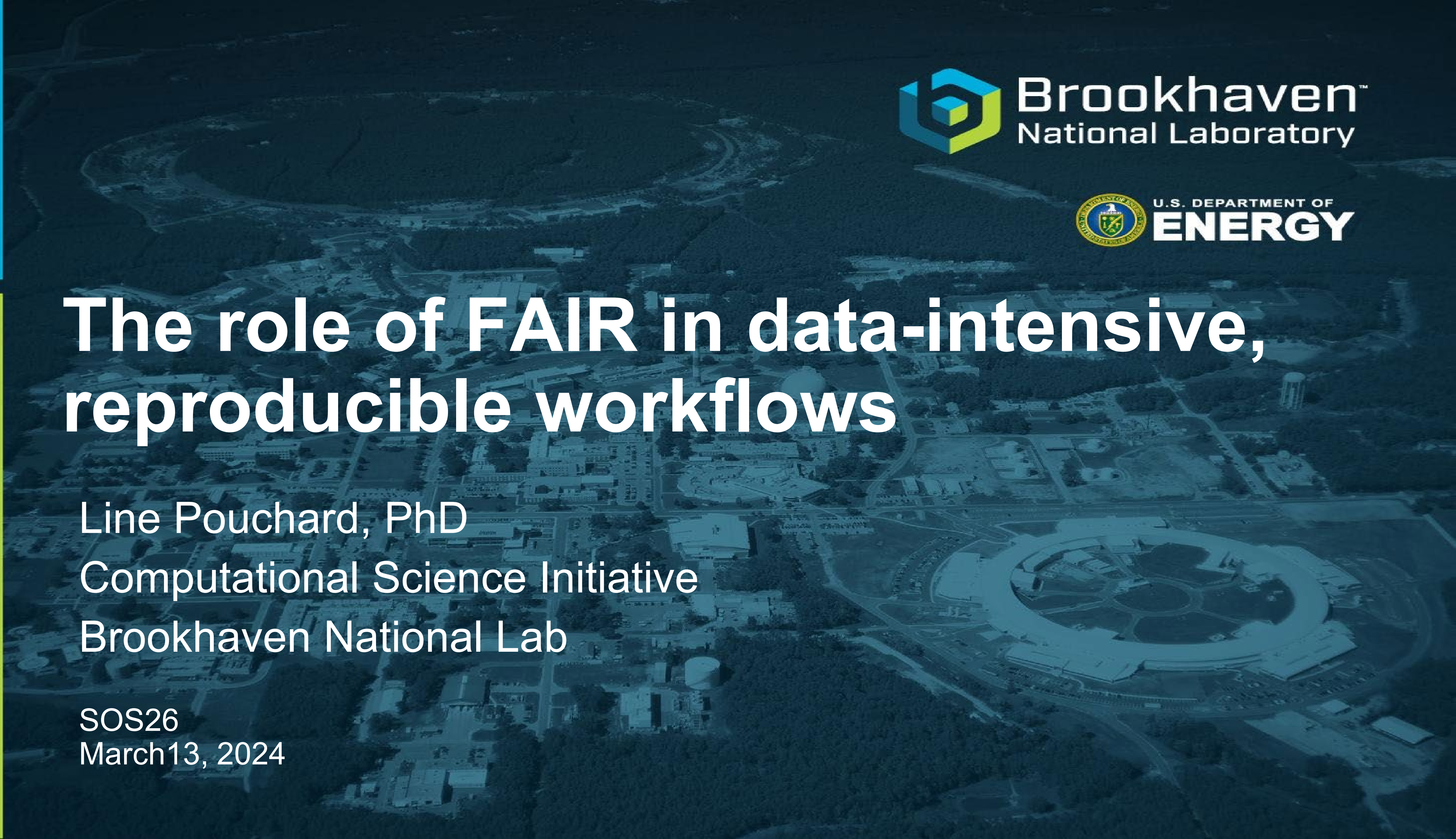


The role of FAIR in data-intensive, reproducible workflows

Line Pouchard, PhD
Computational Science Initiative
Brookhaven National Lab

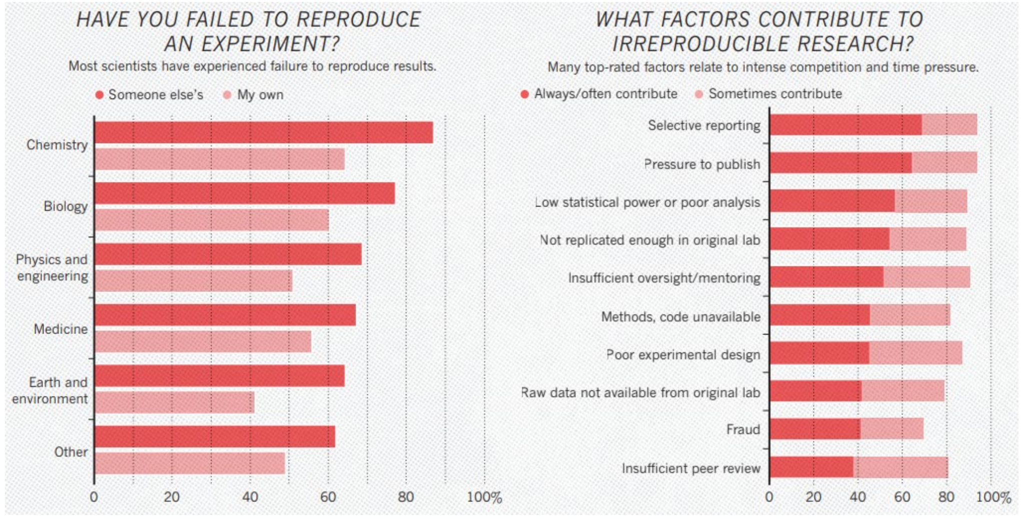
SOS26
March13, 2024

Is there a reproducibility crisis in science?

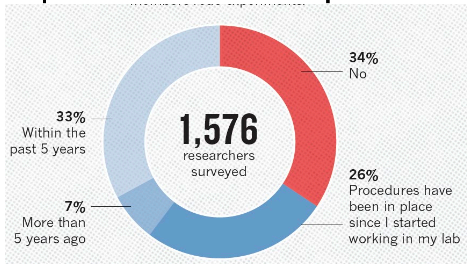
- Nature, 2016: 52% of 1576 respondents believe there is a significant crisis
- Reproducibility survey at SC20
  - ***Awareness***: Participants are aware of issues around reproducibility in science (90%); only 7% think they do not apply to computer or computational science
  - ***Concern:*** As 21% think that publishing code is sufficient to ensure reproducibility
  - 204 respondents out of 9,949 unique SC Technical Program participants, 2016-19. AD/AE became a requirement for SC19.

