scientific data

Check for updates

OPEN

ARTICLE

ML-extendable framework for multiphysics-multiscale simulation workflow and data management using Kadi4Mat

Somnath Bharech[1], Yangyiwei Yang[1]✉, Michael Selzer[2], Britta Nestler[2] & Bai-Xiang Xu[1]✉

As material modeling and simulation has become vital for modern materials science, research data with distinctive physical principles and extensive volume are generally required for full elucidation of the material behavior across all relevant scales. Effective workflow and data management, with corresponding metadata descriptions, helps leverage the full potential of data-driven analyses for computer-aided material design. In this work, we propose a research workflow and data management (RWDM) framework to manage complex workflows and resulting research (meta)data, while following FAIR principles. Multiphysics-multiscale simulations for additive manufacturing investigations are treated as showcase and implemented on Kadi4Mat – an open source research data infrastructure. The input and output data of the simulations, together with the associated setups and scripts realizing the simulation workflow, are curated in corresponding standardized Kadi4Mat records with extendibility for further research and data-driven analyses. These records are interlinked to indicate information flow and form an ontology-based knowledge graph. Automation scheme for performing high-throughput simulations and post-processing integrated with the proposed RWDM framework is also presented.

Introduction

Materials science stands at the forefront of numerous technological innovations spanning across various industries, with a particular emphasis on its engineering background. It has evolved from its empirical and experimental roots, which focused on engineering the chemical composition and the microstructure of materials to achieve specific properties tailored for certain applications, to embracing modeling and simulation as another aspect in the new century, revolutionizing the field with computer-aided material design. This modern approach significantly accelerates the lifecycle of material innovation while reducing costs, time, resources, and energy waste, marking a significant advancement in the pursuit of sustainable and smart material development[1–3]. The vast disparity in scales and the interdisciplinary nature of material modeling and simulation present fresh challenges in this domain, as materials exhibit behaviors across a wide range of spatial and temporal scales, which collectively influence their overall properties. Addressing these phenomena demands a variety of theoretical methodologies, each adhering to certain physical principles at corresponding scale. In other words, the multiphysics and multiscale frameworks are required to fully elucidate material behaviors across all relevant scales. As a result, extensive data bonded with corresponding physical principles at varying scales are normally anticipated in a material modeling and simulation attempt. These data can be roughly classified into three types: (1) input data, which are the pre-requisite quantities and geometries describing the raw/pure materials, intrinsic structures and physical conditions to initiate certain physical processes at corresponding scale; (2) output data, which are the direct/post-processed quantities and geometries presenting the response of the physical process according to certain input data; (3) auxiliary/associated data, which are not related to the input/output of a simulation, but give the necessary information to reproduce the output from a given input. To follow the

[1]Division Mechanics of Functional Materials, Institute of Materials Science, Technical University Darmstadt, Otto-Berndt-Strasse 3, Darmstadt, 64287, Germany. [2]Institute for Applied Materials (IAM), Karlsruhe Institute of Technology (KIT), Kaiserstrasse 12, Karlsruhe, 76131, Germany. ✉e-mail: yangyiwei.yang@mfm.tu-darmstadt.de; xu@mfm.tu-darmstadt.de

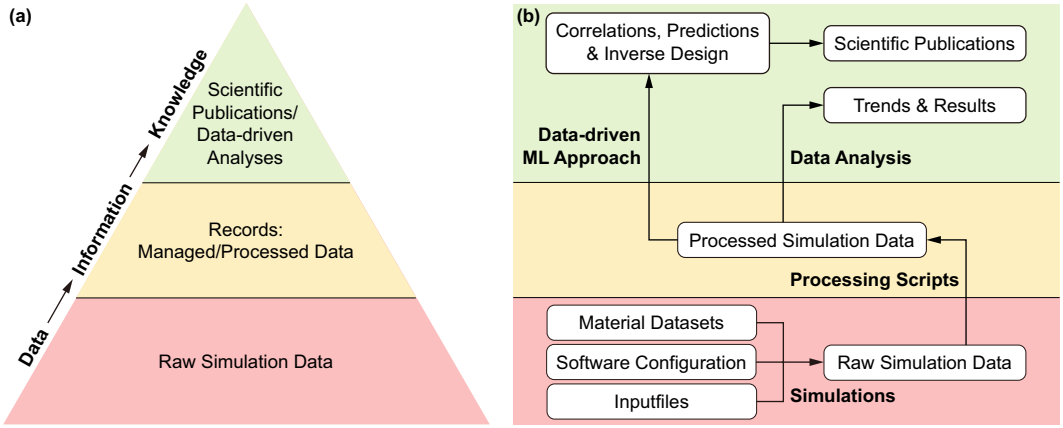


Fig. 1 (**a**) Data-Information-Knowledge (DIK) hierarchy of research data inspired by Chaffey and Wood, 2005 (Chapter 5)[13,44]. This hierarchy is analogous to a typical research data lifecycle, which starts as sets of discrete data. The data is processed into information and then further condensed into knowledge, which is generally documented in form of research articles. (**b**) Workflow of a simulation-based research, where various (meta) data aid the generation of raw simulation data. This data is further processed using scripts and macros on various software. The processed data is further analysed to form meaningful correlations, predictions and trends, which are ultimately published as scientific research.

state-of-the-art FAIR (findable, accessible, interoperable and reusable/reproducible) principles for data sharing[4], all three types of data should be collected and recorded as close to the data producing source as possible[5], leading to the proper design of the data infrastructure with considerations such as the efficiency, readability, extendibility, and reliability.

Meanwhile, with the rapid development of the high-performance computing clusters, it is possible to perform material simulations in a high-throughput fashion, i.e., numerous extensive simulation tasks are simultaneously executed, targeting on the objectives requesting vast volume of data, e.g., delivering the process-microstructure-property (PMP) relationships for the manufacturing of a certain material system[6]. For instance, in additive manufacturing realized by powder bed fusion (PBF) techniques, over one hundred process parameters directly influence the final products[7,8]. The most critical ones, including beam power, scan speed, beam diameter, layer thickness, hatch distance, and scanning strategies, need to be adjusted for each individual build, considering the specific material and geometry. In such cases, data-driven analyses based on statistics and/or machine learning (ML) are generally adopted to extract the PMP relationships of the targeted material system. It has been proven that data management following the FAIR principles is a key to perform scalable ML-based researches, as it readily compacts data describing the raw/pure material, the process parameters and conditions, and the response/effective properties of the processed materials, achieving the data-centric ML analyses[9,10]. Meanwhile, many modeling and simulation methods may have to be integrated as one workflow recapitulating essential factors from various scales in a single material process, it is then essential to manage not just the data involved in a simulation workflow, but also scripts or protocols that realize the workflow in an automatic way enabling the high-throughput computations (HTC)[11]. This can help to adapt the established simulation workflow for similar material systems while retaining complete reproducibility, fostering collaborative research and efficient knowledge transfer. Beyond these, the effective management and curation of data, coupled with simulation workflows adhering to FAIR principles, is also foundational to both scientific accountability and the robust validation and verification of research findings[12].

