

<https://doi.org/10.62405/osiconf.2026.46>

Denysiuk Nataliia,

Frantsevich Institute for Problems of Materials Science
of the NAS of Ukraine, Kyiv, Ukraine,
ORCID: 0000-0002-0492-1945, e-mail: n.denysiuk@ipms.kyiv.ua

Parfionov Oleksandr,

Frantsevich Institute for Problems of Materials Science
of the NAS of Ukraine, Kyiv, Ukraine,
ORCID: 0009-0009-4123-8051, e-mail: o.parfionov@ipms.kyiv.ua

Shevchenko Andrii,

State Research Institution «Kyiv Academic University», Kyiv, Ukraine,
ORCID: 0000-0001-7229-5395, e-mail: a.shevchenko@kau.edu.ua

Shadura Vitalii,

State Research Institution «Kyiv Academic University», Kyiv, Ukraine,
ORCID: 0000-0001-6294-5606, e-mail: v.shadura@kau.edu.ua

Smolyak Svitlana,

State Research Institution «Kyiv Academic University», Kyiv, Ukraine,
ORCID: 0000-0002-5403-4092, e-mail: sssmolyak@gmail.com

Denkov Ivan,

V.M. Glushkov Institute of Cybernetics of National Academy
of Sciences of Ukraine, Kyiv, Ukraine,
ORCID: 0009-0007-7989-2169, e-mail: havoc85@gmail.com

A MODEL FOR MODULAR FAIR DESCRIPTION OF COMPUTATIONAL WORKFLOWS IN MATERIALS SCIENCE

ABSTRACT. *The aim of this study is to develop and demonstrate the applicability of a model for modular FAIR description of multi-stage computational research in materials science. The methodology combines analysis of existing metadata and provenance approaches, structural workflow decomposition, information modelling, and a case-based demonstration. The proposed model separates dataset- and material-level metadata, DFT-, MLIP- and MD-specific metadata, workflow, provenance, and manifest information. Practical applicability is demonstrated on a SiC DFT → MLIP → MD chain. The demonstration showed the value of explicitly representing the DFT-to-MLIP and MLIP-to-MD transitions as distinct components of the model that capture digital-object transformations, provenance, and transfer conditions between computational stages. The planned workflow is thereby separated from the actual execution history. Further work includes schema formalisation, mapping to RO-Crate/W3C PROV, and broader testing.*

KEYWORDS: *FAIR principles; computational materials science; metadata; provenance; scientific workflows; reproducibility.*

INTRODUCTION. The creation of thematic EOSC nodes at the NAS of Ukraine as a development of the Open Science infrastructure at the NAS of Ukraine was discussed in (Zagorodny et al., 2026). One of the main provisions of Open Science is the implementation of the FAIR principles, which are among the fundamental guidelines of modern research data management. They call for structured metadata, persistent identifiers, and semantic relationships needed for the machine-based discovery, interpretation, and reuse of digital objects (Wilkinson et al., 2016). At the same time, FAIR does not define a universal metadata set, so domain communities develop their own profiles suited to the specifics of their methods and data types.

For computational research, open access to final data or code alone does not guarantee reproducibility. It is also necessary to document the software environment, versions of programs, libraries and packages, execution parameters, intermediate digital objects, and the sequence of their transformations (Leipzig et al., 2021). In computational materials science this problem is compounded by the multi-stage nature of the calculations. In particular, modelling the thermomechanical properties of 3C- and 6H-SiC employed a DFT → MLIP → MD sequence, in which DFT results form the reference dataset for training an interatomic potential subsequently applied in molecular dynamics (Parfionov & Vasiliev, 2026). Such workflows combine different data formats, software codes, and computational environments (Ghiringhelli et al., 2023).

The application of FAIR to computational workflows is actively developing in the context of their interoperability, reproducibility, and reuse (Wilkinson et al., 2025). AiiDA supports automated execution and provenance capture (Huber et al., 2020), NOMAD provides a normalised description of materials-science data (Scheidgen et al., 2023), and RO-Crate aggregates data, software, workflow, and execution context into machine-readable research objects (Soiland-Reyes et al., 2022). Workflow Run RO-Crate further combines the description of the planned process with the provenance of its actual execution (Leo et al., 2024). At the same time, for heterogeneous DFT → MLIP → MD workflows, a domain-specific description of the transitions between stages and their associated metadata remains an open issue. Provenance, in turn, ensures the traceability of the digital objects used and generated and of the computational actions performed (Leo et al., 2024).

