A Reproducible Machine Learning Workflow to Characterize the Solid Electrolyte Interphase

Deepalaxmi Rajagopal[1a,1b], Arnd Koeppe[1a,1b], Meysam Esmaeilpour[1c], Michael Selzer[1a,1b,2], Helge Stein[3], Britta Nestler[1a,1b,2]

[1a]Institute for Applied Materials – Microstructure Modelling and Simulation (IAM-MMS);
[1b]Institute for Nanotechnology (INT) - Mikrostruktursimulation (INT-MSS);
[1c]Institute of Nanotechnology (INT);
[1]Karlsruhe Institute of Technology (KIT);
[2]Institute for Digital Materials Science (IDM), Karlsruhe University of Applied Sciences
[3]Helmholtz Institute Ulm

Research data management tools structure the data life cycle and expedite the scientific process. Applied research data management still needs to be incorporated into daily research operations at the institutional level that allows access to the entire data life cycle. The generated warm data along the research operations enable automatic knowledge base generation and interpretation using robust integrated data analysis methods. The open-source research data infrastructure Kadi4Mat provides a generic framework for FAIR data management, efficient scientific workflows, and integrated data analysis. In this use case from virtual material design, we demonstrate how to implement machine learning as a workflow to characterize the virtual Solid Electrolyte Interphase (SEI) formation in Lithium-ion batteries. A better understanding of SEI formation helps to adjust batteries for optimum performance and safety. The workflow combines data and model definition, preprocessing, training, generation, and data analysis. We utilize kinetic Monte Carlo simulations and deep learning (Variational Auto Encoders with a parallel regressor) to structure the complex, high-dimensional, and non-convex design space and demonstrate how integrated data analysis can characterize materials and predict material properties.

1 Introduction

In recent years, there has been a strong push toward digitalizing the materials science field, with data-driven methods being developed to accelerate the development of new materials. For the generation and development of such high-performing material informatics, many syntheses, experiments, and simulations must be performed; this accelerated development

Publiziert in: Vincent Heuveline, Nina Bisheh und Philipp Kling (Hg.): E-Science-Tage 2023. Empower Your Research – Preserve Your Data. Heidelberg: heiBOOKS, 2023. DOI: https://doi.org/10.11588/heibooks.1288.c18061

of material results in multifaceted datasets. These datasets have to be combined and analyzed to discover new knowledge. The material data produced by experiments other than simulation lacks standard formats and corresponding metadata, which makes it difficult to share, visualize, and analyze these data (Ludwig 2019; Hey and Trefethen 2003; Draxl and Scheffler 2020). Incorporating FAIR data principles into material dataset generation methods helps make the data more accessible and easier for the research community to manage, share, and reuse.

A flexible research data management tool is essential for achieving the requirements of the FAIR (Findable, Accessible, Interoperable, and Reusable) principles by providing a wide variety of tasks that can be performed, such as retrieving data from an experimental device, processing data, sharing processed data more efficiently, automating the data handling process, data mining, and visualization. Utilizing research data management tools that provide access to unprocessed information sources from which published data is derived can result in manageable research process chains. Various research data software, such as Zenodo, Dataverse (Crosas 2011), Dspace (Smith et al. 2003), and Nomad, focus specifically on published data. The other important component of research data platforms is electronic lab notebooks to facilitate the digitization of the data from experiments and simulations while offering access to a wide range of data analysis and visualization tools required for the research.

Various Electronic Lab Notebooks (ELNs), including Jupyter Notebooks (Kluyver et al. 2016), Galaxy (Jalili et al. 2020), Fireworks (Jain et al. 2015), ElabFTW (CARPi, Minges, and Piel 2017), and Aiida (Pizzi et al. 2016) exist for documenting research processes. However, their domain-specific nature often prevents researchers from utilizing them fully. Moreover, most of the mentioned ELNs lack interdisciplinary capabilities and necessitate programmatic expertise for establishing workflows that can automate research operations. Hence, it is necessary to implement a research data management system that caters to the requirements of managing interdisciplinary research processes. This system should provide access to "warm data" which refers to unpublished data yet to be analyzed. Incorporating an ELN into the repository-based Research Data Management (RDM) tools with provisions for accessing user interface-based and script-based research process workflow implementations helps to minimize the effort required for daily research activities like data retrieval from the experimental devices, data sharing, data analysis, and visualization.

To meet the aforementioned requirements, Kadi4Mat (Karlsruhe Data Infrastructure for Materials Science; Team 2022; Brandt et al. 2021), an open-source data platform that functions as a communal repository, and the ELN is being developed at the Karlsruhe Institute of Technology. The Kadi4Mat platform's repository component facilitates effectively organizing data from various sources. It expedites data sharing among fellow researchers or research project collaborators. In contrast, their ELN component enables the logging of meaningful information about the research process, the visualization, and the data analysis stored in the repository component. It uses a user and programmatic interface to construct reproducible research workflows. In addition, the Kadi4Mat ecosystem provides access to KadiAI and CIDS (Computational Intelligence and Data Science tools; Koeppe and CIDS Team 2023) to allow the use of interactive dashboards for machine learning process definition and execution, corresponding workflow nodes to define the process of data-driven study, and programmatic interfaces for data preparation into a machine-readable format, data engineering, data-driven model architecture construction, tuning, and training. Integrating Artificial intelligence (AI) toolset and advanced material simulations with research data infrastructure accelerates the discovery of new material configurations and develops traceable research process chains for further study (Koeppe et al. 2022; Mundt et al. 2020; Koeppe et al. 2018).