Plale BA, Malik T, Pouchard LC (2021) Reproducibility Practice in High-Performance Computing: Community Survey Results. Computing in Science & Engineering 23:55–60. https://doi.org/10.1109/mcse.2021.3096678

- Today: "reproducibility crisis in science": 873 articles in google scholar



Established procedures for reproducibility: 34% NO



Baker et al., "1,500 scientists lift the lid on reproducibility". Nature 533, 452–454

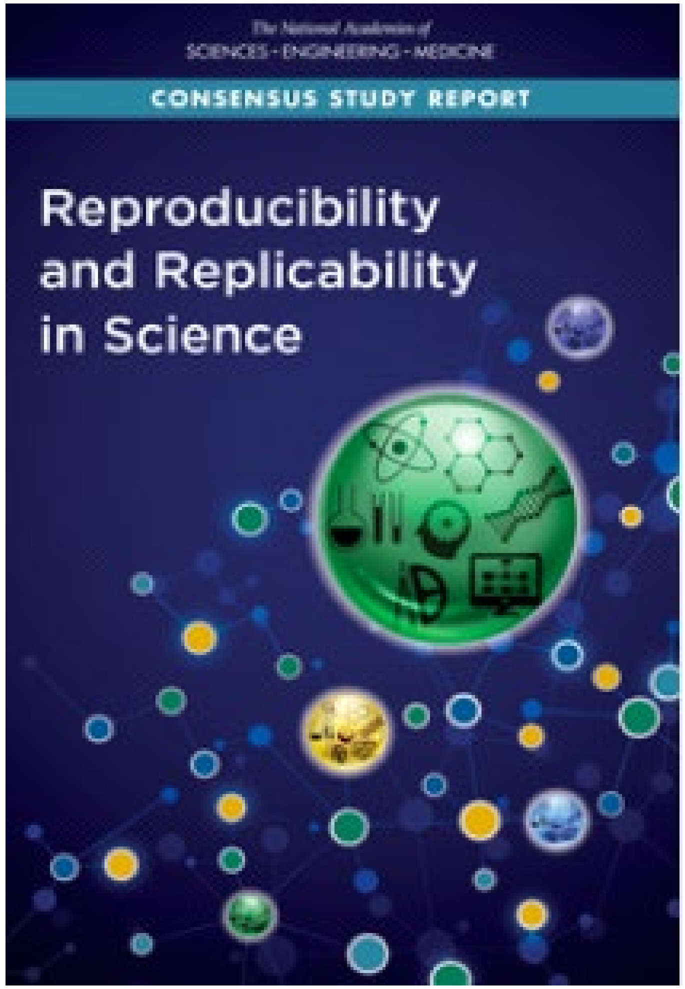
Reproducibility definitions

The official definitions : National Academies of Science, Engineering, & Medicine (NASEM): 2019 Report

Reproducibility: a narrower definition
Obtaining consistent computational results using the same input data, computational steps, methods, code, and conditions of analysis.

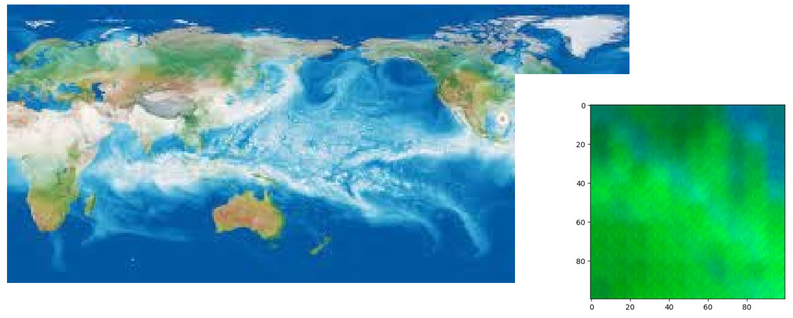
Replicability: a broader definition
Obtaining consistent results across studies aimed at answering the same scientific question, each of which has obtained its own data.

Computational reproducibility: the ability to recreate the reported results of a scientific study from its computational elements (Stodden)


The National Academies of
SCIENCES · ENGINEERING · MEDICINE
CONSENSUS STUDY REPORT
Reproducibility and Replicability in Science

Brookhaven National Laboratory

Reproducibility for ML models at scale



Stengel,et.al: "Adversarial super-resolution of climatological wind and solar data," 2020, doi: 10.1073/pnas.1918964117.

This experiment motivates research on reproducibility for data intensive computing. Most of the predictions are not within the 0.3 Relative Root Mean Square Error threshold for fair performance

Experiment:

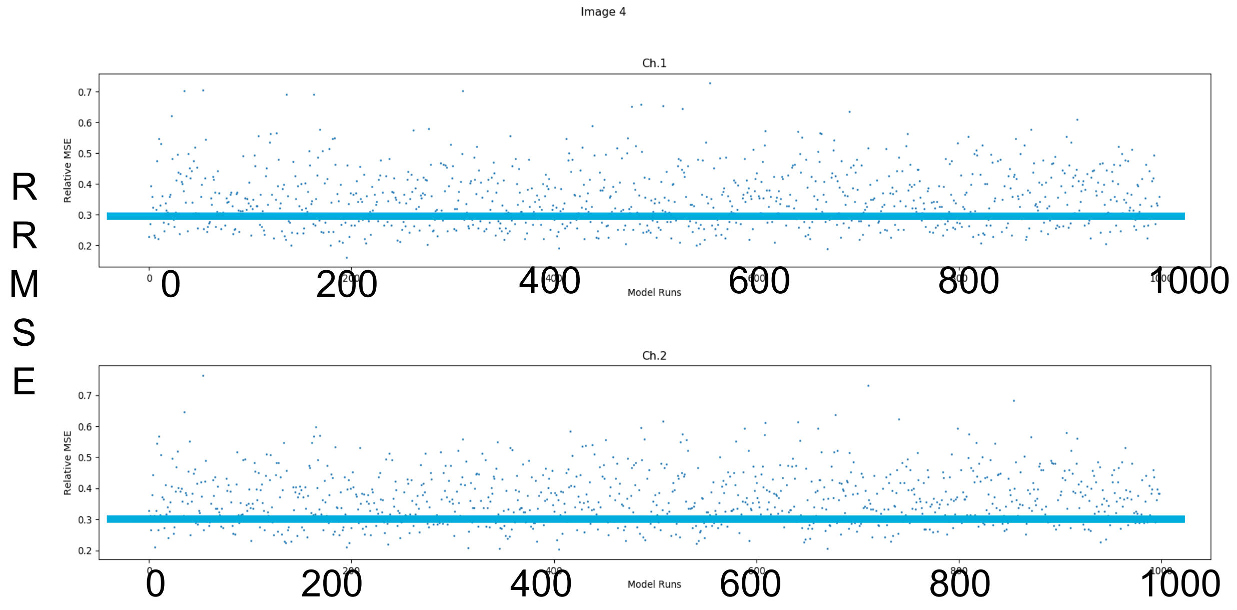
- re-training the model 1000 times
- same data, same system, same code
- wind variable