The aim of this work is to develop and demonstrate the applicability of a model for the modular FAIR description of multi-stage computational workflows, using the DFT → MLIP → MD chain for SiC as an example. The model was developed within the MATSCI-NASU pilot project with the support of EOSC Gravity. Its purpose is to provide a domain-specific description and packaging of a multi-stage computational process, with the possibility of further alignment with RO-Crate, W3C PROV, NOMAD, and systems for automated provenance capture.

METHODS. The methodology of the work combines a comparative analysis of existing approaches, conceptual information modelling, structural decomposition of the workflow, and demonstration of the model on a specific scenario, following the logic of design science for information artefacts (Peppers et al., 2007). First, the information components required to describe the dataset, the material system, stage-specific parameters, software/environment, workflow, provenance, and the physical composition of the package were identified.

The computational process was then decomposed into interrelated stages and digital entities (Fig. 1). For each stage, the input data, computational action, software, parameters, and output data were defined. Transitions between computational stages, which describe the preparation, transformation, and transfer of digital objects to the next stage of the workflow, were identified as separate components. For the demonstration scenario these components are DFT-to-MLIP, which covers the collection and screening of DFT results, the formation of the dataset.extxyz set, and its split into training, validation, and test subsets, and MLIP-to-MD, which describes the transformation of the trained model into a form suitable for use at the molecular-dynamics stage. Data provenance was described through entities, activities, and agents with explicit used/generated/derived relationships, conceptually aligned with W3C PROV. The planned process structure and the actual execution history were separated into a prospective workflow and a retrospective provenance record.

The practical applicability of the model was verified on an assembled package of SiC simulation results, organised into five functionally related parts: `dft/`, `dft_to_mlip/`, `mlip/`, `mlip_to_md/`, and `md/`. For each stage, configuration files, a description of the software environment, input and output data, scripts, and results were preserved; in particular, training/validation/test data are additionally provided for MLIP, and simulation parameters, execution logs, results, and trajectories are provided for MD. The consistency of

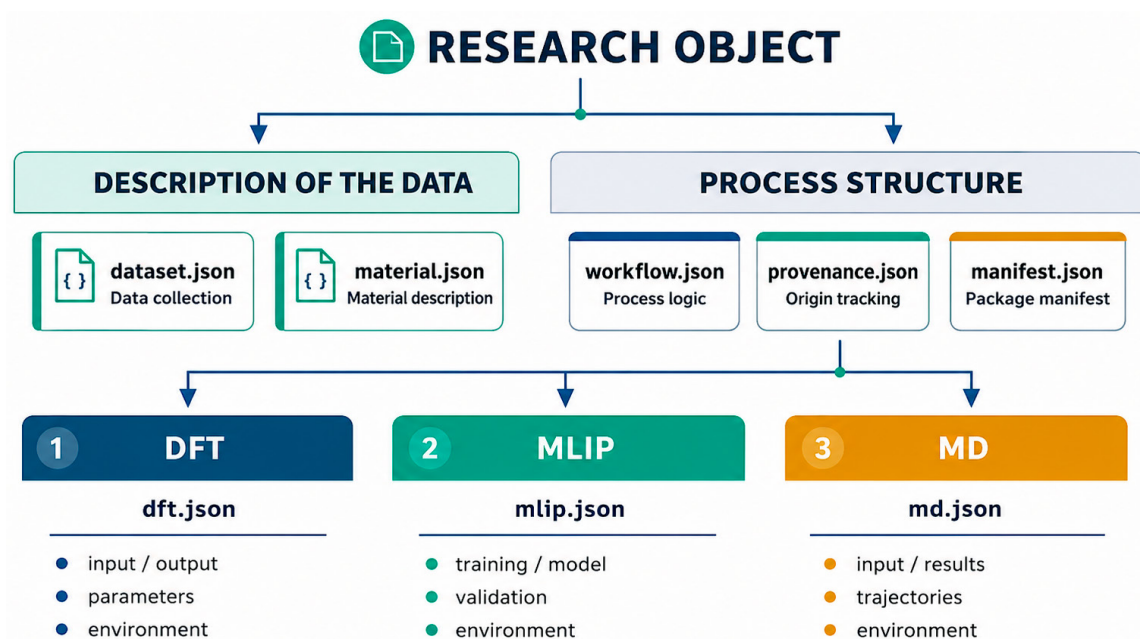


Fig. 1. Modular architecture of a FAIR-oriented research object for a multi-stage workflow in computational materials science

the model was assessed not only in terms of directory structure, but also in terms of the correspondence between the actual files, stage-specific metadata, workflow, provenance, and manifest.