Following the generalized data-information-knowledge (DIK) hierarchy as introduced by Chaffey and Wood[13], a comparison can be made with simulation-based investigations. As shown in the DIK hierarchy in Fig. 1, data is viewed as a discrete set of facts that, when processed, is transformed into information. Further analysis of this information leads to knowledge. This vertical transformation of data is represented using a pyramid, which also signifies the condensation of volume as the data gradually transform into knowledge. Likewise, in a typical simulation-based research, raw simulation data forms the foundation of this hierarchy and needs processing for visualization. Further analysis leads to insightful trends which are usually well-documented in form of scientific publications, as schematically represented in Fig. 1. In order to maintain the comprehensiveness of the recorded knowledge, it is important to identify and recognize supporting items such as material datasets, software configuration and input parameters used in the simulations, along with pre- and post-processing scripts applied to the raw and the processed data. This strategy for research workflow and data management (RWDM) concurs with the input-process-output (IPO) concept introduced by Griem *et al.*[14]. They describe an atomistic approach where research processes can be iteratively structured as tasks and those tasks are further represented as horizontal transformation with three generic components: (1) Input, (2) Process and (3) Output. Applying the IPO concept throughout the different stages of the DIK hierarchy ultimately enables us to represent the complete research process. Therefore, it becomes evident that an effective RWDM framework must include both horizontal as well as vertical components of a research investigation.

Kadi4Mat, the Karlsruhe Data Infrastructure for Materials Science, is an open-source research data infrastructure developed by Karlsruhe Institute of Technology[15,16]. It utilizes *records*, which are essentially digital objects, as the fundamental building blocks for the infrastructure to store and manage research data. The records are uniquely identified with their *persistent identifiers (PID)* and can hold associated metadata alongside the data itself. Kadi4Mat offers various features to organize and manage (meta)data effectively. Records from an investigation can be grouped together to form *collections* with further sub-categorization using *child collections*. Additionally, customizable *templates* help maintain consistency and standardization within records. The curated research data, in the form of records, can be visualized as a knowledge graph, where individual records are linked based on their relationships. This promotes data exploration and understanding of complex relationships within the research data. Kadi4Mat implements a role-based access control for the records. Owners can set permissions for *users* or *groups* based on their predefined roles like administrator, editor, collaborator and member. This selective access control ensures data security during the course of the investigation, meanwhile enhancing collaboration among researchers and scientific staff at different access levels. On the other hand, the research data records can be published for broader accessibility, and can even be published on open repositories such as Zenodo for universal access. At present, Kadi4Mat hosts 100+ users from a variety of institutions, who have contributed more than 480 publicly accessible records. Apart from Kadi4Mat, there is a variety of electronic laboratory notebook (ELN) based research data management (RDM) tools available, such as LabArchives[17], labfolder[18], NOMAD[19] and eLabFTW[20,21]. Although most of them provide common functionalities for RDM as Kadi4Mat such as data integrity, data and research security, version control and team collaboration, they lack in one way or the other in comparison with Kadi4Mat. Some of them are commercial software, while Kadi4Mat is open source. In addition to serving as a ELN, Kadi4Mat also functions as the data repository distinguishing it from the RDM tools that primarily function as lab notebooks. In addition to the web interface, Kadi4Mat also provides programmatic access through its python-based application programming interface (API) called Kadi^{APY}[22]. This enables potentially automated interaction with Kadi4Mat using personal access tokens (PAT), facilitating seamless integration of RWDM workflows with HTC workflows. The research data generated from such investigations often needs to be exported for ML analyses. The data can be fetched and processed directly into the ML models/algorithms using the API access or can simply be transferred to Kadi4Mat-hosted ML utilities and applications such as KadiStudio[14], KadiAI and CIDS[23]. These features make Kadi4Mat the most appropriate infrastructure for the data management needs of complex investigations, like the showcase presented in this work.

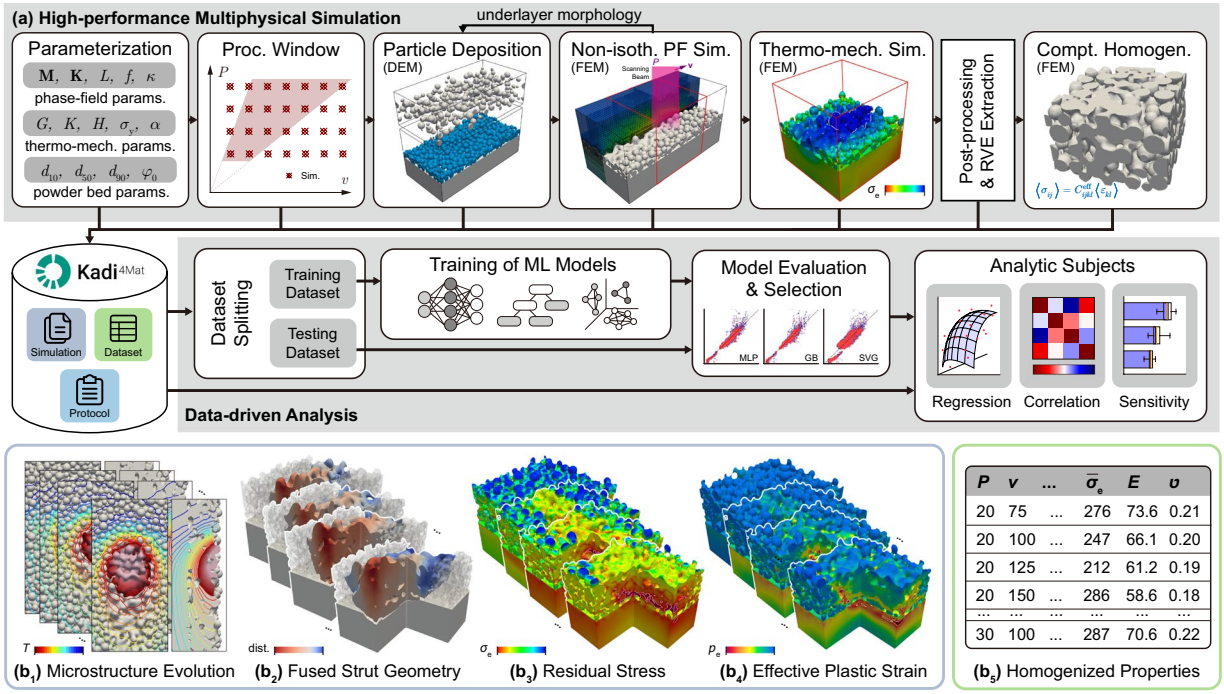
In this work, we present a RWDM framework which is implemented for our recent numerical investigation on establishing PMP relationships during PBF process using Kadi4Mat[24,25]. During the RWDM process of this investigation, the crucial steps involved in the workflow, the identification, collection and organization of (meta)data, their recording and crosslinking to indicate the information flow, will be discussed. Ontology-based knowledge representation of the overall investigation using the records and capturing their relation promotes further expansion and usage of the research database, as the relations are machine-readable and can be machine-actionable as well. Automation in the implementation of the proposed RWDM framework and the further usage of the curated data will also be discussed.