Expectation:

- even with stochastic models, there should be less variation in 1000 runs
- Potentially not enough training data provided

=> Better guidance on the necessary documentation and annotations provided by authors is needed

Reproducing the predictions of an adversarial super resolution model: a case study, **Pouchard**, Yoon, Van Dam, CoDA 2023



Variations in results over 1000 model runs. Blue line is the 0.3 threshold, above which performance is only considered poor

Numerous challenges for reproducibility in AI/ML

43 sources of irreproducibility in ML identified in https://doi.org/10.48550/arXiv.2204.07610

Model related

- Inherently stochastic models
- Random initialization
- Random splitting of training data
- Local optimization
- Incorrect Gaussian assumptions

Computational reproducibility issues

- Availability of specialized architectures
- Numerical reproducibility, floating point precision
- Digitalization of continuous phenomena
- Machine Learning platform releases
  - TensorFlow, PyTorch, etc.

Data related

- Documentation, metadata, provenance
- Appropriate representation of a phenomena in training data: geographical and ethnic diversity
- Data quality, sparsity, and mitigation
- Appropriate use of synthetic data and transfer models

Theory related issues

- Systematic errors due to fundamental model approximations
- Choices in the discretization
- different representations of some basis functions
- Coding errors
- Bugs

Reproducibility at scale: what for? and what kind?

Numerical Reproducibility: to validate code correctness

Climate models during model development phase by large independent teams

A goal in QCD to avoid additional fluctuations in MC simulations

Partial reproducibility: Key quantities but not all are reproduced

Classical Molecular Dynamics

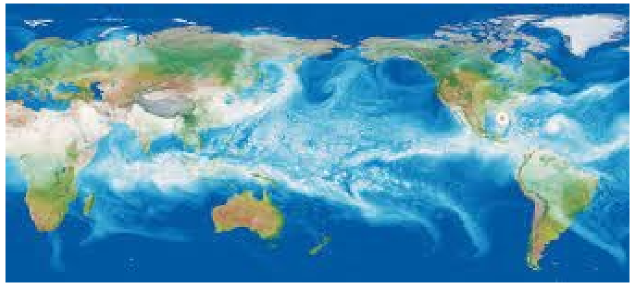
Molecular ab-initio calculations

Materials

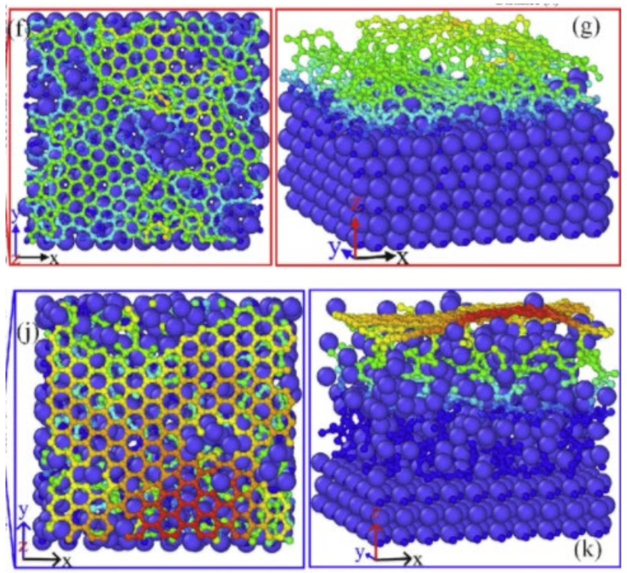
Performance reproducibility: minimal run-to-run variation across multiple runs of the same application using a consistent set of configurations – different from performance portability

T. Patki, et.al. Performance optimality or reproducibility: that is the question, SC '19. [10.1145/3295500.3356217](https://doi.org/10.1145/3295500.3356217)

Pouchard, LC. Reproducibility in Extreme Heterogeneous Architectures White Paper
Extreme Heterogeneity DOE Workshop participant, 2018



E3SM



Tingwei Hu - LAMMPS

FAIR is a set of principles to provide guidelines for digital assets supporting discovery through good data management with machine actionability

Findable
- Rich metadata
- Indexed
- Persistent idendifiers

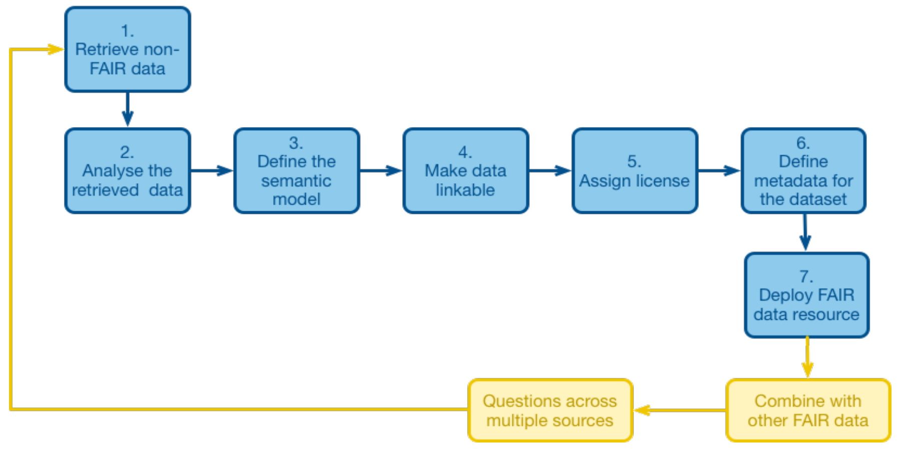
Accessible
- Standard communication protocol
- Free protocol
- Allows for authentication and authorization

Interoperable
- Formal, shared, broadly applicable language for knowledge representation
- Metadata vocabularies
- Qualified references to other metadata

Reusable
- Plurality of accurate and relevant attributes
- Detailed provenance
- Meets domain-relevant community standards

The FAIR Guiding Principles for scientific data management and stewardship Mark D. Wilkinson et al.2018. 10.1038/sdata.2016.18