This article explores Kadi4Mat ecosystem functionalities and their potential for implementing reproducible integrated data analysis. Specifically, we use a deep generative model to examine the data-driven use case based on characterizing solid electrolyte interphase in batteries.

2 Tools and Methods

2.1 The Kadi4Mat ecosystem

Kadi4Mat represents a comprehensive platform incorporating a community repository to organize and share data from various sources efficiently, enriched by an ecosystem of applications and interfaces that facilitate efficient and automatized RDM. The ELN of the Kadi4Mat ecosystem consists of both web-based and desktop-based workflow editors to streamline scientific workflows. Kadi4Mat aims to digitally document the scientific workflow in daily research, which facilitates researchers in reproducing and utilizing identical data and research workflows more efficiently. The generic architecture of Kadi4Mat enables the replication of nearly all stages involved in the research data lifecycle, except planning and publishing. However, it is possible to effectively execute these two procedures by integrating established frameworks into the Kadi4Mat. For instance, RDMO (Klar et al. 2017) may be utilized for managing research data plans, while Zenodo can serve as a platform for publishing data. In the following sections, some components of the Kadi4Mat ecosystem are discussed to provide an overview of their role in implementing reproducible integrated data analysis.

2.2 Kadi's core components: KadiWeb and Kadistudio

KadiWeb is a web-based interface that provides a user-friendly platform to access the repository and ELN components of Kadi4Mat. The resources stored in the Kadi4Mat can be accessed via the web and programmatic interfaces called KadiAPY (Schoof and Brandt 2020) based on the command line interface and Python library. Upon accessing Kadi4Mat, the web interface provides access to create and structure data with the help of different components available in the interface. The essential features of the Kadi4Mat are as follows:

Records: instances to structure resources, including data and associated metadata. The metadata includes general information such as title, description, identifier, record type, tags, and other additional metadatum denoted by key/value pairs specific to the stored data, which can be customized based on user needs.

Collections: Collect and arrange records of relevance together.

Templates: defines the record's metadata beforehand. It can contain the information required to create a record or the extra metadata specific to an individual record.

Users: displays registered users of logged-in Kadi4Mat instance.

Group: organizes users into workspaces based on their roles or research projects and facilitates transparent and efficient access management.

Kadistudio is a desktop-based workflow editor enabling the seamless implementation of scientific workflows. The software platform provides access to a wide range of pre-installed tools conveniently tailored to create and implement heterogeneous scientific workflows comprising diverse data types and sources (Griem et al. 2022; Zschumme 2021b, 2021a). To add new tools in the workflow environment, user can define their corresponding XML tool settings (Zschumme et al. 2020). The functionalities of KadiAPY (Schoof and Brandt 2020), workflow nodes that mimic user interface commands, and additional tools such as CIDS and KadiAI for data-based analysis are also available in the form of workflow nodes accessible through the Kadistudio integrated workflow editor.

2.3 KadiAI and CIDS

As an interface between RDM and machine learning applications, KadiAI aims to standardize and integrate AI projects, work packages, and workflows into the Kadi ecosystem. The interface streamlines data-driven research through structuring and automatization and offers interactive dashboards to implement data-driven studies and provide feedback to the user. CIDS is a Python-based framework to develop, implement, and standardize learning algorithms in AI workflows. The framework is integrated within the Kadi4Mat ecosystem to aid in preprocessing and converting the data, as well as develop data-driven models that learn from the defined datasets stored within the Kadi4Mat repository. Together, KadiAI and CIDS enable data-integrated AI within the Kadi4Mat ecosystem to enable a quantitative and qualitative analysis of the collected data.

The generic machine learning workflow concept in Kadi4Mat consists of four essential process elements: source and data preparation, model development, and share model for prediction. There are two different ways to implement this machine-learning workflow concept within Kadi4Mat. One method uses KadiAI user interactive dashboards, which require no programming background to establish and perform machine learning process steps. In contrast, the other method utilizes tailored Python scripts within the CIDS framework for optimized automation of node execution. Figure 1 shows the tools required to implement the machine learning process steps using the Kadi4Mat ecosystem.

