XAIR research data and epistemic metadata for molecular methods

Martin Thomas Horsch,[1,2] Björn Schembera,[3] and Simon Stephan[4]

[1]Institutt for datavitenskap, Norges miljø- og biovitenskapelige universitet, Ås, Norway

[2]STFC Daresbury Laboratory, UK Research and Innovation, Daresbury, UK

[3]Institut für Angewandte Analysis und Numerische Simulation, Universität Stuttgart, Germany

[4]Lehrstuhl für Thermodynamik, Rheinland-Pfälzische Technische Universität, Kaiserslautern, Germany

Epistemic metadata and reproducibility claims

Research data infrastructures promise to support good practice in dealing with research data, making the research outcomes *findable, accessible, interoperable, and reusable* (**FAIR**) and *explainable AI ready* (**XAIR**). These goals make it necessary to document the knowledge status of data by providing **epistemic metadata** [1], *cf.* **Fig. 1**. If the required annotation is missing, data become dark [2, 3]. This occurs for a substantial amount of data in scientific computing, turning **dark data** into a challenge for computational engineering at large [4]. Making data FAIR and XAIR will support researchers at reproducing others' work, corroborating or refuting their findings and communicating the outcome, *cf.* **Fig. 2**. This will make the *"hard road to reproducibility"* [5] less hard, particularly for simulation methods and tools that are seen as epistemically opaque [6] or where validation has been said to require a holistic approach, defying decomposition into individual steps [7].

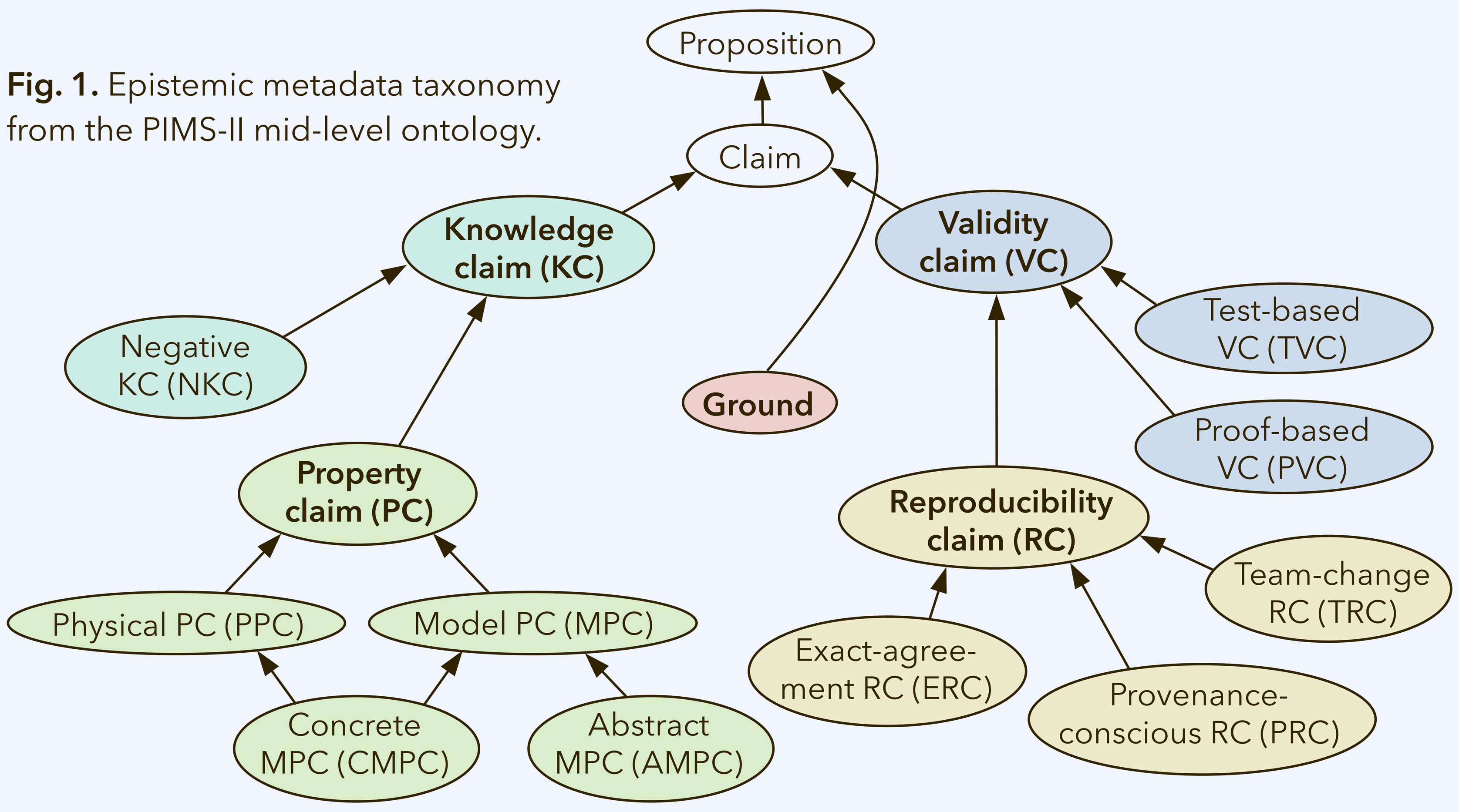

Fig. 1. Epistemic metadata taxonomy from the PIMS-II mid-level ontology.