The FAIRification process

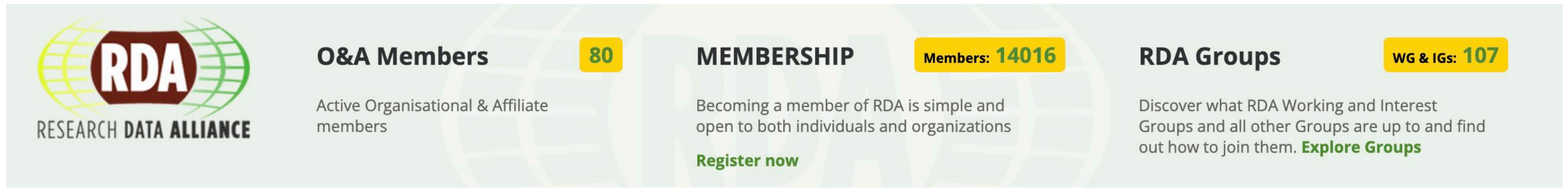


1. Retrieve non-FAIR data
2. Analyse the retrieved data
3. Define the semantic model
4. Make data linkable
5. Assign license
6. Define metadata for the dataset
7. Deploy FAIR data resource
Combine with other FAIR data
Questions across multiple sources

GO FAIR

Applies to Data and Metadata

https://go-fair.org and https://go-fair.us


RESEARCH DATA ALLIANCE

O&A Members 80

Active Organisational & Affiliate members

MEMBERSHIP Members: 14016

Becoming a member of RDA is simple and open to both individuals and organizations

Register now

RDA Groups WG & IGs: 107

Discover what RDA Working and Interest Groups and all other Groups are up to and find out how to join them. **Explore Groups**

Brookhaven National Laboratory

Can the FAIR-ification of digital objects help?

What should be made available for SW, data, and workflows to become FAIR? And How?

Computing environments, submission scripts, libraries and their version numb
Scientific metadata, performance counters, instrumentation choices, instrume
metadata

Metadata exist and is captured in various *non-interoperable* data formats, schemas, and services

- data services, containers
- machine learning platforms and their versions: Tensorflow, pytorch..

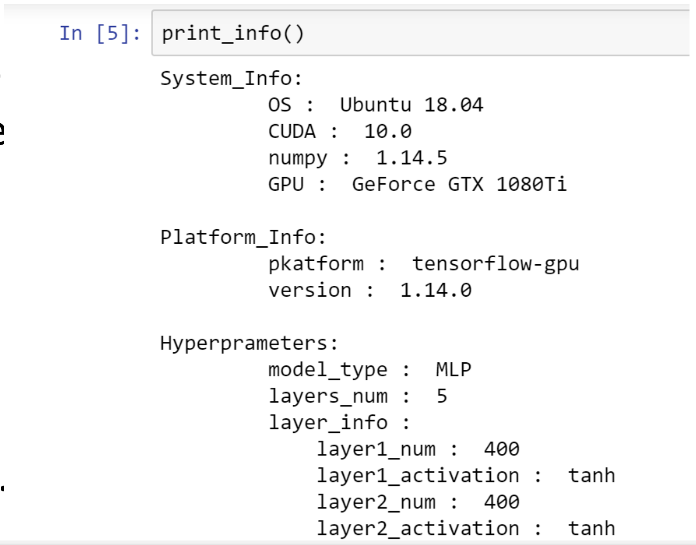
Metadata standards: WFCommons, Common Workflow Language

Automatic capture of provenance:

- Metadata and their relationship to data
- SW dependencies
- Workflow behavior

Persistent Identifiers: many schemes e.g. ARK, DOI, minIDs, easyIDs, dPIDs

Scripts, FAIR, model cards (Google ML, huggingface, etc.), NeurIPS and SC appendices

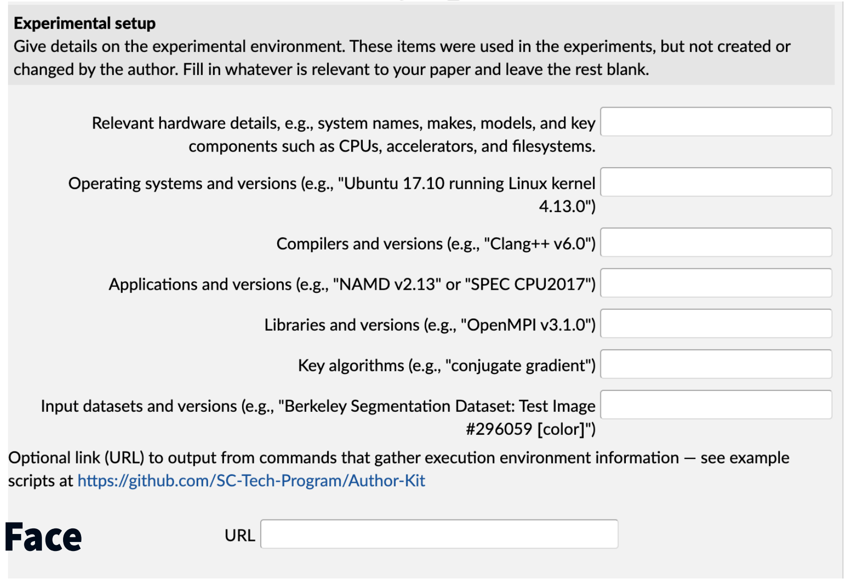


```
In [5]: print_info()

        System_Info:
                 OS :  Ubuntu 18.04
                 CUDA :  10.0
                 numpy :  1.14.5
                 GPU :  GeForce GTX 1080Ti

        Platform_Info:
                 pkatform :  tensorflow-gpu
                 version :  1.14.0

        Hyperprameters:
                 model_type :  MLP
                 layers_num :  5
                 layer_info :
                     layer1_num :  400
                     layer1_activation :  tanh
                     layer2_num :  400
                     layer2_activation :  tanh
```



Experimental setup
Give details on the experimental environment. These items were used in the experiments, but not created or changed by the author. Fill in whatever is relevant to your paper and leave the rest blank.

Relevant hardware details, e.g., system names, makes, models, and key components such as CPUs, accelerators, and filesystems.	
Operating systems and versions (e.g., "Ubuntu 17.10 running Linux kernel 4.13.0")	
Compilers and versions (e.g., "Clang++ v6.0")	
Applications and versions (e.g., "NAMD v2.13" or "SPEC CPU2017")	
Libraries and versions (e.g., "OpenMPI v3.1.0")	
Key algorithms (e.g., "conjugate gradient")	
Input datasets and versions (e.g., "Berkeley Segmentation Dataset: Test Image #296059 [color]")	

Optional link (URL) to output from commands that gather execution environment information — see example scripts at https://github.com/SC-Tech-Program/Author-Kit

URL

Brookhaven National Laboratory

Hugging Face

FAIR for Research Software (FAIR4RS)

Findable, Accessible, Interoperable, Re-usable

F1. Software is assigned a globally unique and persistent identifier.

F1.1. Components of the software representing levels of granularity are assigned distinct identifiers: implementation dependent

F1.2. Different versions of the software are assigned distinct identifiers

I.2. Software tools can interoperate via common support for the data they exchange.

R2: The ultimate goal of FAIR is to enable and encourage the use and reuse of software. To achieve this, software should be well described (by metadata) and appropriately structured so that it can be replicated, combined, reinterpreted, reimplemented, and/ or used in different settings.