RESULTS AND DISCUSSION. The result of the structural analysis is a proposed modular architecture of a research digital object (the DFT → MLIP → MD workflow), in which different levels of research information have distinct functions. Machine-readable files were produced describing: dataset.json — the dataset as a whole; material.json — the material system; dft.json, mlip.json, and md.json — the metadata of the respective computational stages; workflow.json — the logic of the workflow; provenance.json — the actual history of the origin of the digital objects; and manifest.json — the composition of the package and control of its integrity. This separation minimises duplication and allows the model to be extended with new stages without restructuring the entire architecture.

The practical implementation of the proposed model on the SiC compound (Fig. 2) refined the initial DFT → MLIP → MD workflow scenario. Instead of a simple three-block sequence, five functional packages were identified: DFT → DFT-to-MLIP → MLIP → MLIP-to-MD → MD.

At the DFT stage, Quantum ESPRESSO generated reference data for the 3C-SiC and 6H-SiC polytypes using a single set of parameters (PBE-PAW pseudopotentials, $ecutwfc = 100$ Ry, $ecutrho = 800$ Ry). The dataset contains 1604 configurations and 66,248 atomic environments, with each configuration storing geometry, energy, forces, and the stress tensor. A deterministic split based on structure hashing produced 1349 training, 167 validation, and 88 test configurations.

At the MLIP stage, the NequIP framework was used to train a machine-learning interatomic potential, with the following set of key hyperparameters: $r_{\text{max}} = 4.5$ Å, three interaction layers, $l_{\text{max}} = 2$, 64 features. On an independent test set, the following MAE values were obtained: 3.617 meV/atom for energy, 25.99 meV/Å for forces, and 0.2823 GPa for stresses. This study prioritizes the structural role of the metrics within the package over their specific values: the validation results exist as a separate machine-readable object linked to the test set, the checkpoint, the training configuration, and the deployable artefact. This closes a traceable chain from the DFT dataset to the validated model and confirms that it was actually executed, rather than merely declared.