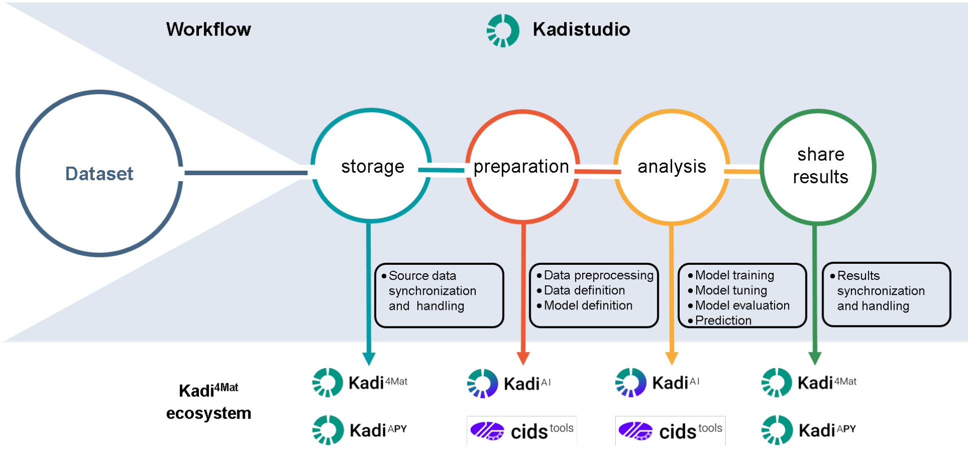

Figure 1: Concept of generic machine learning workflow in Kadi4Mat ecosystem.

3 Use case and workflow

In subsequent sections, we will address the use case research problem based on the characterization of solid electrolyte interphase using a data-driven strategy and its implementation using the Kadi4Mat ecosystem. Specifically, we used a collection of CIDS tools incorporated within Kadistudio to define and execute each processing step of machine learning workflows. Figure 2 shows the machine workflow layout using the Kadi4Mat ecosystem. Some CIDS workflow nodes can communicate with the Kadi4Mat repository to retrieve original data, update transformed data, or upload analysis results. This feature helps with tracking and ensuring the reproducibility of data provenance.

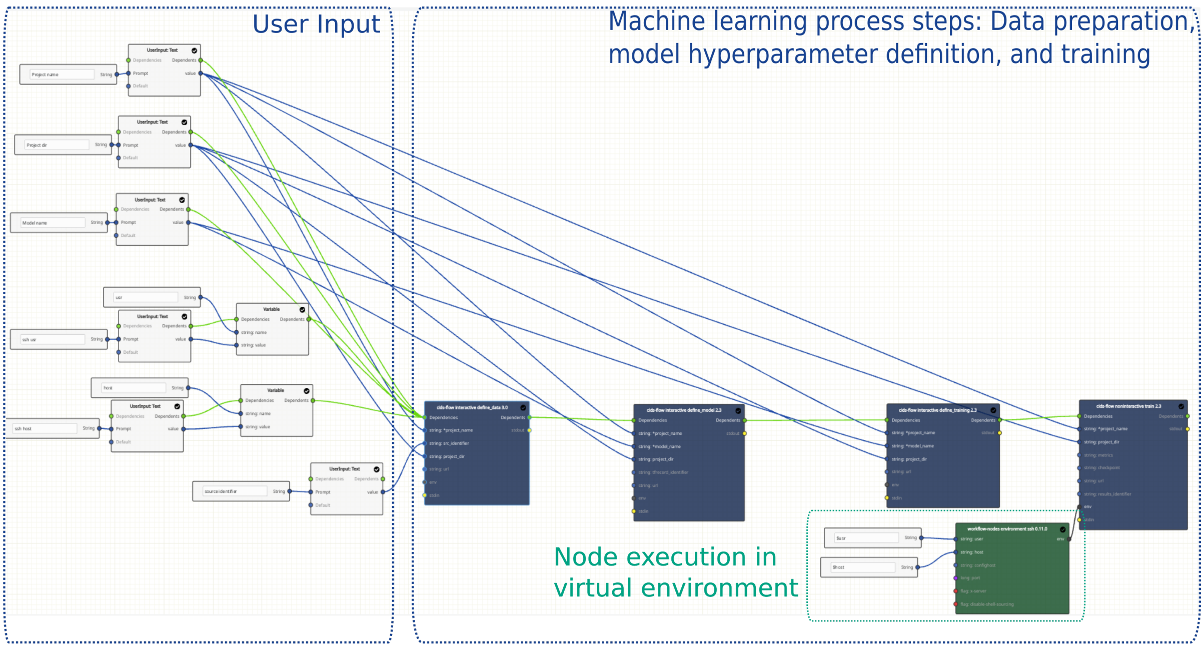

Figure 2: Machine learning workflow layout visualization in Kadistudio.

3.1 Background

The current use case involves a comprehensive investigation of the dataset acquired from kinetic Monte Carlo simulation of solid electrolyte interphase growth in batteries.

The solid electrolyte interphase is formed on the electrode of a battery as a reduction product of the electrolyte. These interphases play a vital role in performance, cycling, and even the safety of the battery (Dunn, Kamath, and Tarascon 2011). The complex interdependent relationship between the composition and morphology of the evolving solid electrolyte interphase (SEI) makes the traditional modeling approaches struggle to capture essential information to predict the behavior of SEI accurately. The recently developed Kinetic Monte Carlo protocol based on the multiscale approach with corresponding reaction kinetics as input can capture the growth of SEI and can determine the essential physical properties of the evolving SEI (Esmaeilpour et al. 2023). However, the above-proposed protocol still takes a long time to complete each simulation and lacks structure-property linkage to interpret and understand the underlying mechanism of SEI growth.