Restricted examples:

FAIR executable, not source code, is available

FAIR source code with platform is available within a restricted community

Katz, 2021. doi: 10.1016/j.patter.2021.100222.
Hong, 2022. https://doi.org/10.15497/RDA00068

FAIR4ML Research Data Alliance Interest Group
https://www.rd-alliance.org/groups/fair-machine-learning-fair4ml-ig

FAIR Research Objects: A little bit of packaging goes a long way

Practical lightweight packaging

Aggregate files/directories

+ any content with URI

Embed contextual information

Archive with rich structured metadata

Developer friendly

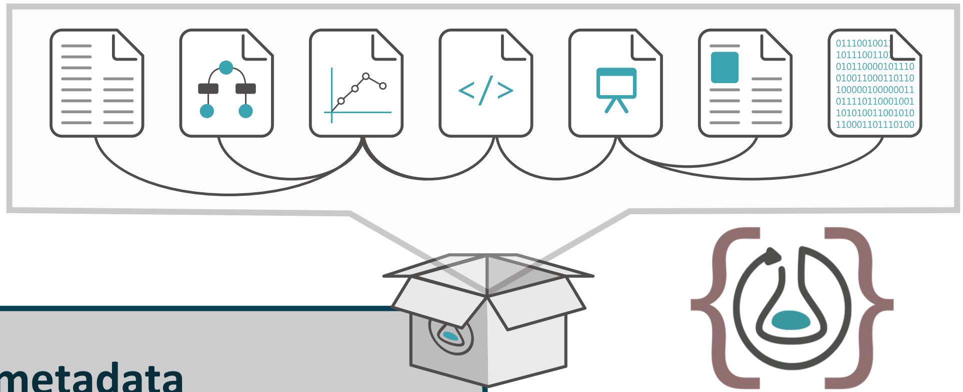
RO-Crate

http://www.researchobject.org/ro-crate/

MANCHESTER 1824
The University of Manchester

UNIVERSITY OF AMSTERDAM
Informatics Institute

Aims of FAIR Research Objects



Describe and **package** data collections, datasets, software etc. **with their metadata**

Platform-independent object exchange between repositories and services

Support **reproducibility** and **analysis**: link data with codes and workflows

Transfer of **sensitive/large** distributed datasets with persistent identifiers

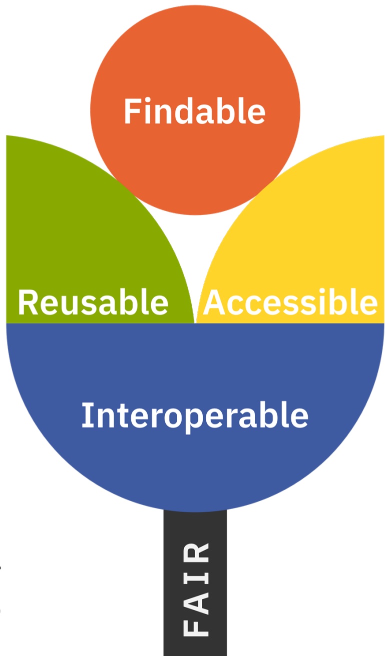
Aggregate **citations** and **persistent identifiers**

Propagate **provenance** and **existing metadata**

Publish and archive **mixed objects** and references

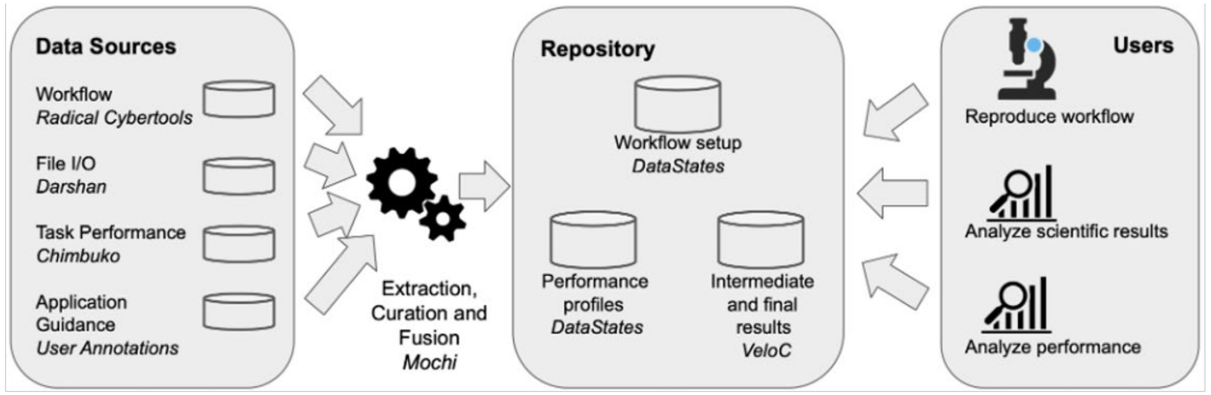
Reuse existing **standards**, but hide their complexity

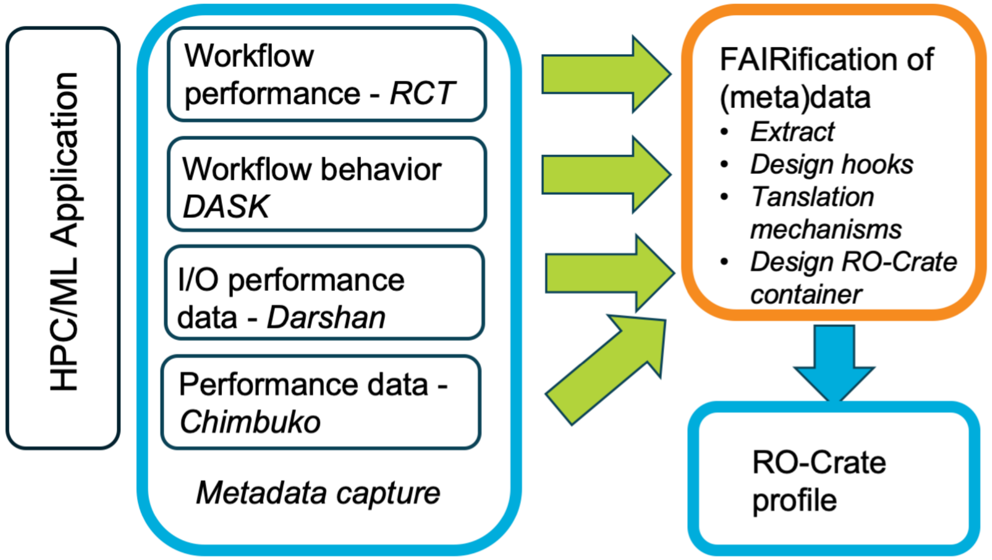
Courtesy Stian Roiland-Seyes, the U of Manchester



researchobject.org

FAIRification of performance data in RECUP





HPC/ML Application
Workflow performance - RCT
Workflow behavior DASK
I/O performance data - Darshan
Performance data - Chimbuko
Metadata capture
FAIRification of (meta)data
• Extract
• Design hooks
• Tanslation mechanisms
• Design RO-Crate container
RO-Crate profile