Research package structure for a DFT → MLIP → MD workflow

Example package aligned with the MATSCI-NASU modular structure

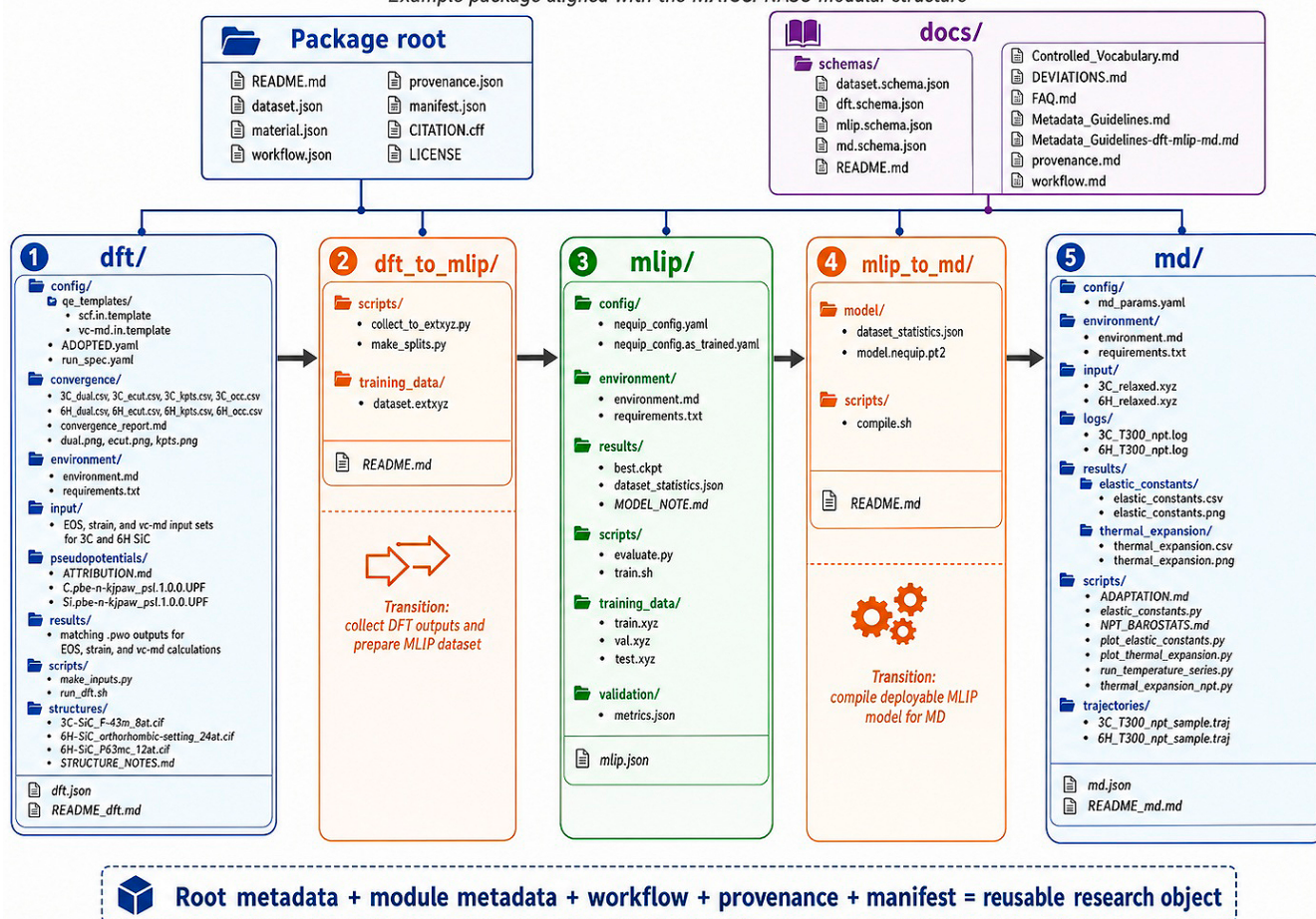


Fig. 2. Structure of the research package for the multi-stage DFT → MLIP → MD workflow, highlighting the DFT-to-MLIP and MLIP-to-MD transition components

The MD component contains the input structures for 3C- and 6H-SiC, simulation parameters, a description of the software environment, NPT-run logs, sample trajectories, and the results of elastic-property and thermal-expansion calculations. Thus, the package structure covers not only the preparation of the MLIP model for use in MD, but also the digital artefacts of the executed molecular-dynamics stage.

One of the significant innovations of the proposed modular FAIR-description model is the finding that a transition between computational stages cannot be reduced to a simple file transfer. It is precisely at the boundary between stages that format transformations, data selection and filtering, formation of dataset subsets, model compilation, and changes of software environment occur. Accordingly, two separate transition packages were introduced into the structure: `dft_to_mlip` and `mlip_to_md`. The first package captures the transformation of raw DFT outputs into a training dataset. A key decision was to store geometry, energy, forces, and stresses in a single extxyz representation at the level of an individual configuration. Splitting these labels across parallel directories would create an implicit dependence on file or row order; after filtering or partial regeneration this could lead to an undetected mismatch between the structure and its corresponding labels. Thus, the practical implementation led to a deviation from a formally possible file structure in favour of preserving the semantic integrity of the research object.

The second transition, `mlip_to_md`, revealed a different category of problems. The trained checkpoint and the model ready for use in MD turned out to be different digital objects. The AOT-compiled NequIP

artefact depends on the PyTorch/CUDA versions and the hardware environment, while recompiling the checkpoint in the configuration used requires access to the training dataset via the path recorded during training. Consequently, simply storing the model weights does not ensure full transferability. Reuse therefore also requires describing the compilation activity, the execution environment, dataset statistics, and the interdependence between the checkpoint and the training data.

This result shows that, for the FAIR description of computational workflows, it is appropriate to model data handover as part of workflow provenance. In a multi-stage workflow, reproducibility can be lost precisely between two correctly documented computational stages if the transformation of the object between them remains undescribed (Fig. 3).