3.2 Objective and approach for characterization using deep generative models

This work aims to accelerate the design of solid electrolyte interphase according to target physical properties and to understand the underlying mechanisms that dominate its growth. We implemented a data-driven approach to expedite the Kinetic Monte-Carlo simulation to characterize the SEI configuration concerning existing physical properties for further optimization. For this approach, we used a variational auto-encoder model and a regressor to understand the underlying representation of higher dimensional SEI configurations and predict the key physical properties of SEI formation. The proposed approach demonstrates the ability to accurately predict SEI properties and structure, which can be used to optimize battery performance and safety.

3.3 Source data and preprocessing

The study utilizes input data derived from Kinetic Monte Carlo simulations that model SEI growth. The obtained SEI dataset contains 50000 samples of the final configuration of SEI growth. Each sample consists of spatial features and non-spatial features. The spatial features of an SEI configuration represent the electrolyte reduction species distinguished by colors according to their reaction product type. The non-spatial features of the SEI configuration refer to physical properties such as thickness, density, volume fraction, and porosity.

The foremost step in promoting data-oriented research using machine learning models is gathering information from experimental observations or simulations. The repository component of Kadi4Mat facilitates efficient sourcing and collection of datasets from the collaborators. Once collected, it is essential to preprocess and clean the data to eliminate any inconsistencies or errors that may compromise the accuracy of the analysis. For instance, we received SEI configuration data generated with Kinetic Monte Carlo simulations from a collaborator via Kadi4Mat (ibid.). To proceed with analysis using this information, we need to define attributes of the obtained dataset such as feature name, data format, data type, data shape, and decoding used during data conversion. Additionally, it is necessary to preprocess the source data to achieve an efficient learning process using machine learning algorithms. The preprocessed data of each sample are then stored in the form of TF records (TensorFlow records) for efficient data serialization and storage. We utilized the CIDS framework-based interactive data definition node to streamline this process to facilitate user input requirements and functionalities for defining the data attributes. Figure 3 visualizes the interactive data definition process using CIDS interactive nodes and KadiAI dashboards.

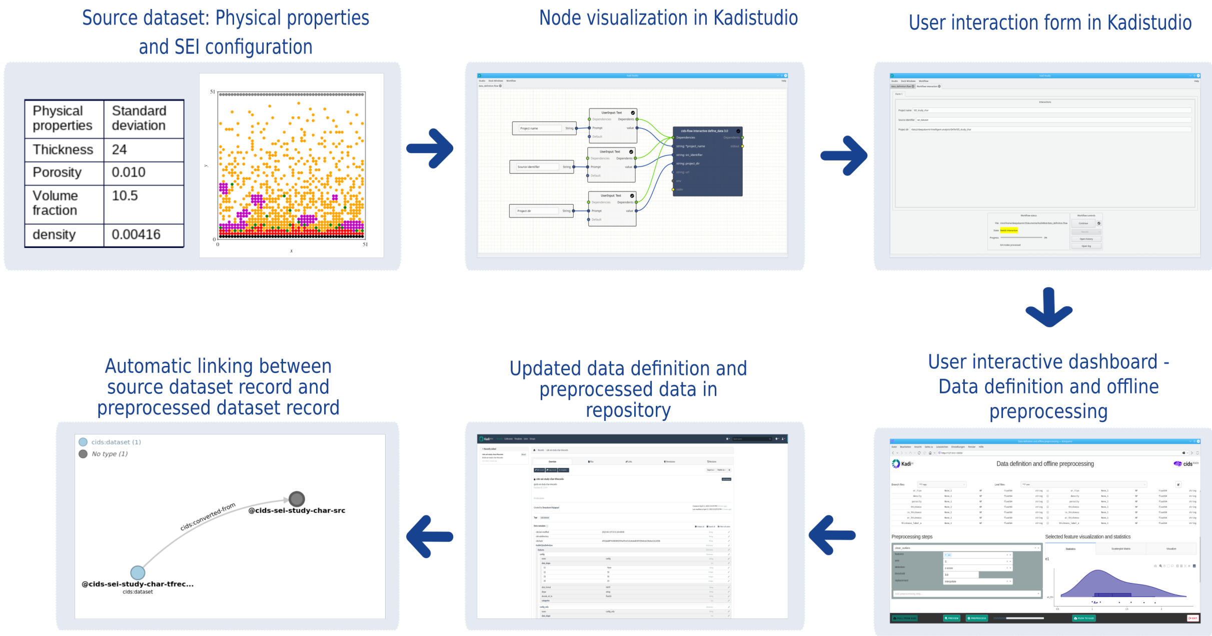

Figure 3: Data preparation process for machine learning study using CIDS interactive nodes and KadiAI dashboards.