1) identify and capture the rich information necessary for reproducing hybrid workflows at scale: fuse, organize, store, index
2) make the captured information FAIR to enable key workflow reproducibility tasks: re-runs, re-use workflows, data
3) use the (meta)data to isolate where one workflow's execution deviated relative to another
4) design reproducibility metrics for scientific and performance results

https://sites.google.com/view/recup-reproducibility/home

Brookhaven National Laboratory

Nicolae B, **Pouchard** L, Ross R, et al (2023) Building the I (Interoperability) of FAIR for Performance Reproducibility of Large-Scale Composable Workflows in RECUP. 2023 IEEE 19th International Conference on e-Science (e-Science). https://doi.org/10.1109/e-science58273.2023.10254808

Chimbuko provenance and data analysis

The Provenance DB supports post-hoc performance analysis queries for optimization and debugging

Collects information asynchronously about detected anomalies from an anomaly detection algorithm at each worker node

- Total execution time and links to parent and child execution
- Function execution time
- Associated communication events
- Full function stack
- Execution environments, code version, compile and runtime configurations
- For GPU kernels: device information and parent CPU function
- Window of N(5) events occurring prior and post anomaly
- The Provenance DB is connected to AD nodes with iSonata from the MOCHI team

C. Kelly, L. **Pouchard**, *et al.*, "Chimbuko: A Workflow-Level Scalable Performance Trace Analysis Tool," in Best Paper, *ISAV'20* collocated SC20 Nov. 2020, pp. 15–19. doi: 10.1145/3426462.3426465.

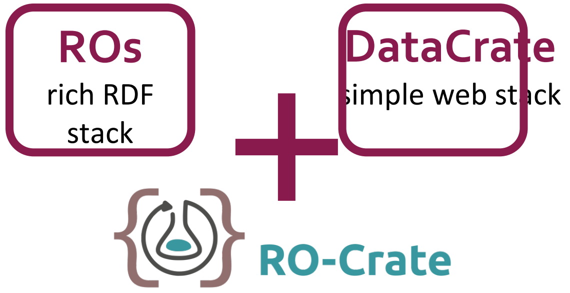
Pouchard *et al.*, "Prescriptive provenance for streaming analysis of workflows at scale," in *2018 New York Scientific Data Summit (NYSDS)*, Upton, NY, Aug. 2018, pp. 1–6. doi: 10.1109/NYSDS.2018.8538951

FAIR Metadata Collection with a Workflow Management System

Goal: Enable a unified, FAIR-enabled metadata format that captures task details (dependencies, execution order, performance metrics, inputs and outputs, etc.)

Tool development

- A schema based on RO-Crate profiles (packaging system for Research Objects that annotates ZIP archives using JSON-LD)
- A method to reconcile the diversity of tools and unify metadata using *translations* and *mappings:*
  - RCT is task-focused and has no awareness of files
    - Service feature added
    - Enable task data annotations (inputs, outputs) that are automatically captured into RO-Crates
  - RCT uses a Pipeline-Stage-Task model where Stages have fixed order of execution, Tasks don't
    - Hierarchical element added in RO-Crate profile
  - Chimbuko/Darshan are data-focused
    - More direct implementation
  - Reconcile partial profiles into one



```
"@context": "https://w3id.org/ro/crate/1.1/context",
"@graph": [
    {
        "@id": "./",
        "@type": "Dataset",
        "datePublished": "2023-08-14T17:34:48+00:00"
    },
    {
        "@id": "ro-crate-metadata.json",
        "@type": "CreativeWork",
        "about": {
            "@id": "./"
        },
        "conformsTo": {
            "@id": "https://w3id.org/ro/crate/1.1"
        }
    },
    {
        "@id": "#f1b5f334-6338-45f5-9052-939f62636698",
        "@type": [
            "ComputationalWorkflow"
        ],
        "hasPart": [
            {
                "@id": "#fapi.c"
            },
            {
                "@id": "#nwchem.F"
            },
            {
                "@id": "#274fb538-dd65-455b-abb2-23d90965a126"
            }
        ],
```

https://github.com/infispiel/RECUP-ROCrate-Library
Polina Shpilker, PhD candidate, Tufts U

Performance Reproducibility Analytics

Goal: Understand what performance knobs affect performance reproducibility using workflow management systems.

Development:

- Metadata available across three sources of information: Darshan logs, Yappi profiles, and Dask's own report of function call streams.
- Metadata identified: filename, line no, #of times, total inclusive time, cumulative time, percall time, who called this function, nbytes, POSIX counters, and many more in Darshan logs.
- Are these enough to identify run-to-run performance variation?

➢ *11 Per-set PAPI counters on Polaris, 108 different configurations*

➢ *Extending LLNL perfdump library*

➢ *Exploring new graph-based method of analysis*

Takeaways

FAIR is not synonymous with Open Science, FAIR basically equates to Persistent Identifiers and metadata, customization depending on purpose is necessary

A systematic approach and tools are needed to manage the entire life cycle of FAIR data and metadata in the DOE complex

To enable efficient analytics and multiple runs from various perspectives, numerous composable tools are needed

To design for the composability of tools and with efficient metadata collection from inception are key to FAIR interoperability

Thank you for your attention

Line Pouchard, BNL
PI

Tanzima Islam, TxState

Bogdan Nicolae, ANL

Rob Ross, ANL

Phil Carns, ANL

Matthieu Dorier, ANL

Amal Gueroudji, ANL

Huub Van Dam, BNL

Shantenu Jha, BNL

Mikhail Titov, BNL

Tianle Wang, BNL

Byung-Jun Yoon, BNL

Ozgur Kilic, BNL

Chris Kelly, BNL

Kevin Assogba, RIT

Polina Shpilker, Tufts

Omnia Sharan, TAMU

Nafiz Abeer, TAMU

Chase Phelps, TxSt

Ankur Lahiry, TxSt

6 Graduate Students

Brookhaven National Laboratory

The authors gratefully acknowledge the funding support from the U.S. Department of Energy Office of Science and employees of Brookhaven Science Associates, LLC under Contract No. DESC0012704