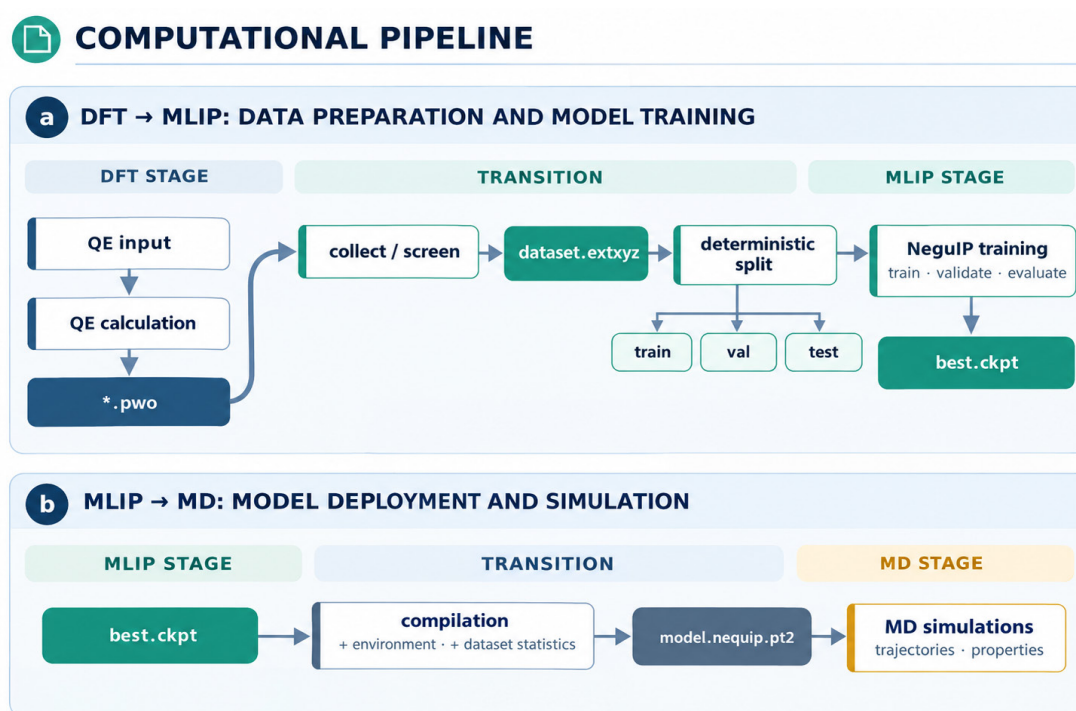


Fig. 3. Transfer of digital objects between computational stages and the role of transition packages

The practical implementation of the modular FAIR-description model confirmed the value of separating two informational representations: workflow, as the plan of the study, and provenance, as the record of the computational actions actually performed. This separation corresponds to the distinction between prospective workflow description and retrospective execution provenance used in contemporary models of computational workflows (Leo et al., 2024).

In workflow.json, the process is represented at two levels. The abstract level contains five functional stages: DFT generation, dataset preparation, MLIP training, model preparation, and MD simulation. The concrete level details these into a sequence of operations including generation of DFT inputs, execution of DFT, dataset formation, its split into the corresponding subdatasets, MLIP training and evaluation, model compilation, and MD calculations for both SiC polytypes.

In contrast, provenance.json describes the actual execution history through related activities, entities, and agents. During validation, it was compared with the digital objects actually present in the package, including DFT results, the resulting dataset, training/validation/test subsets, MLIP training and model-compilation outputs, and digital artefacts of the MD stage. This comparison confirmed consistency between the performed computational stages, the generated digital objects, and their provenance representation.

The separation of prospective workflow and retrospective provenance is important for the practical use of the model: the former defines the structure and dependencies of the computational process, whereas the latter documents its actual execution. Their joint analysis makes it possible to verify the correspondence between planned stages, computational activities actually performed, and generated digital results, thereby supporting the traceability of the multi-stage study (Fig. 4).