3.4 Deep generative models

Variational autoencoder (VAE; Kingma and Welling 2013) is a deep generative model consisting of two main parts: encoder and decoder. The function of the encoder is to compress the higher dimensional input into a distribution over the lower dimensional latent space. The lower dimensional latent space defines the bottleneck of the VAE model. Thus, the encoder produces new feature space from the old feature space through extraction or selection of essential features. On the other hand, the decoder part takes a point in the lower dimensional latent space and decompresses it back to the original higher dimensional input space. An additional part called regressor at the bottleneck of VAE aids in incorporating the information of physical properties into the learned latent space. Training the regressor component involves predicting the desired physical properties acquired latent space representation and backpropagation to automatically arrange VAE's bottleneck based on physical attributes (Gómez-Bombarelli et al. 2018). The acquired knowledge of the representations assists in predicting desired physical properties. Here, we used the CIDS interactive model definition node to aid in defining the architecture of a model. Figure 4 visualizes the model definition and hyperparameter selection using the CIDS interactive model definition node. This node aims to enable efficient input feature and output feature definition, facilitate the selection of appropriate model architectures, and specify ranges for corresponding hyperparameters. The hyperparameters defined here determine the aspects of model architecture, such as ranges for the numbers of convolution layers within the encoder and decoder, ranges and choices of latent dimensionality, and choice of activation function for the layers.

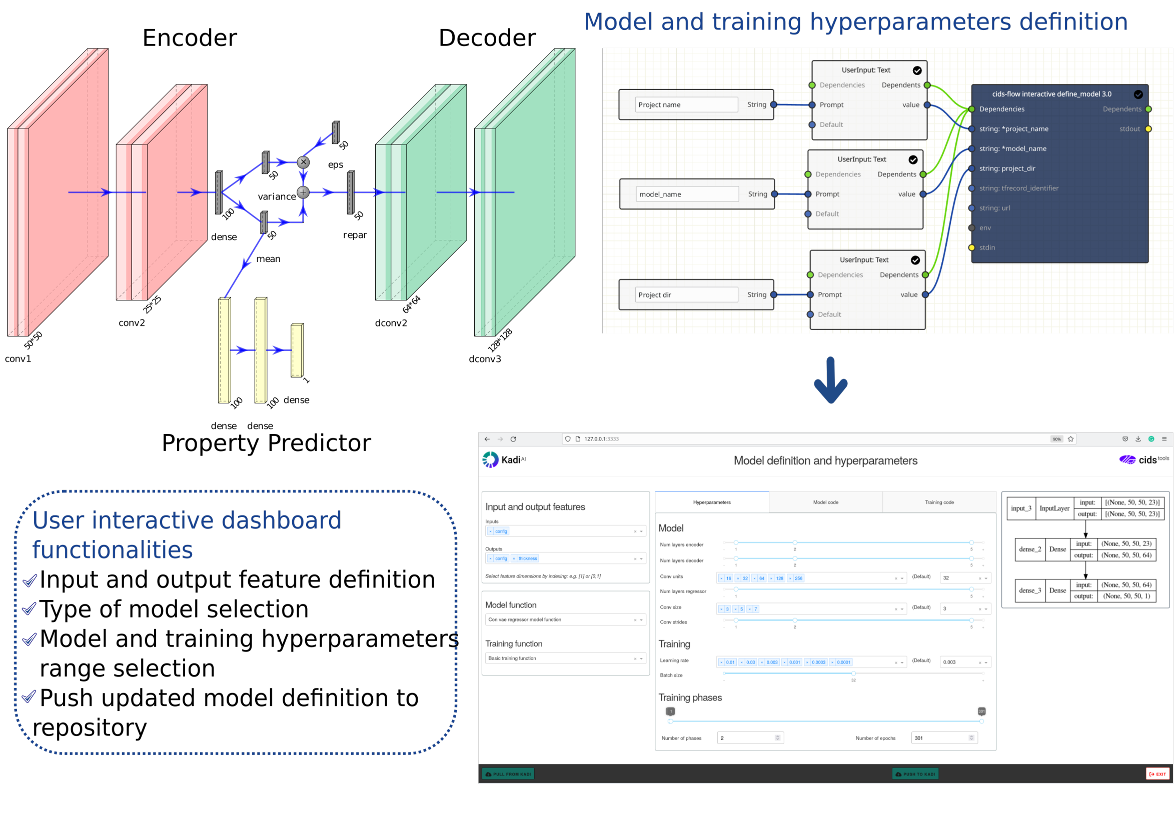

Figure 4: Model and training hyperparameter selection using CIDS interactive nodes and KadiAI dashboards.

In a machine learning workflow, the next step is to define the hyperparameters of the training, such as batch size, number of epochs, learning rate, optimizer choice, and number of training phases. These settings should be carefully selected and fine-tuned to ensure optimal model performance during training. A user-interactive training definition node assists in the definition of these training hyperparameters and their corresponding search range. On defining the required hyperparameters for training, the next step is to start the model training with the help of a non-interactive training node which automates the training process by utilizing the hyperparameters and model architecture defined in the previous steps.

3.5 Data provenance tracking in executed machine learning workflows

The KadiAI workflow nodes developed to define and execute machine learning processes using the CIDS framework record all the essential artifacts obtained throughout the workflow execution. Each CIDS workflow node integrated with the KadiAI interface has its machine-learning metadata library. The metadata library captures and retrieves metadata containing essential information about the various workflow process steps, their executions, and the artifacts produced during the machine learning process. Logging metadata includes everything from setting up KadiAI projects to model training and evaluation. KadiAI nodes generate the corresponding record type in the Kadi4Mat repository for each machine learning process result and metadata storage.