Results

RWDM infrastructure design. The RWDM framework outlined in this study comprises of workflow (including simulation sub-routines) management, (meta)data identification and curation. The implementation of this framework is illustrated by curating the research workflow and the data generated from our recent works on multilayer PBF simulations[24,25]. The simulation workflow is explained in the methods section and is also visually summarized in Fig. 2(a). Sample results from this workflow are schematically displayed in Fig. 2(b_{1-5}). The four-layer PBF process simulations using non-isothermal phase-field model results in the evolution of the thermo-microstructure containing the temporal information of the fused strut. Subsequent thermo-elasto-plastic calculations are performed to capture the evolution of residual stress and plastic strain in the thermal microstructure from the process simulations. The effective mechanical properties of the PBF processed microstructure are calculated using the computational homogenization method. As implied in Fig. 2(b), the research data generated during the multilayer PBF investigation was carefully collected, organized and stored into Kadi4Mat in form of records. Three distinct record types are utilized to curate the research data from the multilayer PBF work: (1) dataset, (2) protocol and (3) simulation records. From technical point of view, these various types of records are on the same level in the RWDM infrastructure. However, from managerial aspect, they are distinguished by their data content, as outlined in Table 1. The usage and integration of these different record types in this RWDM infrastructure will be further discussed in detail in the following subsection.

An overview of the curated data from the multilayer PBF work can be seen in form of an ontology-based knowledge graph, in Fig. 3. The research workflow and data curated from multilayer PBF simulations consists of several records, indicated by circular nodes. These nodes can be seen arranged prominently, to visualize the child collections representing the three clusters of simulations: (1) phase-field simulations, (2) thermo-mechanical simulations and (3) computational homogenization. The records are labeled with their respective identifiers and the record types. The record type can also be identified by the node colors.

Data records design. The dataset records, for example `@mfm_materials_ss316l` contains the temperature-dependent material properties of the material SS316L that were used for phase-field simulations, thermo-mechanical simulations and computational homogenization, as shown in Fig. 4(a_3). Temperature dependent material properties of SS316L, such as, Poisson's ratio, Young's modulus, yield stress, plastic modulus and thermal expansion coefficient, employed in the thermo-mechanical anaylses and computational homogenization simulations are stored in the metadata fields. Other material properties such as melting temperature, thermal conductivity, specific heat capacity, latent heat of fusion, diffusivities and mobilities, employed in the



P	v	...	$\bar{\sigma}_e$	E	υ
20	75	...	276	73.6	0.21
20	100	...	247	66.1	0.20
20	125	...	212	61.2	0.19
20	150	...	286	58.6	0.18
...	...	...	...	...	...
30	100	...	287	70.6	0.22

Fig. 2 Schematic of the research framework consisting of (**a**) Multiphysical simulation scheme for multilayer powder bed fusion simulations including paramterization for phase-field model, thermomechanical analysis, process parameter selection for PBF, powder bed parameters and deposition, process simulations using non-isothermal phase-field model, thermomechanical analysis, RVE selection for homogenization of mechanical properties; (**b**) Curating the relevant (meta)data in form of several record types, namely simulation, dataset and protocol in Kadi4Mat, to be further used in data-driven analysis. Illustration of (b_{1-4}) the simulation results stored in simulation records and (b_5) data stored in form of dataset records.

Record type	Data stored
Protocol	Workflows, sub-routine, technical SOPs, relevant macros/scripts
Simulation	Inputfiles, output files, relevant parameters
Dataset	Material-specific parameters, data from secondary sources, aggregated output data

Table 1. Types and the corresponding research (meta)data stored.

phase-field process simulations are curated as well. The effective mechanical properties obtained from the computational homogenization simulations, such as Young's modulus, bulk modulus, shear modulus and Poisson's ratio, along with their process parameters (beam power and scan speed) and their resulting porosities are summarized in a CSV file attached to the dataset record. Protocol records, such as `@inputfile_pf` utilize the description field to document the workflow to generate input files with a new layer of deposited powder particles, as shown in Fig. 4(a_1). This particular workflow involves several sub-routines, including: conversion of phase-field based microstructure to voxel based one, importing the voxelized microstructure into discrete element method (DEM) software (e.g., GeoDict[26], YADE[27]), depositing a new layer of powder over the previously processed layer, exporting the center and radii information of the newly formed powder bed, and finally, using it to generate inputfile to process another layer of powder bed. The relevant metadata like software versions and powder characteristics are stored in the metadata field, while the supporting files like, macros and processing scripts are uploaded to the protocol record as attachments with their usage sufficiently documented in the description field. Simulation records store data related to simulations based on the IPO model. They include inputfiles, process parameters used in the simulations, and summarized output files. For phase-field simulations, simulation records such as `@pf_20-100` illustrated in Fig. 4(a_2), contain inputfiles for each layer, process parameters and setup such as beam power, scan speed, powder bed characteristics, simulation domain size, and simulated variables defining the transient thermo-microstructure. Simulation records for thermo-mechanical simulations include their corresponding inputfiles, process parameters, and transient thermo-microstructure along with the localized stress and strain response. Temporal values of von Mises stress, stress components, as well as plastic strain and its components are stored in CSV files for each scanned layer. Simulation records for computational homogenization curate their respective inputfiles, process parameters as well as the components of stress and strain tensors, which were utilized to calculate the stiffness tensor and thereafter, the effective mechanical properties.