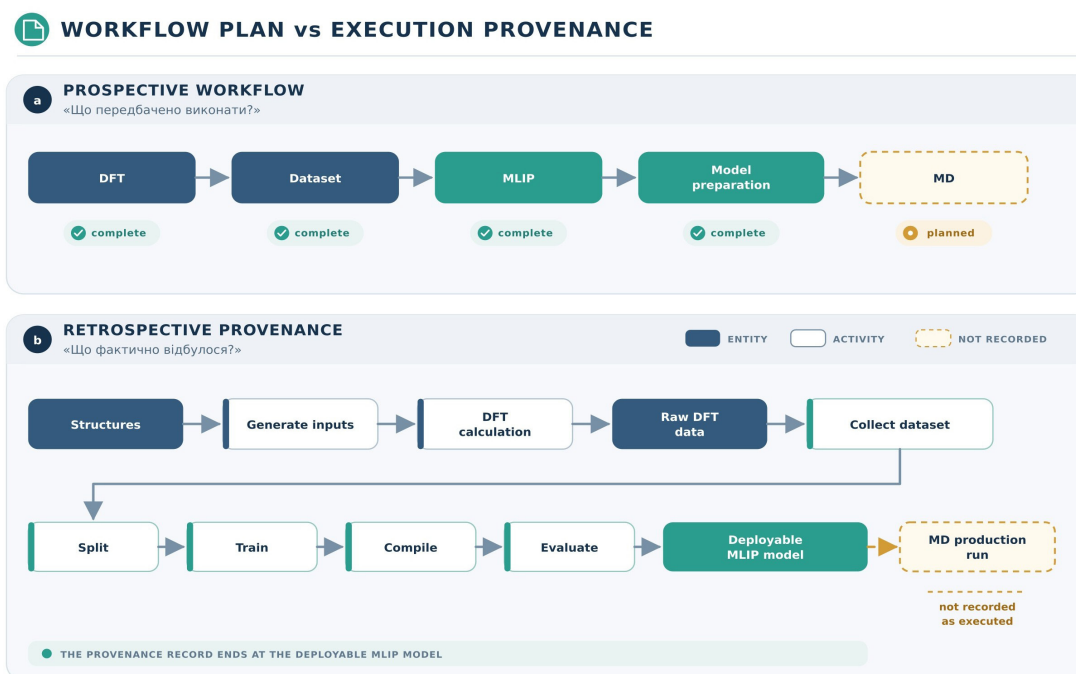


Fig. 4. Separation of prospective workflow and retrospective provenance in the proposed model

This approach is conceptually compatible with the development of workflow-oriented Research Objects and Workflow Run RO-Crate, in which process execution is considered together with its inputs, outputs, software, and workflow definition (Soiland-Reyes et al., 2022; Leo et al., 2024). At the same time, the proposed model is independent of any specific workflow management system, which is particularly important in computational materials science, where individual stages may be performed using Quantum ESPRESSO, Python scripts, machine-learning frameworks, and molecular-dynamics environments that do not belong to a single orchestration system. Unlike approaches in which provenance is generated automatically within a particular infrastructure, such as AiiDA (Huber et al., 2020), the proposed scheme is intended for packaging a heterogeneous, already established computational research process. These approaches should therefore be regarded not as mutually exclusive, but as complementary. In the proposed model, 'workflow.json' defines the sequence and dependencies between the stages of the computational process, 'provenance.json' records the activities actually performed and the origin of the resulting objects, while the transition components provide traceability of digital-object transformations between successive stages.

Findability is supported by structured metadata describing the workflow and its individual stages, as well as by the identification of digital objects and the relationships between them. Accessibility is ensured through human- and machine-readable descriptions of the workflow, computational parameters, software environment, and provenance, together with stable repository-level access. Interoperability is achieved through the use of JSON/JSON Schema, controlled stage types, explicit inter-module relationships, and the Entity–Activity–Agent model, with the possibility of further mapping to RO-Crate and W3C PROV. Reusability is supported by the traceability of the inputs → activities → outputs chain, the recording of

software parameters and versions, and the explicit description of the DFT-to-MLIP and MLIP-to-MD inter-stage transformations.

CONCLUSIONS. This work proposes a model for the modular FAIR description of multi-stage computational workflows in materials science. Rather than introducing a new data format, the proposed approach differentiates and interlinks different levels of research information: general dataset- and material-level metadata, the metadata of individual computational stages, workflow, provenance, and manifest. This approach enables the representation of the results of a computational study as a coherent research object while preserving the specific character of its individual stages.

Demonstration on the DFT → MLIP → MD scenario for SiC showed that, for reproducibility, not only the individual computational stages matter, but also the transitions between them. Identifying DFT-to-MLIP and MLIP-to-MD as independent components facilitates the formalisation of digital objects transformations, the conditions of their transfer, and the relationships between the results of successive methods. The separation of prospective workflow and retrospective provenance, in turn, ensures the representation of the logic of the planned process and of the actual history of its execution.