The KadiAI projects are locally executed at the user level, which is analogous to classic research processes. Through a structured upload, KadiAI projects synchronize the results with the repository, where each processing step is saved as a record with corresponding metadata and data. As the admin of the uploaded project, the user can grant access to other members with specific roles such as member, editor, admin, or collaborator. KadiAI projects also record the tool versions used for each processing step, allowing users to go forward or backward with tool versions or infrastructures as required. For each new model within a KadiAI, the ML model-creating function as Python code and tunable model settings are recorded in separate records, which are linked to the corresponding project record. Execute in Kadi workflows, users can skip process nodes and reuse existing derived datasets for new models if no reprocessing is required, ultimately streamlining the workflow. The created records of the KadiAI project are linked using unidirectional record links to define the context of the relationship between them. The knowledge graphs of the executed machine learning workflow aid in traceability and provenance tracking of the data used, models developed, and results obtained during the machine learning project.

Figure 5 visualizes the data provenance for the characterization of solid electrolyte interphase using the deep generative model in KadiAI as knowledge graphs. Knowledge graphs turn data collected along your machine-learning process into machine-understandable knowledge. These graphs help you to define the context of the relationship between the two methods and adapt situational changes. On the other hand, it allows you to incorporate real-world knowledge into your study, which data-driven lacks.

3.6 Data and code availability

The code for Kadi4Mat is readily available on the public Gitlab repository (Kadi4Mat Team and Contributors 2023) for seamless sharing and collaboration within the community. Additionally, we publish tool versions on Zenodo with a Digital Object Identifier

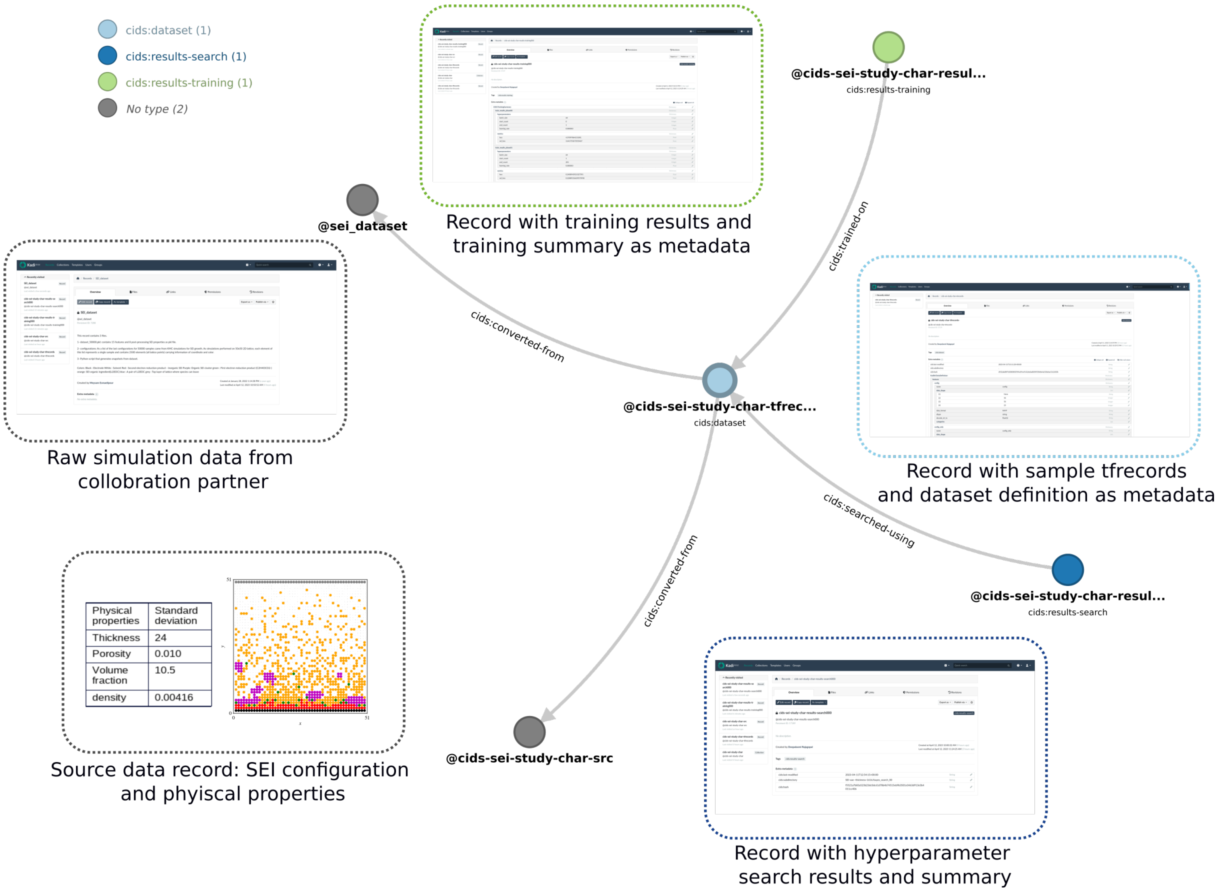

Figure 5: Visualization of tracked data provenance during machine learning process steps as knowledge graphs.