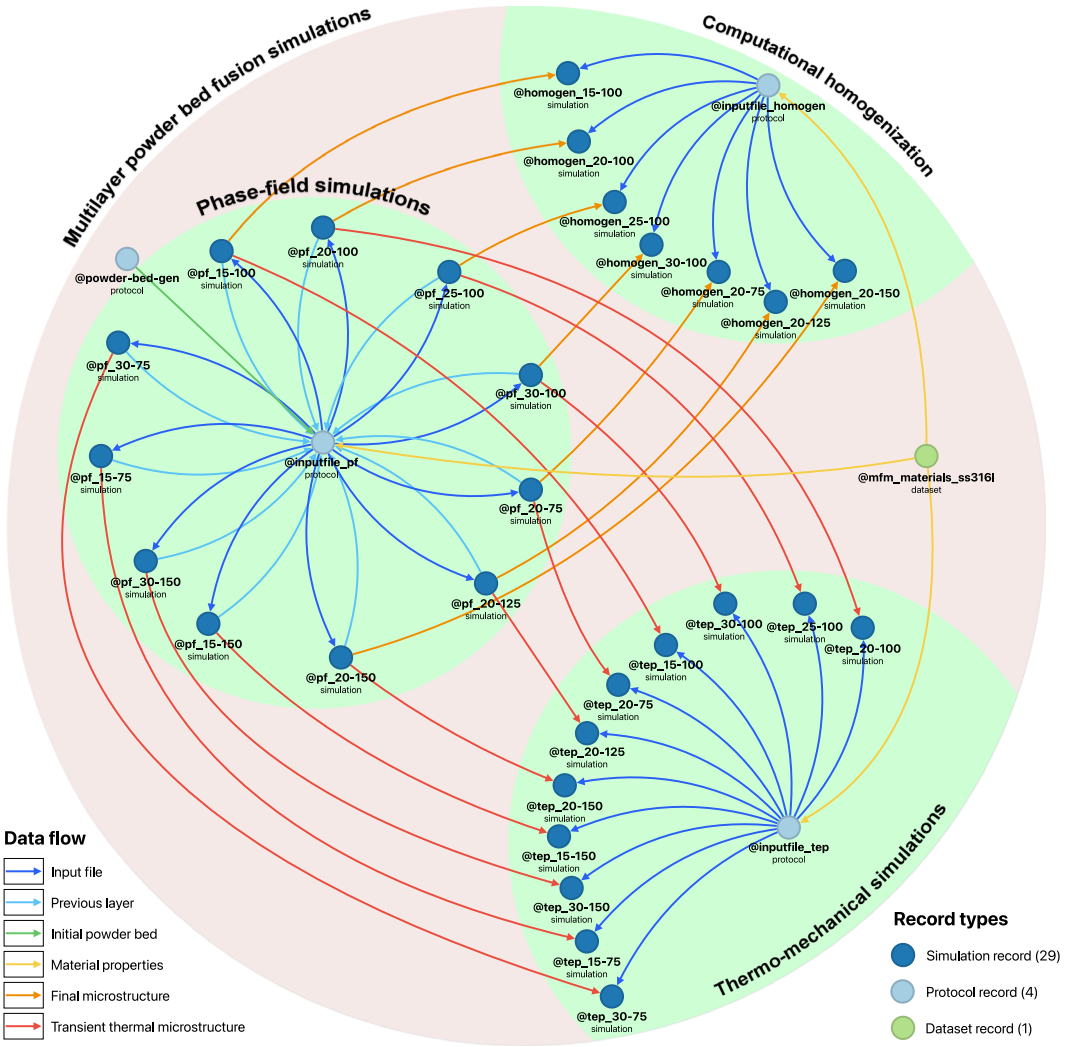


Fig. 3 Knowledge graph visualizing research data from multilayer PBF simulations, including phase-field simulations, computational homogenization, and thermo-mechanical analysis. The data is stored in various record types, such as simulation records, protocol records, and dataset records. The data flow within the records is also indicated.

Records on Kadi4Mat have persistent and unique identifiers that distinguish them and ensure efficient retrieval and management. Upon creation, records are automatically assigned with a numeric PID, which remains unchanged. However, users can assign unique alphanumeric identifiers to further distinguish the records. The nomenclature for these identifiers is chosen to concisely represent the content of the record, such as material properties or simulation setup and results. In case of dataset records, the identifier name reflects the type of data as well as the material system. The identifiers of protocol records represent the procedure documented within them. Data from the three types of simulations with varying process parameters is curated in form of simulation records. Their identifiers take the form `@<type>_<P>-<v>`, to represent the simulation type as well as the distinguishing process parameters, in this case, the beam power `<P>` and scan speed `<v>`. `<type>` is a placeholder for the simulation type. For instance, `@pf_20-100` would represent the simulation record for a multilayer phase-field simulation with beam power 20 W and scan speed 100 mm s^{-1}. Likewise, `@tep_20-100` and `@homogen_20-100` would be the records for thermo-mechanical simulation and computational homogenization respectively.

The interlinking of records represents the relationships between them and the data flow within them as shown in the edges connecting the nodes in Fig. 3. These linkages can be used to understand the overall simulation workflow. For instance, the protocol record `@powder-bed-gen` explains the procedure to generate the initial layer of powder bed. The centre and radii information of the initial powder bed is then exported to `@inputfile_pf` for generating the inputfiles to simulate the PBF process. `@inputfile_pf` also receives temperature-dependent material properties of SS316L alloy from the dataset record `@mfm_materials_ss316l`. The inputfiles generated from `@inputfile_pf` is sent to the phase-field simulation records, which also store the simulated microstructures of the processed layers. `@inputfile_pf` adds another layer of powder on the previously processed microstructure and subsequently generates another inputfile and sends

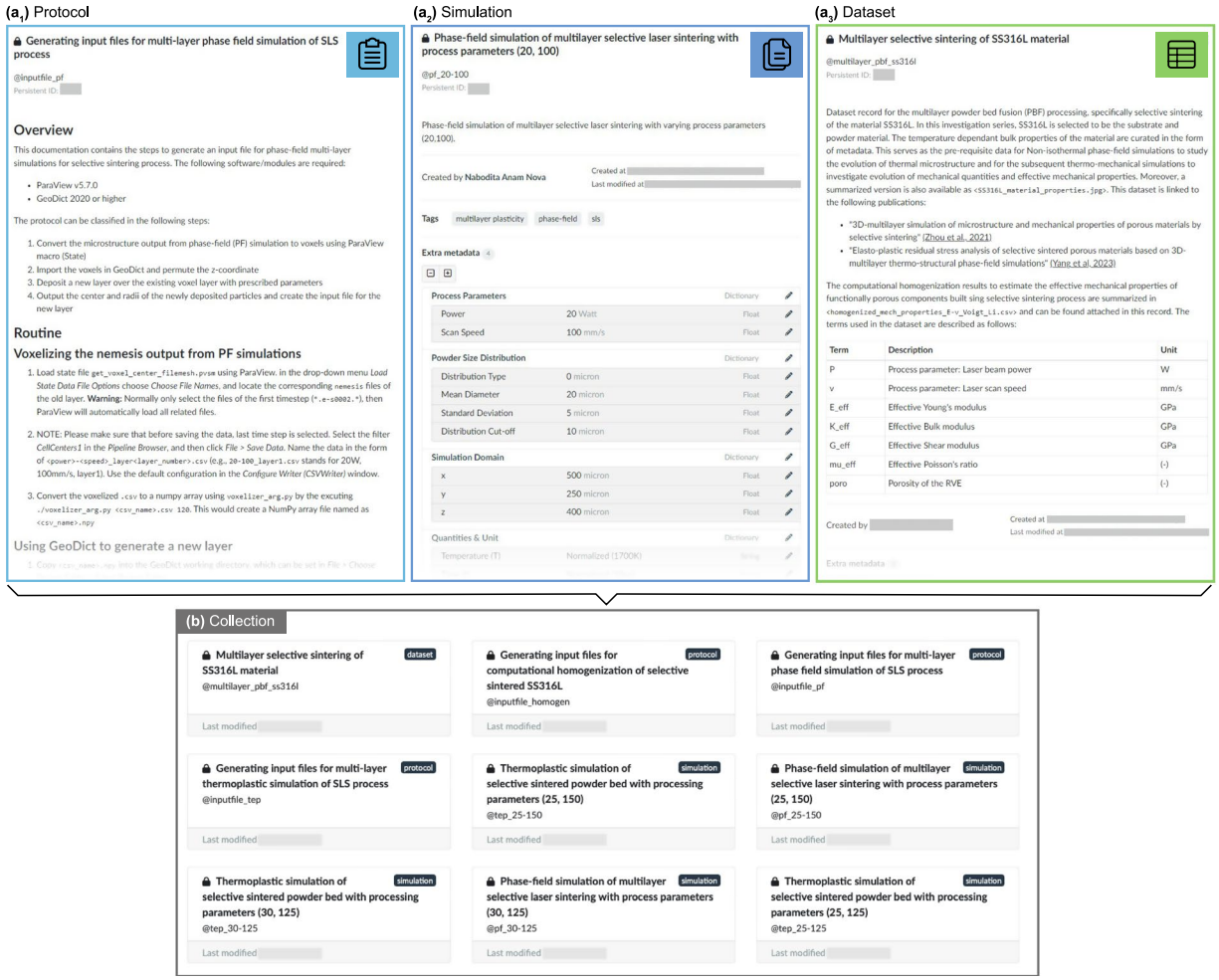


Fig. 4 Snapshot of different record types: ($\mathbf{a}_1$) Protocol record, ($\mathbf{a}_2$) Simulation record and ($\mathbf{a}_3$) Dataset record. These records along with other records form a collection as illustrated in (**b**).

to the phase-field simulation records to process the PBF scan of new layer of powder. The transient thermo-microstructure is delivered to the thermo-mechanical simulation records, where the relevant inputfiles are sent by `@inputfile_tep` upon receiving the material properties from `@mfm_materials_ss316l`. Likewise, for the computational homogenization, the final microstructure is transferred by the phase-field simulation records and the inputfile is created by `@inputfile_homogen` to simulate the effective mechanical properties of the PBF processed parts with varying process parameters.