The demonstration also outlined the limits of the current version of the model. In particular, inter-stage transformations do not yet have a separate Transition type in the controlled vocabulary, and the JSON Schema formalisation for material, workflow, provenance, and manifest still needs to be completed. Further development also envisages mapping to RO-Crate and W3C PROV, as well as testing the model on branched, cyclic, and active-learning workflows. The proposed approach is not intended as an alternative to NOMAD, AiiDA, or other platforms, but as a domain-specific level of description that can be aligned with them.

The results obtained suggest that the object of FAIR description in computational materials science can be not only the final dataset, but also a system of interrelated digital objects, computational actions, and inter-stage transformations formed over the course of a study. At the same time, the conclusions of this work are based on a linear SiC scenario, so further testing on other materials, software tools, and more complex workflows is necessary to assess the transferability of the model.

FUNDING. The work was supported by the MATSCI-NASU project (EOSC Gravity project) and the project “Updating and expanding the use of the capabilities of the Center of Competence for Research Data Management” implemented within the framework of the Targeted Scientific and Technical Project of the NAS of Ukraine “Implementation and Support of Open Science in Institutions of the NAS of Ukraine (OPENS2)”.

REFERENCES

- Ghiringhelli, L. M., Baldauf, C., Bereau, T., Brockhauser, S., Carbogno, C., Chamanara, J., Cozzini, S., Curtarolo, S., Draxl, C., Dwaraknath, S., Fekete, Á., Kermode, J., Koch, C. T., Kühbach, M., Ladines, A. N., Lambrix, P., Himmer, M.-O., Levchenko, S. V., Oliveira, M., ... Scheffler, M. (2023). Shared metadata for data-centric materials science. *Scientific Data*, 10, 626. <https://doi.org/10.1038/s41597-023-02501-8>
- Huber, S. P., Zoupanos, S., Uhrin, M., Talirz, L., Kahle, L., Häuselmann, R., Gresch, D., Müller, T., Yakutovich, A. V., Andersen, C. W., Ramirez, F. F., Adorf, C. S., Gargiulo, F., Kumbhar, S., Passaro, E., Johnston, C., Merkys, A., Cepellotti, A., Mounet, N., ... Pizzi, G. (2020). AiiDA 1.0, a scalable computational infrastructure for automated reproducible workflows and data provenance. *Scientific Data*, 7, 300. <https://doi.org/10.1038/s41597-020-00638-4>
- Leipzig, J., Nüst, D., Hoyt, C. T., Ram, K., & Greenberg, J. (2021). The role of metadata in reproducible computational research. *Patterns*, 2(9), 100322. <https://doi.org/10.1016/j.patter.2021.100322>
- Leo, S., Crusoe, M. R., Rodríguez-Navas, L., Sirvent, R., Kanitz, A., De Geest, P., Wittner, R., Pireddu, L., Garijo, D., Fernández, J. M., Colonnelli, I., Gallo, M., Ohta, T., Suetake, H., Capella-Gutierrez, S.,