Additional slides

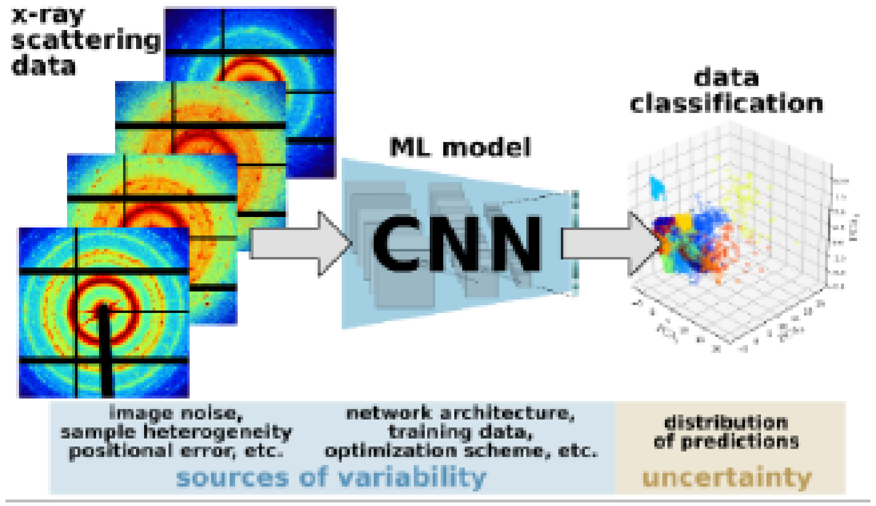
Uncertainty-aware reproducibility metric

- We consider the distribution of prediction results in complex workflows involving AI/ML
  (i.e., characterize the uncertainty of QoI)

- Method: uncertainty quantification (UQ) based on a Bayesian paradigm - quantify the impact on the reproducibility of QoI

- Advantages:
  - Efficacy, uncertainty-aware, goal-oriented
  - available physics-informed quantities
  - knowledge exists that can contribute to the design of priors

- Assess the impact of various potential factors of variability for inverse problems

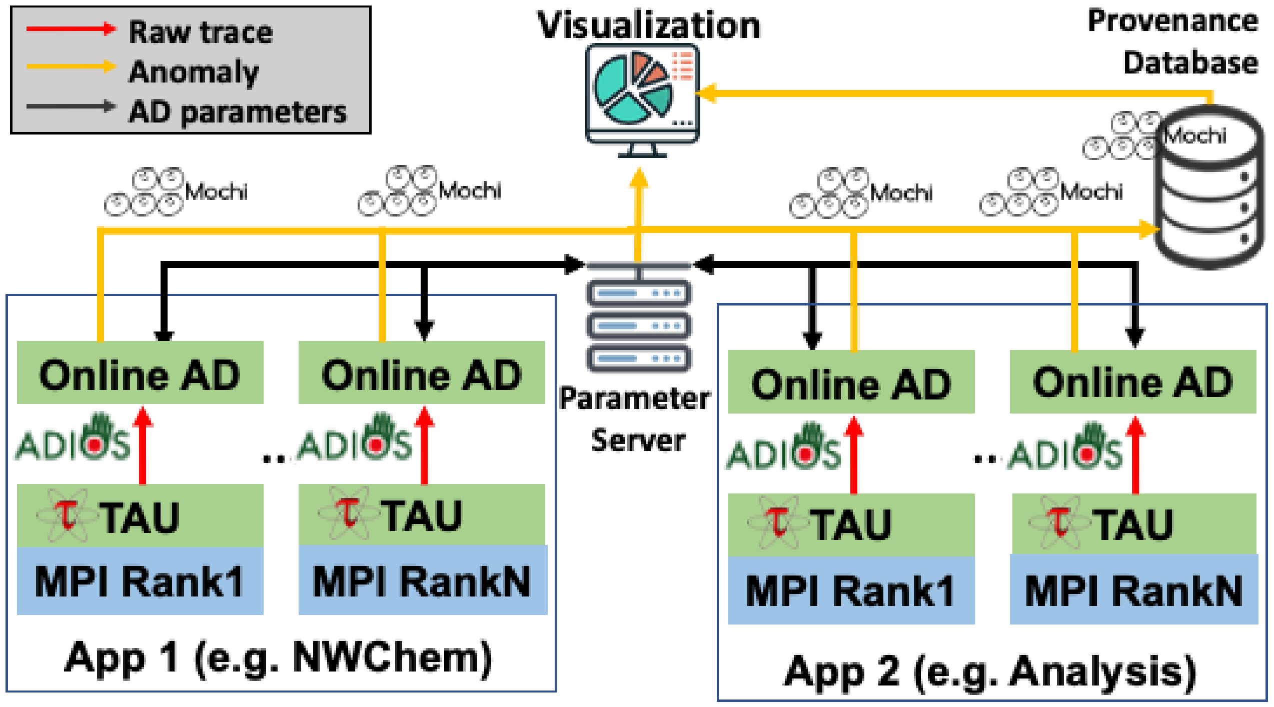

x-ray scattering data
ML model
CNN
data classification
image noise, sample heterogeneity positional error, etc.
network architecture, training data, optimization scheme, etc.
sources of variability
distribution of predictions
uncertainty

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

Pouchard, Reyes, Alexander, Yoon (2023) A rigorous uncertainty-aware quantification framework is essential for reproducible and replicable machine learning workflows. DOI: 10.1039/D3DD00094J

The ECP Chimbuko performance analysis framework

- Chimbuko is a performance analysis framework that provides real-time, distributed, in-situ anomaly detection on application traces.
- Where are the performance anomalies coming from?
- From which workflow component, node, function?
- Can I quickly find the environmental conditions where anomalies are detected?
- Has a similar error been encountered before in previous runs?
- Have the HW and SW stack on this system changed since my last run?