(DOI; Kadi4Mat Team and Contributors 2023) for enduring accessibility while prioritizing backward compatibility of both tools and infrastructure through the versioning tools on Zenodo. Implementing the FAIR principles made our code easily discoverable and accessible through our public Gitlab repository. Furthermore, we ensure that the code is interoperable through the commonly used Python language, and our Zenodo DOI persistence guarantees the reproducibility of tool versions.

4 Conclusion

In this article, we discussed the importance of research data management in daily scientific research activities and the role of existing research data infrastructure in tackling the challenges related to the organization, storage, and sharing of research data. Then the infrastructure and functionalities of the Kadi4Mat ecosystem, a generic research data management system to handle heterogeneous data in interdisciplinary material science, and the development of virtual research environments through scientific workflows are described. To understand the functionalities of the Kadi4Mat ecosystem, a research problem focused on the data-driven study to characterize and predict the properties of solid electrolyte configuration of interest is addressed. The demonstrated machine learning workflow using the CIDS framework and KadiAI user interactive dashboards interfaced with the Kadi4Mat repository showed how data management like Kadi4Mat can be integrated with machine learning processes to ensure reproducibility, transparency, and traceability of artifacts generated during the execution of the workflow.

Acknowledgements

The authors gratefully acknowledge the financial support for the research provided by the Deutsche Forschungsgemeinschaft's "Cluster of Excellence" POLiS (project number 390874152), the BMBF's "FestBatt" competence cluster (project number 03XP0174E), the Ministry of Science, Research, and Art Baden-Württemberg (MWK-BW) in the project MoMaF-Science Data Center, with funds from the state digitization strategy digital@bw (project number 57). We would also like to thank the German Federal Ministry of Education and Research (BMBF) for its financial support of the AQuaBP project under grant number 03XP0315B. The authors acknowledge support by the Helmholtz association through the program MTET, no: 38.02.01.

References

Brandt, Nico, Lars Griem, Christoph Herrmann, Ephraim Schoof, Giovanna Tosato, Yinghan Zhao, Philipp Zschumme, and Michael Selzer. 2021. *Kadi4Mat: A Research Data Infrastructure for Materials Science.* 20:8. 1. Ubiquity Press. DOI: `https://doi.org/10.5334/dsj-2021-008`.

CARPi, Nicolas, Alexander Minges, and Matthieu Piel. 2017. "eLabFTW: An open source laboratory notebook for research labs". *The Journal of Open Source Software* 2 (12): 146. ISSN: 2475-9066. DOI: `https://doi.org/10.21105/joss.00146`.

Crosas, Mercè. 2011. "The Dataverse Network: An Open-source Application for Sharing, Discovering and Preserving Data". *D-Lib Magazine* 17 (1/2). DOI: `https://doi.org/10.1045/january2011-crosas`.

Draxl, Claudia, and Matthias Scheffler. 2020. "Big data-driven materials science and its FAIR data infrastructure". In *Handbook of Materials Modeling,* 49–73. Springer International Publishing. DOI: `https://doi.org/10.1007/978-3-319-44677-6_104`.

Dunn, Bruce, Haresh Kamath, and Jean-Marie Tarascon. 2011. "Electrical energy storage for the grid: a battery of choices". *Science* 334 (6058): 928–935. DOI: `https://doi.org/10.1126/science.1212741`.

Esmaeilpour, Meysam, Saibal Jana, Hongjiao Li, Mohammad Soleymanibrojeni, and Wolfgang Wenzel. 2023. "A Solution-Mediated Pathway for the Growth of the Solid Electrolyte Interphase in Lithium-Ion Batteries". *Advanced Energy Materials* 13 (14): 2203966. DOI: `https://doi.org/10.1002/aenm.202203966`.

Gómez-Bombarelli, Rafael, Jennifer N. Wei, David Duvenaud, José Miguel Hernández-Lobato, Benjamín Sánchez-Lengeling, Dennis Sheberla, Jorge Aguilera-Iparraguirre, Timothy D. Hirzel, Ryan P. Adams, and Alán Aspuru-Guzik. 2018. "Automatic Chemical Design Using a Data-Driven Continuous Representation of Molecules". *ACS Central Science* 4 (2): 268–276. DOI: https://doi.org/10.1021/acscentsci.7b00572.

Griem, Lars, Philipp Zschumme, Matthieu Laqua, Nico Brandt, Ephraim Schoof, Patrick Altschuh, and Michael Selzer. 2022. "KadiStudio: FAIR Modelling of Scientific Research Processes". *Data Science Journal* 21 (1): 16. ISSN: 1683-1470. DOI: https://doi.org/10.5334/dsj-2022-016.

Hey, Anthony, and Anne Trefethen. 2003. *The Data Deluge: An e-Science Perspective.* DOI: https://doi.org/10.1002/0470867167.CH36.