The research data curated from the aforementioned three simulation clusters is organised into collections on Kadi4Mat. Each collection is identified by the following identifiers: `@pf_sls`, `@tep_sls` and `@homogen_sls`. These collections are published on Zenodo to boost the findability and accessibility of the curated research data. The Zenodo entries contain records in JSON, RDF, and PDF formats, along with their corresponding metadata and file uploads. This comprehensive approach ensures that the records are stored in both human-readable and machine-readable formats, therefore, increasing their interoperability.

Automation of data recording integrated with high-throughput simulations. Collecting and managing (meta)data from high-throughput investigations is of paramount importance for its further analysis and potential data-driven studies, however it can be a daunting task, if not automated. An automated scheme integrating the data curation along with the reproducible workflows for high-throughput simulations is presented here. Automating the data collection step in a HTC workflow not only boosts the efficiency of a laboratory by eliminating the need for an intermediary (i.e. human operator), but also ensures consistency of the data records and compliance with the community-agreed FAIR data standards.

Figure 5 depicts a typical HTC workflow, implemented on a master Jupyter session, with sub-routine scripts executed on the worker computer via ssh and the seamless interaction with Kadi4Mat powered by Kadi^{APY}. In a high-throughput investigation, batch simulations are performed with an array of combinations of processing parameters. The selection of process parameter combinations, often aided by design of experiments, is a critical step and depends on the processing window for the particular process and the process-material relationship. The process parameters and setup are normalized and formulated as metadata by the master console and are then fed into the inputfile composing scripts of the worker console along with the inputfile template pulled

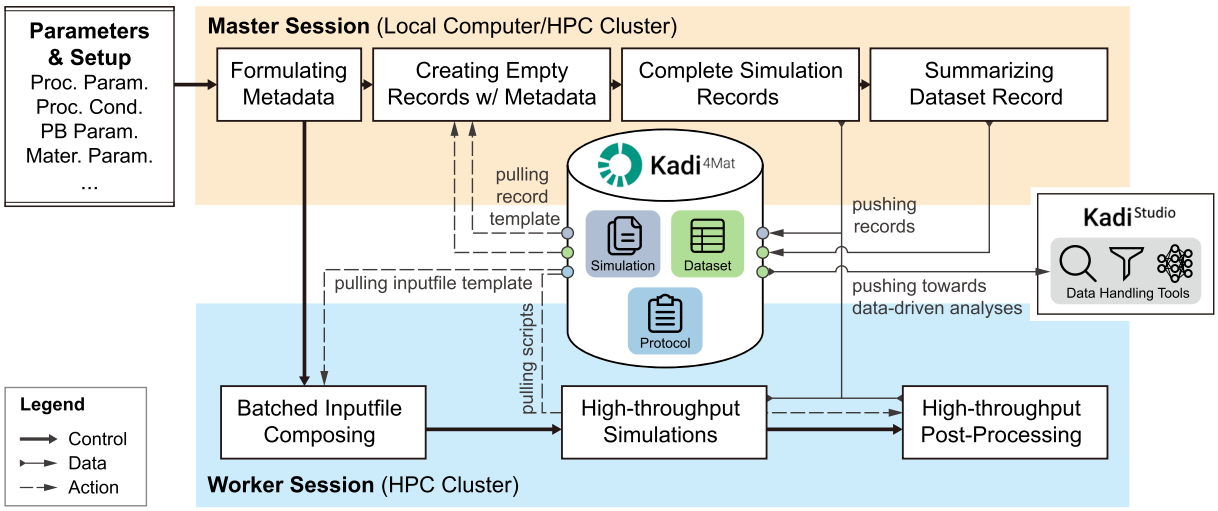


Fig. 5 Flowchart of integrated automation for performing high-throughtput simulation and post-processing together with RWDM framework.

from the protocol records in data repository, to create batch inputfiles corresponding to the parameters. These inputfiles are submitted as batch jobs on the computing cluster. Meanwhile, the master console pulls simulation record template from Kadi4Mat repository to create empty records for each simulation with the parameter information as metadata. Simple Linux utility for resource management (SLURM) workload manager is used to submit the batch simulation jobs on the high-performance computing (HPC) cluster, which assigns each job a unique ID. Timed python scripts are used to check the status of the simulation jobs, completed jobs are verified for successful completion. The successful completion of a simulation job is verified by conducting at least two checks on different levels, i.e. the SLURM report generated using the job ID and the output log file from the corresponding simulation. The data from successfully completed jobs are further sent for processing, while the failed jobs are marked for error resolution. Most commonly occurring errors in the execution of simulation jobs often result from insufficient memory or wall time (time allocated in the job scheduler). The progress of a simulation is curated at particular timestep intervals using checkpoints. Checkpoints are essentially snapshots of the simulation data including meshes, solutions and stateful object data, which are used to resume prematurely terminated simulation jobs. Doing so prevents total loss of progress in case of failed simulation jobs and minimizes the wastage of computational resources. The error handling of failed simulation jobs require human intervention, but handling of common faults such as job cancellation due to calculation ran out of wall time could also be automated using dedicated workflow managers such as AiiDA[28]. The post-processed data is pushed to their corresponding simulation records on Kadi4Mat. The data is further analyzed and summarized into a dataset, which would serve as an end-result of the HTC investigation. This dataset is also pushed as a dataset record in Kadi4Mat and could be retrieved for further data-centric machine learning analysis. In this context, Kadi4Mat serves as a community repository as well as an ELN. Alternatively, the data could be directly fetched into Kadi4Mat's ML workflow suite called KadiStudio[14,23]. KadiStudio has ML modules such as KadiAI and cids-tools to facilitate the development and implementation of data-driven models in ML workflows. These workflows can be documented as Kadi4Mat records to ensure their reproducibility.

Data transfer for machine learning. Once the data is collected and summarized into datasets, it can be utilized for machine learning. Various ML algorithms can be employed based on the data and the desired objective. The selection of an appropriate ML algorithm would be directly affected by the amount and type of available data. As presented in Fig. 6(a), the summarized data from the HTC investigation stored as dataset records on Kadi4Mat, are pulled for further data-driven analysis using python-based Kadi^{APY} commands. The fetched data is scaled and then split into training and testing datasets. The training dataset is used to train the selected ML model, simultaneously the model is tuned by optimizing the hyperparameters to improve the model accuracy, and the testing dataset is used to evaluate the employed ML algorithms. To demonstrate the utilization of summarized dataset from a HTC investigation in data-driven analyses, data from 3375 simulations from a previous investigation based on phenomenological model of PBF is employed[29]. Figure 6(b) presents the schematic of the phenomenological PBF simulation[30], where the thermal evolution in a homogenized powder bed is simulated during a single layer scan of PBF. The ML algorithms used to predict the fused track width from the various input parameters are multilayer perceptron (MLP) and gradient boosting (GB). Figure 6(c),(d) present the plot between the predicted fused track width with the ground truth for MLP and GB algorithms respectively, with their corresponding R^2 values of 99.53% and 98.77%, suggesting good model accuracy.