- de Wit, R., Kinoshita, B. P., & Soiland-Reyes, S. (2024). Recording provenance of workflow runs with RO-Crate. *PLOS ONE*, 19(9), e0309210. <https://doi.org/10.1371/journal.pone.0309210>
- Parfionov, O., & Vasiliev, O. (2026). Temperature dependent thermo-mechanical properties of 3C and 6H silicon carbide from atomistic simulations on large-scale systems with DFT-derived machine learning interatomic potential. *Functional Materials*, 33(1), 98–105. <https://doi.org/10.15407/fm33.01.98>
- Peffer, K., Tuunanen, T., Rothenberger, M. A., & Chatterjee, S. (2007). A design science research methodology for information systems research. *Journal of Management Information Systems*, 24(3), 45–77. <https://doi.org/10.2753/MIS0742-1222240302>
- Scheidgen, M., Himanen, L., Ladines, A. N., Sikter, D., Nakhaee, M., Fekete, Á., Chang, T., Golparvar, A., Márquez, J. A., Brockhauser, S., Brückner, S., Ghiringhelli, L. M., Dietrich, F., Lehmborg, D., Denell, T., Albino, A., Näsström, H., Shabih, S., Dobener, F., ... Draxl, C. (2023). NOMAD: A distributed web-based platform for managing materials science research data. *Journal of Open Source Software*, 8(90), 5388. <https://doi.org/10.21105/joss.05388>
- Soiland-Reyes, S., Sefton, P., Crosas, M., Castro, L. J., Coppens, F., Fernández, J. M., Garijo, D., Grüning, B., La Rosa, M., Leo, S., Ó Carragáin, E., Portier, M., Trisovic, A., RO-Crate Community, Groth, P., & Goble, C. (2022). Packaging research artefacts with RO-Crate. *Data Science*, 5(2), 97–138. <https://doi.org/10.3233/DS-210053>
- Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., Bonino da Silva Santos, L. O., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., ... Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, 3, 160018. <https://doi.org/10.1038/sdata.2016.18>
- Wilkinson, S. R., Aloqalaa, M., Belhajjame, K., Crusoe, M. R., de Paula Kinoshita, B., Gadelha, L., Garijo, D., Gustafsson, O. J. R., Juty, N., Kanwal, S., Khan, F. Z., Köster, J., Peters-von Gehlen, K., Pouchard, L., Rannow, R. K., Soiland-Reyes, S., Soranzo, N., Sufi, S., Sun, Z., ... Goble, C. (2025). Applying the FAIR Principles to computational workflows. *Scientific Data*, 12, 328. <https://doi.org/10.1038/s41597-025-04451-9>
- Zagorodny, A. G., Vasiliev, O. O., Milenin, O. S., Svistunov, S. Ya., Shadura, V. M., Shevchenko, A. Yu. (2026). Open Science at the NAS of Ukraine: the path from infrastructure to participation in the EOSC Federation. *Visn. Nac. Akad. Nauk Ukr.*, 9, 3-18. <https://doi.org/10.15407/visn2026.09.003>

Денисюк Н. М.,

Інститут проблем матеріалознавства ім. І. М. Францевича
НАН України, Київ, Україна,
ORCID: 0000-0002-0492-1945, e-mail: n.denysiuk@ipms.kyiv.ua

Парфьонов О. В.,

Інститут проблем матеріалознавства ім. І. М. Францевича
НАН України, Київ, Україна,
ORCID: 0009-0009-4123-8051, e-mail: o.parfionov@ipms.kyiv.ua

Шевченко А. Ю.,

Державна наукова установа «Київський академічний університет»,
Київ, Україна,
ORCID: 0000-0001-7229-5395, e-mail: a.shevchenko@kau.edu.ua

Шадур В. М.,

Державна наукова установа «Київський академічний університет»,
Київ, Україна,
ORCID: 0000-0001-6294-5606, e-mail: v.shadura@kau.edu.ua

Смоляк С. С.,

Державна наукова установа «Київський академічний університет»,

Київ, Україна,

ORCID: 0000-0002-5403-4092, e-mail: sssmolyak@gmail.com

Денков І. Д.,

Інститут кібернетики імені В. М. Глушкова НАН України, Київ, Україна,

ORCID: 0009-0007-7989-2169, e-mail: havoc85@gmail.com

МОДЕЛЬ МОДУЛЬНОГО FAIR-ОПISY ОБЧИСЛЮВАЛЬНИХ РОБОЧИХ ПРОЦЕСІВ У МАТЕРІАЛОЗНАВСТВІ

АНОТАЦІЯ. Мета роботи — розробити та продемонструвати застосовність моделі модульного FAIR-опису багатостадійних обчислювальних досліджень у матеріалознавстві. Методологія поєднує аналіз існуючих підходів до метаданих і provenance, структурну декомпозицію workflow, інформаційне моделювання та апробацію на окремому дослідницькому сценарії. Запропонована модель розділяє загальні метадані набору даних і матеріалу, метадані DFT-, MLIP- та MD-етапів, workflow, provenance і manifest. Практичну застосовність продемонстровано на SiC-ланцюжку $DFT \rightarrow MLIP \rightarrow MD$. Апробація показала доцільність явного представлення переходів DFT-to-MLIP і MLIP-to-MD як окремих компонентів моделі, що фіксують перетворення цифрових об'єктів, їх походження та умови передавання між обчислювальними етапами. Запланований workflow при цьому відокремлено від фактичної історії виконання. Визначено напрями подальшої формалізації схем, mapping до RO-Crate/W3C PROV та розширення апробації.

КЛЮЧОВІ СЛОВА: принципи FAIR; обчислювальне матеріалознавство; метадані; provenance; наукові робочі процеси; відтворюваність.