Jain, Anubhav, Shyue Ping Ong, Wei Chen, Bharat Medasani, Xiaohui Qu, Michael Kocher, Miriam Brafman, Guido Petretto, Gian-Marco Rignanese, Geoffroy Hautier, et al. 2015. "FireWorks: a dynamic workflow system designed for high-throughput applications". *Concurrency and Computation: Practice and Experience* 27 (17): 5037–5059. DOI: https://doi.org/10.1002/cpe.3505.

Jalili, Vahid, Enis Afgan, Qiang Gu, Dave Clements, Daniel Blankenberg, Jeremy Goecks, James Taylor, and Anton Nekrutenko. 2020. "The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2020 update". *Nucleic Acids Research* 48 (W1): W395–W402. DOI: https://doi.org/10.1093/nar/gkaa434.

Kadi4Mat Team and Contributors. 2023. *Kadi-Karlsruhe Data Infrastructure for Materials Science.* https://gitlab.com/iam-cms/kadi.

Kadi4Mat Team and Contributors. 2023. *kadi: 0.39.3.* Version 0.39.3. DOI: https://doi.org/10.5281/zenodo.8233992.

Kingma, Diederik P, and Max Welling. 2013. "Auto-encoding variational bayes". *arXiv preprint arXiv:1312.6114.*

Klar, Jochen, Claudia Engelhardt, Heike Neuroth, Harry Enke, and Jens Ludwig. 2017. "RDMO-Research Data Management Organiser". In *EGU General Assembly Conference Abstracts,* 15760.

Kluyver, Thomas, Benjamin Ragan-Kelley, Fernando Pérez, Brian E Granger, Matthias Bussonnier, Jonathan Frederic, Kyle Kelley, Jessica B Hamrick, Jason Grout, Sylvain Corlay, et al. 2016. "Jupyter Notebooks – a publishing format for reproducible computational workflows." In *Positioning and Power in Academic Publishing: Players, Agents and Agendas,* 87–90. IOS Press. DOI: https://doi.org/10.3233/978-1-61499-649-1-87.

Koeppe, Arnd, Franz Bamer, Michael Selzer, Britta Nestler, and Bernd Markert. 2022. "Explainable artificial intelligence for mechanics: physics-explaining neural networks for constitutive models". *Frontiers in Materials* 8:636. DOI: https://doi.org/10.48550/arXiv.2104.10683.

Koeppe, Arnd, and CIDS Team. 2023. "CIDS and KadiAI GitLab Repository". Visited on September 5, 2023. https://gitlab.com/intelligent-analysis/cids.

Koeppe, Arnd, Carlos Alberto Hernandez Padilla, Maximilian Voshage, Johannes Henrich Schleifenbaum, and Bernd Markert. 2018. "Efficient numerical modeling of 3D-printed lattice-cell structures using neural networks". *Manufacturing Letters* 15:147–150. DOI: https://doi.org/10.1016/j.mfglet.2018.01.002.

Ludwig, Alfred. 2019. "Discovery of new materials using combinatorial synthesis and high-throughput characterization of thin-film materials libraries combined with computational methods". *NPJ Computational Materials* 5 (1): 70. DOI: https://doi.org/10.1038/s41524-019-0205-0.

Mundt, Marion, Arnd Koeppe, Sina David, Tom Witter, Franz Bamer, Wolfgang Potthast, and Bernd Markert. 2020. "Estimation of Gait Mechanics Based on Simulated and Measured IMU Data Using an Artificial Neural Network". *Frontiers in Bioengineering and Biotechnology* 8 (41): 1–16. DOI: https://doi.org/10.3389%2Ffbioe.2020.00041.

Pizzi, Giovanni, Andrea Cepellotti, Riccardo Sabatini, Nicola Marzari, and Boris Kozinsky. 2016. "AiiDA: automated interactive infrastructure and database for computational science". *Computational Materials Science* 111:218–230. DOI: https://doi.org/10.1016/j.commatsci.2015.09.013.

Schoof, Ephraim, and Nico Brandt. 2020. *IAM-CMS/kadi-apy: Kadi4Mat API Library.* Version 0.2.1. DOI: https://doi.org/10.5281/zenodo.4088276.

Smith, Mackenzie, Mary Barton, Mick Bass, Margret Branschofsky, Greg McClellan, Dave Stuve, Robert Tansley, and Julie Walker. 2003. "DSpace: An Open Source Dynamic Digital Repository". *D-Lib Magazine* 9. DOI: https://doi.org/10.1045/january2003-smith.

Team, Kadi4Mat. 2022. *kadistudio: 0.1.0.alpha1.* DOI: https://doi.org/10.5281/zenodo.6810891.

Zschumme, Philipp. 2021a. *IAM-CMS/process-engine.* Version 0.1.0. DOI: https://doi.org/10.5281/zenodo.4442563.

———. 2021b. *IAM-CMS/process-manager.* Version 0.1.0. DOI: https://doi.org/10.5281/zenodo.4442553.

Zschumme, Philipp, Patrick Altschuh, Nico Brandt, Lars Griem, and Ephraim Schoof. 2020. *IAM-CMS/workflow-nodes.* Version 0.1.0. DOI: https://doi.org/10.5281/zenodo.4094719.