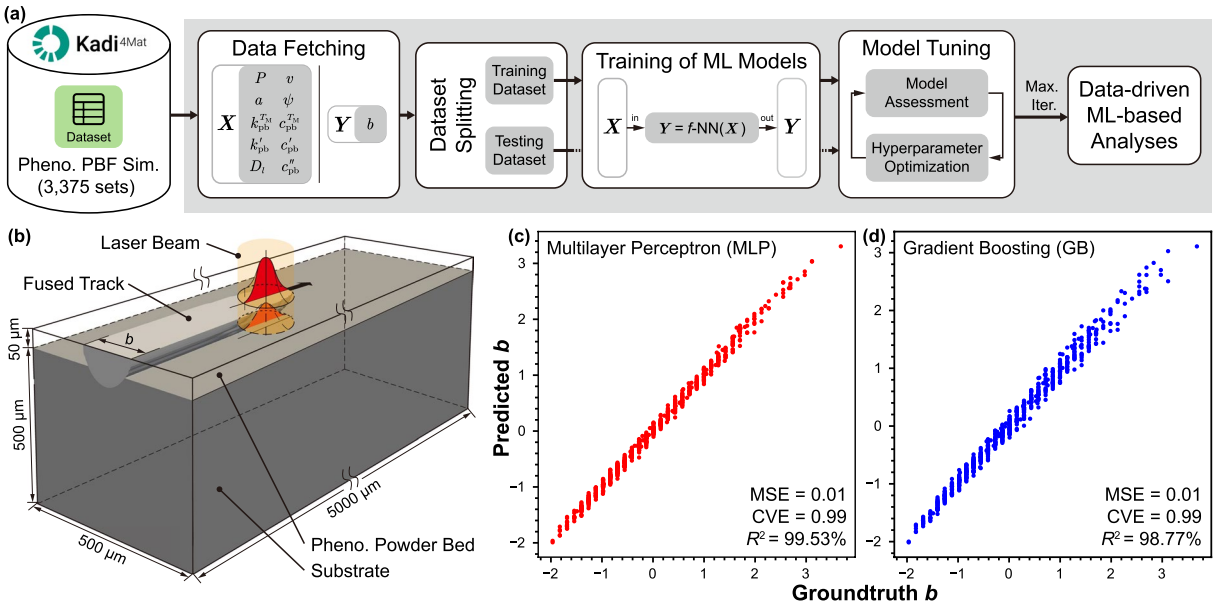


Fig. 6 (**a**) Schematic for utilizing dataset records from Kadi4Mat for data-driven analyses. (**b**) Schematic of a phenomenological PBF simulation used to train models based on (**c**) multilayer perceptron (MLP) and (**d**) gradient boosting (GB) algorithms.

Discussion

The suggested RWDM framework was implemented for the simulation based multilayer PBF investigations. With respect to our investigations on PBF, a total of 47 records are curated in the collection `@multilayer_sls` on Kadi4Mat, part of which is illustrated in Fig. 3, showcasing the interlinking and data transfer between these records. This collection includes 39 simulation records, 5 protocol records, 2 dataset records and a measurement record. The description of each record is stored in the description field, which may explain the metadata fields and attachments, while the associated data is stored as attachments and additional supporting (or auxiliary) data is stored in the metadata fields. There are 16 simulation records each from the multilayer phase-field simulations and corresponding thermo-mechanical simulations, and 7 from the computational homogenization simulations. Standardization and consistency of the records was maintained by employing record templates. Python scripts based on Kadi^{APY} library were used to create multiple records, making the framework capable of scaling up for even larger datasets, with the possibility of automating the data curation step, especially for HTC investigations. The current RWDM database for multilayer PBF studies can be expanded for an even wider range of processing parameters and can be extended to similar material systems.

As the temporal thermo-microstructure resulting from the process simulation was utilized as input in computational homogenization and thermo-mechanical analysis, they can be easily transferred for other multiphysics-multiscale investigations such as nanoparticle migration behavior during PBF[31,32], influence of process parameters on the magnetic properties of AM produced parts[33], thermal anisotropy in porous AM parts[34]. Data from the Kadi4Mat records can be exported in machine-readable formats such as JSON, and can be fetched automatically using Kadi^{APY} scripts. The existing data can be extrapolated and utilized for further data-driven analysis as illustrated in Fig. 2(b). These studies could possibly optimize the fused strut geometry by manipulating the volumetric energy input, predict the part properties for a particular set of processing parameters or tailoring the mechanical properties. Finally, reusing research data generated from computationally intensive simulations for further investigation is a step towards sustainable research.

The data curated through this RWDM workflow could later on serve as reference learning material for bachelor and master's degree students. Access to extensive research data can expedite the learning process for the students[35]. In conjunction with the learning process, access to a reference data would make it easier for other researchers to benchmark their simulation code and setup. The meticulously curated research workflow presented in this work underwent a rigorous test of reproducibility when a master's student was tasked with replicating the results solely by adhering to the documented procedures. With minimal guidance, the student successfully executed the workflows and reproduced the results, demonstrating the robustness and accessibility of the framework. The framework can be customized and extended to fulfill the data management needs of various other research works, and could even facilitate collaboration within multiple levels of researchers. Principal investigators can initialize the customization of the RWDM framework by identifying the needs of their corresponding research projects. Followed by the breakdown of the project goals into tasks, and further breakdown into input, process and output components. These components can be managed using records, with detailed instruction of their usage along with automated macros/scripts in form of protocol records describing sub-workflows. Notably, it is important to set standards such as nomenclature of records, files and metadata descriptors, particularly in the case of collaborative research works. Upon customization of the framework, it can be implemented by research assistants/students or automated, especially for repetitive tasks, thereby

accelerating the overall research workflow. Implementing a similar RWDM framework would be a necessity for large projects like inter-laboratory study (ILS) involving numerous researchers from various research institutes, dealing with wide processing windows and multiple material systems[36].

Methods

Data generation: Simulation scheme and workflow. In this section, the simulation methods used to simulate the evolution of microstructure and the mechanical quantities (like stress and strain) during PBF processing are introduced. During PBF processing, a variety of complex physical phenomena occur that are crucial for developing an understanding of the process. These phenomena have a significant impact on the resultant morphology and properties. Some of these phenomena include thermal behavior (laser beam interaction with powder bed and associated heat transfer), phase transformation (like melting and solidification), mass transfer (through diffusion or advection), and grain coalescence and coarsening. Furthermore, the thermal expansion during the heterogeneous heating and cooling cycles, causing non-uniform thermal stresses in the powder bed during the processing, is also captured. These values have been employed to further calculate the evolution of residual stress and plastic strain in the simulation domain. The characteristic length and time used in the simulations is 1 μm and 1 μs, respectively. Even though most studies approximate homogeneous layers for their single-layer or part-scale multilayer models to analyze thermal history and mechanical response in additively manufactured parts[37–39], this investigation employs powder-resolved layers to preserve the heterogeneity arising from the complex morphology of the powder bed. Due to the presence of stochastic voids in the powder bed, the local microstructure exhibits microscopic variations in the thermal profile, which influence the thermal stresses and the corresponding mechanical responses, such as plastic deformation and residual stress, during PBF processing. These microscopic variations reveal significant physical aspects such as high thermal gradients in the necking regions of the powder bed, concentration of residual stress in the necking regions as well as between two subsequent layers. Furthermore, the homogenized properties obtained from these mesoscopic simulations are comparable to the experimental observations and/or part-scale model results. Based on our former research, a powder resolved non-isothermal phase-field model was employed to simulate the microstructural evolution and corresponding thermo-mechanical response, while considering the aforementioned physical phenonema during single-scan multilayer PBF processing of SS316L parts[24,25,33,40].