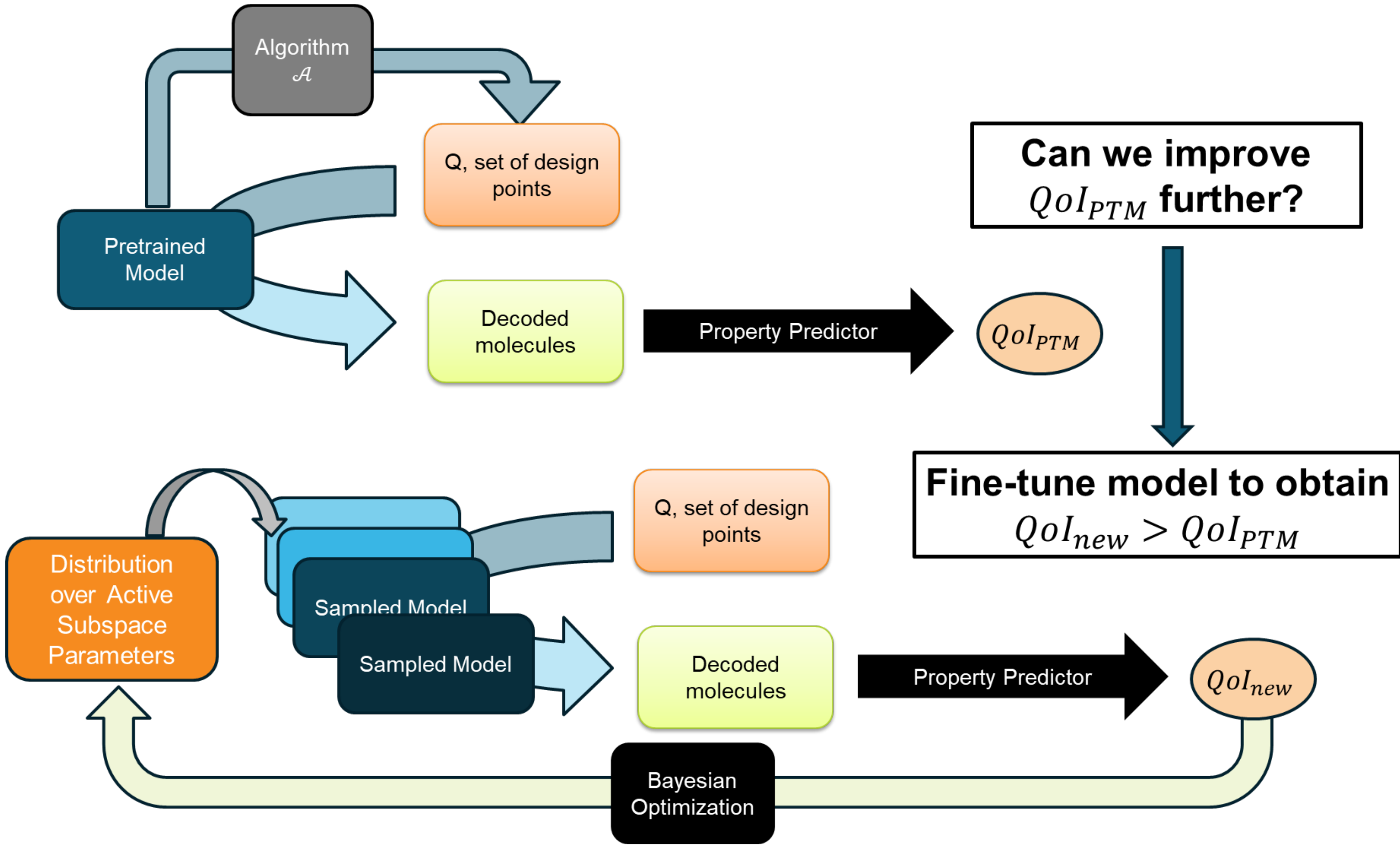


C. Kelly, L. **Pouchard**, *et al.*, “Chimbuko: A Workflow-Level Scalable Performance Trace Analysis Tool,” in Best Paper, *ISAV’20* collocated SC20 Nov. 2020, pp. 15–19. doi: 10.1145/3426462.3426465.

Scientific results reproducibility | Stage-1 UQ

Leveraging active subspace for improving QoI

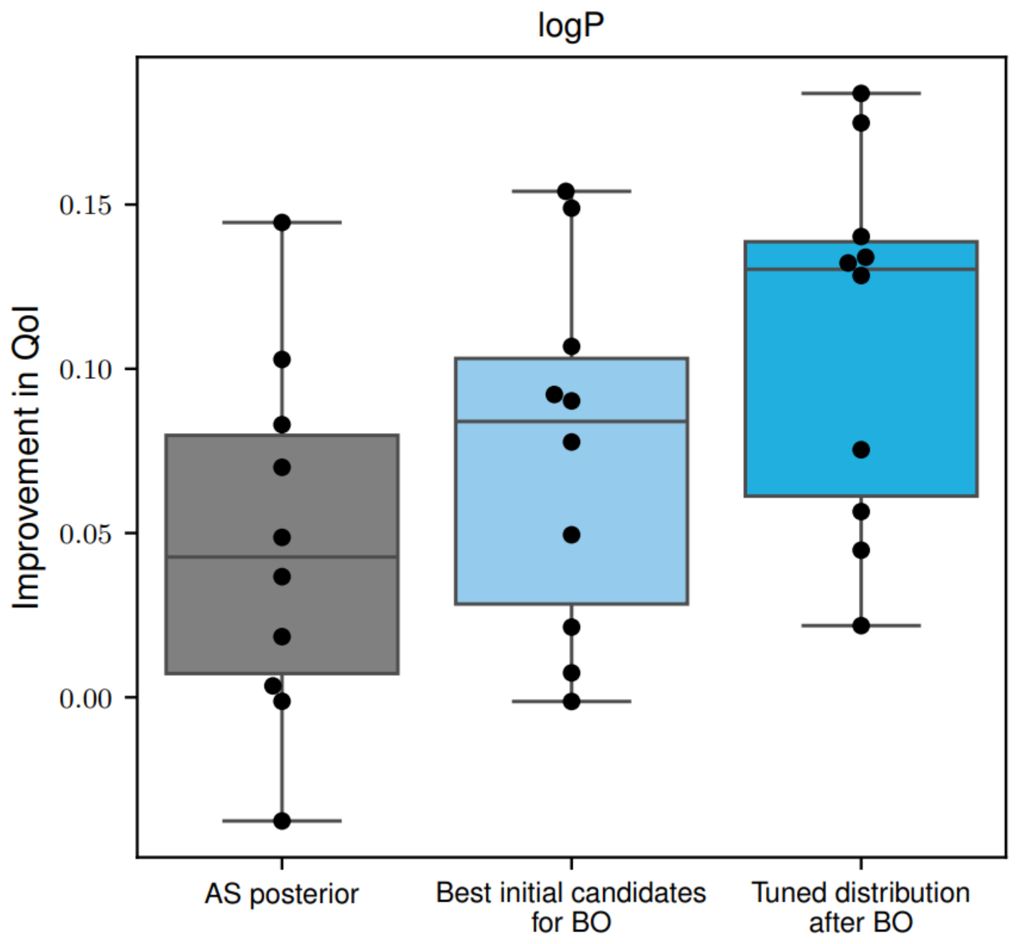
- Active subspace to **capture model uncertainty**
- **Exploration of model uncertainty** class via Bayesian optimization to **boost downstream performance** in molecular design tasks with JT-VAE.



Brookhaven National Laboratory

Can we improve QoI for fixed Q by tuning pretrained model?

- **Improvement in QoI:** $QoI_{new} - QoI_{PTM}$
- Each boxplot (from left to right) shows the improvement due to
  - AS **posterior** distribution
  - best **initial candidate** distribution
  - **tuned distribution** found by Bayesian optimization
- Positive improvement corresponds to increase in the property values.


logP
Improvement in QoI
AS posterior
Best initial candidates for BO
Tuned distribution after BO