The simulation scheme used to comprehensively investigate microstructural evolution and thermo-mechanical analysis is arranged in multiple stages, as shown in Fig. 2(a). It starts with the parameterization and normalization of quantities used by the simulation models, such as phase-field parameters, temperature-dependent material properties of SS316L and Argon atmosphere for thermo-mechanical analysis, and the powder characteristics for powder bed deposition[24,25]. A process window was selected with variation of the two most important process parameters for PBF: Beam power and scan speed. The non-isothermal simulator (NIsoS) program based on the MOOSE framework is employed to implement the non-isothermal phase-field model using finite element method[41]. A simulation subdomain is selected and further imported into the thermo-elasto-plastic model and further into the computational homogenization scheme for calculating the mechanical quantities of the printed parts. The details of the models used and the simulation workflow is sufficiently reported in our previous works[24,25,40].

Data sorting and organization. Diverse multifaceted data is generated throughout the simulation workflow as explained in previous section. The generated data is categorized in three main clusters, each representing a simulation stage from the workflow: (1) Phase-field simulations, (2) Thermo-mechanical simulations and (3) Computational homogenization. Apart from the data generated from these simulations, there is a variety of supporting metadata containing crucial information needed to reproduce these data. Therefore, it is imperative to identify, collect and organize the metadata from each stage of the simulations. Typical metadata for these simulations would be the material- and processing parameters, and their normalization, simulation setup (such as boundary conditions, initial conditions, numerical solver configuration, powder characteristics for the generation of powder bed, software versions, etc). Furthermore, a comprehensive documentation of the sub-routines entailing the techniques and usage of software is necessary to achieve the reproducibility goal from the FAIR guiding principles[4].

Data recording and linking. The huge variety of data generated from the series of simulations during the multilayer PBF investigation is recorded and curated in manner that it follows the FAIR principles[4]. Kadi4Mat is employed to curate our research data in form of records. These records have various fields, such as title, identifier, description, metadata and file attachments to store the data and relevant metadata containing crucial information about the stored data. The records are uniquely identified using an alphanumeric identifier and a persistent numeric identifier. Standardization and consistency among the records are maintained by creating templates, which were then used to create records to store data from simulations with varying process parameters. Kadi4Mat offers the capability to store data in various record types, such as simulation records, protocol records and dataset records, as listed in Table 1. Protocol records were used to document the sub-routines and/or standard operating protocols (SOPs) used for a particular task from the workflow. Dataset records, as the name suggests, were used to store datasets, for example, material properties of SS316L used in the multilayer PBF investigation. The organized data from each simulation cluster was documented in form of simulation records using python scripts based on Kadi^{APY}. This ensured the scalability of our RDM framework by automating the data fetching and recording step to an extent. Linking of records with other records enables the visualization of data flow and exchange between the records, sometimes even depicting usage of particular data in the overall workflows.

Accessibility and publishing of data. The records created on Kadi4Mat have the capability of being shared with the researchers within the Kadi4Mat consortia of institutions. However, for sharing the data with researchers outside the consortia, the records can be exported in various formats, including PDF, RDF and JSON. The records are also exported on Zenodo and is linked to a digital object identifier (DOI), thereby making the data findable and accessible to all. Zenodo supports data storage of up to 50 GB per record. For datasets exceeding this limit, multiple Zenodo records can be created or a third party data repositories can be utilized.

Data availability

The authors declare that the data curated using the RWDM framework described in this study for our multilayer powder bed fusion investigations are available on Kadi4Mat under the collection `@multilayer_sls` with PID: 592. The summarized version of the dataset is published on Zenodo[29,42].

Code availability

Source code for MOOSE-based application NIsoS and related utilities are available via the online repository bitbucket.org/mfm_tuda/nisos.git. The corresponding authors can be contacted for granting access. Exemplary python scripts, based on KadiAPY library, used to automate the data recording and fetching steps are available on Zenodo[43].

Received: 3 July 2024; Accepted: 8 April 2025;
Published online: 09 June 2025

References

1. Gu, D., Shi, X., Poprawe, R., Bourell, D., Setchi, R. & Zhu, J. Material-structure-performance integrated laser-metal additive manufacturing. *Science.* **372**, eabg1487 (2021).
2. Deagen, M., Brinson, L., Vaia, R. & Schadler, L. The materials tetrahedron has a -digital twin-. *MRS Bulletin.* **47**, 379–388 (2022).
3. Frazier, W. Metal Additive Manufacturing: A Review. *J. Mater. Eng. Perform.* **23**, 1917–1928 (2014).
4. Wilkinson, M. *et al.* The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data.* **3**, 1–9 (2016).
5. Garabedian, N. *et al.* Generating FAIR research data in experimental tribology. *Scientific Data.* **9**, 315 (2022).
6. McElfresh, C., Wang, Y. & Marian, J., Fast-throughput simulations of laser-based additive manufacturing in metals to study the influence of processing parameters on mechanical properties. *Heliyon.* 10 (2024).
7. Oliveira, J., Santos, T. & Miranda, R. Revisiting Fundamental Welding Concepts To Improve Additive Manufacturing: From Theory To Practice. *Prog. Mater. Sci.* **107**, 100590 (2020).
8. King, W. *et al.* Laser Powder Bed Fusion Additive Manufacturing Of Metals; Physics, Computational, And Materials Challenges. *Appl. Phys. Rev.* **2**, 041304 (2015).
9. Miranda, L., Towards data-centric machine learning: a short review. Ljvmiranda921.github.io. (2021).
10. La, H. & La, K., Comparing model-centric and data-centric approaches to determine the efficiency of data-centric AI. *Journal Of Emerging Investigators*. https://api.semanticscholar.org/CorpusID:263322523 (2023).
11. Yang, Y. *et al.* High-throughput computational homogenization of porous materials produced by selective laser sintering: a diffuse-interface-based workflow. Computational Mechanics. https://link.springer.com/10.1007/s00466-025-02610-8 (2025).
12. Russell, K., FAIR Principles and Why They Matter. *Digital Transformation Of The Laboratory: A Practical Guide To The Connected Lab*. 101–105 (2021).
13. Chaffey, D. & Wood, S., Business information management: improving performance using information. (Pearson Education Ltd., Upper Saddle River, 2005).
14. Griem, L. *et al.* KadiStudio: FAIR Modelling of Scientific Research Processes. *Data Science Journal.* **21**, 16–16 (2022).
15. Brandt, N. *et al.* Kadi4Mat: A research data infrastructure for materials science. *Data Science Journal.* **20**, 8–8 (2021).
16. Team, K., & Contributors kadi: 0.25.1. (Zenodo, 2022, 6), https://doi.org/10.5281/zenodo.6623521.
17. Dunie, M. The importance of research data management: The value of electronic laboratory notebooks in the management of data integrity and data availability. *Information Services & Use.* **37**, 355–359 (2017).
18. Colabroy, K. & Bell, J., Lab eNotebooks. *Biochemistry Education: From Theory To Practice*. pp. 173-195 (2019).
19. Shabih, S. *et al.* Development of a FAIR Data Management Infrastructure. *Microscopy And Microanalysis.* **28**, 2930–2932 (2022).
20. CARP, N., Minges, A. & Piel, M. eLabFTW: An open source laboratory notebook for research labs. *J. Open Source Softw.* **2**, 146 (2017).
21. Hewera, M., Hänggi, D., Gerlach, B. & Kahlert, U. eLabFTW as an Open Science tool to improve the quality and translation of preclinical research. *F1000Research.* **10**, 292 (2021).
22. Team, K., & Contributors kadi-apy: 0.23.0. (Zenodo, 2022, 6), https://doi.org/10.5281/zenodo.6623518.
23. Tosato, G., Koeppe, A., Xu, B., Selzer, M. & Nestler, B., Bayesian optimization framework for data-driven materials design. *Tage 2023*. pp. 306 (2023).
24. Zhou, X. *et al.* 3D-multilayer simulation of microstructure and mechanical properties of porous materials by selective sintering. *GAMM-Mitteilungen.* **44**, e202100017 (2021).
25. Yang, Y. *et al.* Elasto-plastic residual stress analysis of selective laser sintered porous materials based on 3D-multilayer thermo-structural phase-field simulations. *Npj Computational Materials.* **10**, 117 (2024).
26. Hilden, J., Rief, S. & Planas, B., GeoDict 2023 User Guide. GrainGeo handbook. (Math2Market GmbH, Germany, 2023).
27. Et al, V. Yade Documentation 3rd ed. http://yade-dem.org/doc/ (The Yade Project, 2021).
28. Huber, S. *et al.* AiiDA 1.0, a scalable computational infrastructure for automated reproducible workflows and data provenance. *Scientific Data.* **7**, 300 (2020).
29. Yang, Y., Bondi, P. & Xu, B., Physics-based data-driven analyses on melt pool control in laser powder bed fusion. (Zenodo, 2024, 12), https://doi.org/10.5281/zenodo.14501390.
30. Yang, Y. *et al.* Validated dimensionless scaling law for melt pool width in laser powder bed fusion. *Journal Of Materials Processing Technology.* **299**, 117316 (2022).
31. Yang, Y., Doñate-Buendía, C., Oyedeji, T., Gökce, B. & Xu, B. Nanoparticle tracing during laser powder bed fusion of oxide dispersion strengthened steels. *Materials.* **14**, 3463 (2021).
32. Goßling, M. *et al.* Towards enhancing ODS composites in laser powder bed fusion: Investigating the incorporation of laser-generated zirconia nanoparticles in a model iron-chromium alloy. *Journal Of Materials Research* **39**, 1–15 (2023).
33. Yang, Y., Oyedeji, T., Zhou, X., Albe, K. & Xu, B. Tailoring magnetic hysteresis of additive manufactured Fe-Ni permalloy via multiphysics-multiscale simulations of process-property relationships. *Npj Computational Materials.* **9**, 103 (2023).
34. Yang, Y. *et al.* A diffuse-interface model of anisotropic interface thermal conductivity and its application in thermal homogenization of composites. *Scripta Materialia.* **212**, 114537 (2022).
35. Scheffler, M. *et al.* FAIR data enabling new horizons for materials research. *Nature.* **604**, 635–642 (2022).

36. Kusoglu, I. *et al*. Nanoparticle Additivation Effects on Laser Powder Bed Fusion of Metals and Polymers-A Theoretical Concept for an Inter-Laboratory Study Design All Along the Process Chain, Including Research Data Management. *Materials*. **14**, 4892 (2021).
37. Vastola, G., Zhang, G., Pei, Q. & Zhang, Y. Controlling of residual stress in additive manufacturing of Ti6Al4V by finite element modeling. *Additive Manufacturing*. **12**, 231–239 (2016).
38. Yi, M., Xu, B. & Gutfleisch, O. Computational study on microstructure evolution and magnetic property of laser additively manufactured magnetic materials. *Computational Mechanics*. **64**, 917–935 (2019).
39. Carraturo, M. *et al*. Modeling and experimental validation of an immersed thermo-mechanical part-scale analysis for laser powder bed fusion processes. *Additive Manufacturing*. **36**, 101498 (2020).
40. Yang, Y., Ragnvaldsen, O., Bai, Y., Yi, M. & Xu, B. 3D non-isothermal phase-field simulation of microstructure evolution during selective laser sintering. *Npj Computational Materials*. **5**, 81 (2019).
41. Permann, C. *et al*. MOOSE: Enabling massively parallel multiphysics simulation. *SoftwareX*. **11**, 100430 (2020).
42. Yang, Y. & Bharech, S. Elasto-plastic residual stress analysis of selective laser sintered porous materials based on 3D-multilayer thermo-structural phase-field simulations. *npj Computational Materials* **10**, 117, https://doi.org/10.5281/zenodo.10940626 (2024).
43. Nova, N., Bharech, S., Yang, Y. & Xu, B. Research Data Management framework for simulation based materials science research: A Kadi-APY approach. https://doi.org/10.5281/zenodo.8419354 (Zenodo, 2023, 9).
44. Rowley, J. The wisdom hierarchy: representations of the DIKW hierarchy. *Journal Of Information Science*. **33**, 163–180 (2007).

Acknowledgements

Authors acknowledge the financial support of German Research Foundation (DFG) in the framework of the Collaborative Research Centre Transregio 270 (CRC-TRR 270, project number 405553726, sub-projects A06 and B13), and the Priority Program 2122 (SPP 2122, project number 493889809). The authors also greatly appreciate the access to the Lichtenberg II high-performance computing (HPC) cluster and the technical support from the HHLR, Technische Universität Darmstadt. This work was partly carried out with the support of the Karlsruhe Nano Micro Facility (KNMFi, www.knmf.kit.edu), a Helmholtz Research Infrastructure at Karlsruhe Institute of Technology (KIT, www.kit.edu). The computing time on the HPC cluster is granted by the NHR4CES Resource Allocation Board under the project "special00007". S.B. also thanks the master's student Ms. Nabodita Anam Nova for her assistance in preparing showcase for proposed framework.

Author contributions

Conceptualization: B.-X.X., Y.Y. and S.B.; methodology: S.B. and Y.Y.; software: S.B. and Y.Y.; data curation: S.B. and Y.Y.; visualization: S.B. and Y.Y.; writing-original draft preparation: S.B., Y.Y. and B.-X.X.; writing-review and editing: S.B., Y.Y., M.S., B.-X.X. and B.N.; supervision: B.-X.X.; consultation and discussion: M.S. and B.N.; funding acquisition: B.-X.X. All authors have read, reviewed and agreed to the published version of the manuscript.

Funding

Open Access funding enabled and organized by Projekt DEAL.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to Y.Y. or B.-X.X.