Knowledge claims (KCs) and **reproducibility claims** (RCs) need to be included among the epistemic metadata, *cf.* **Fig. 1**. Knowledge bases containing need to rely on formal semantics, as illustrated for RCs in **Fig. 2** and for KCs in **Fig. 3**. To this end, the present work employs the PIMS-II mid-level ontology [1, 8], which is aligned with the EMMO [9] and Metadata4Ing [10] from NFDI4Ing. Common ways of expressing reproducibility or replicability, *cf.* Plesser [11], can be understood as instantiating $\Box(\varphi'' \mid \kappa'')$ as a pattern, *cf.* **Fig. 2**; therein, φ'' and κ'' are **orthodata** concerning the knowledge claims and the data provenance.

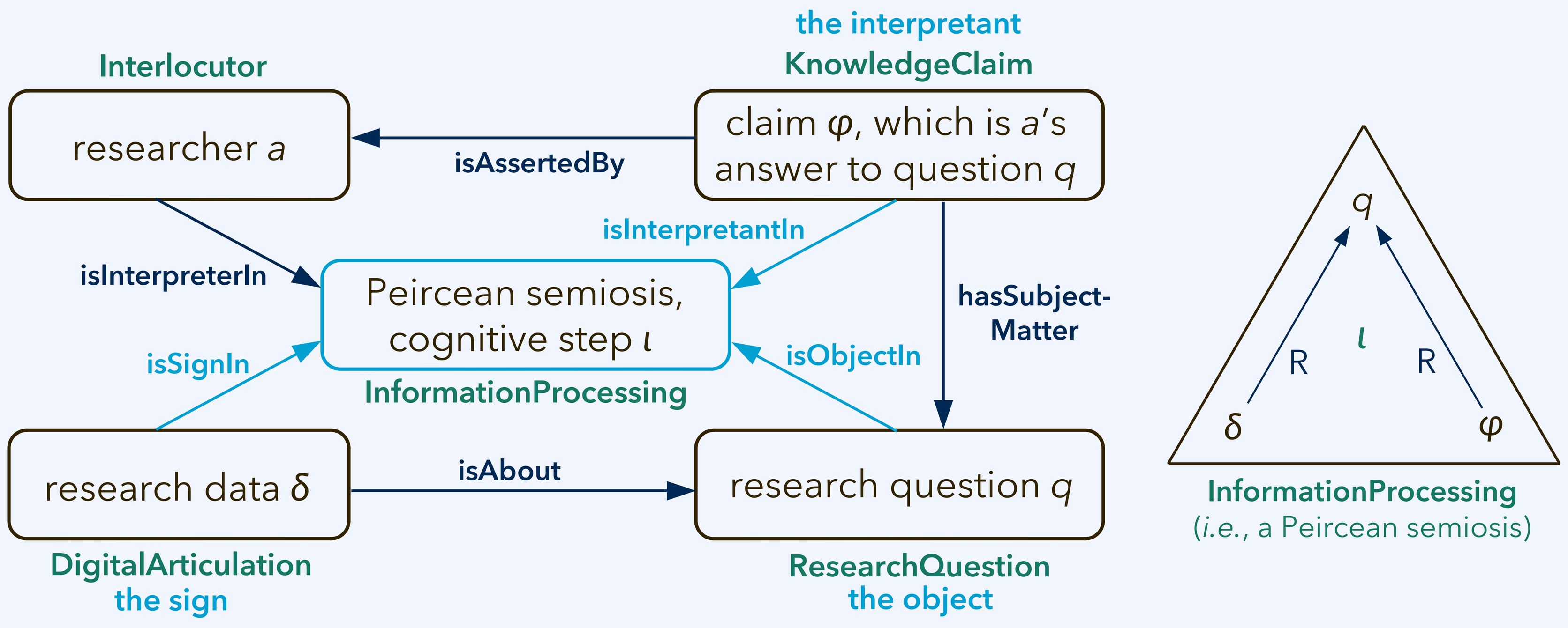

Fig. 3. Knowledge claim schema (*i.e.*, graph shape constraint) using the PIMS-II ontology.

Molecular thermodynamics case study

Reproducibility can become complicated for molecular simulation [12, 13]. The character of their scientific foundation has made numerical methods in statistical mechanics prone to being called *epistemically opaque* [6]. The expectation for outcomes to be reproducible is rooted in disciplinary conventions which are usually unwritten. We are working jointly with partners toward extending the **MolMod DB** model repository [14] to a molecular modelling interoperability infrastructure that complies with European recommendations for *data spaces* and *FAIR digital objects*. As a prerequisite for this, we have conducted a **case study on knowledge claims** in molecular modelling. Therein, researchers engaged in a disciplinary dialogue on knowledge claims, discussing requirements for documenting epistemic metadata [15, 16].

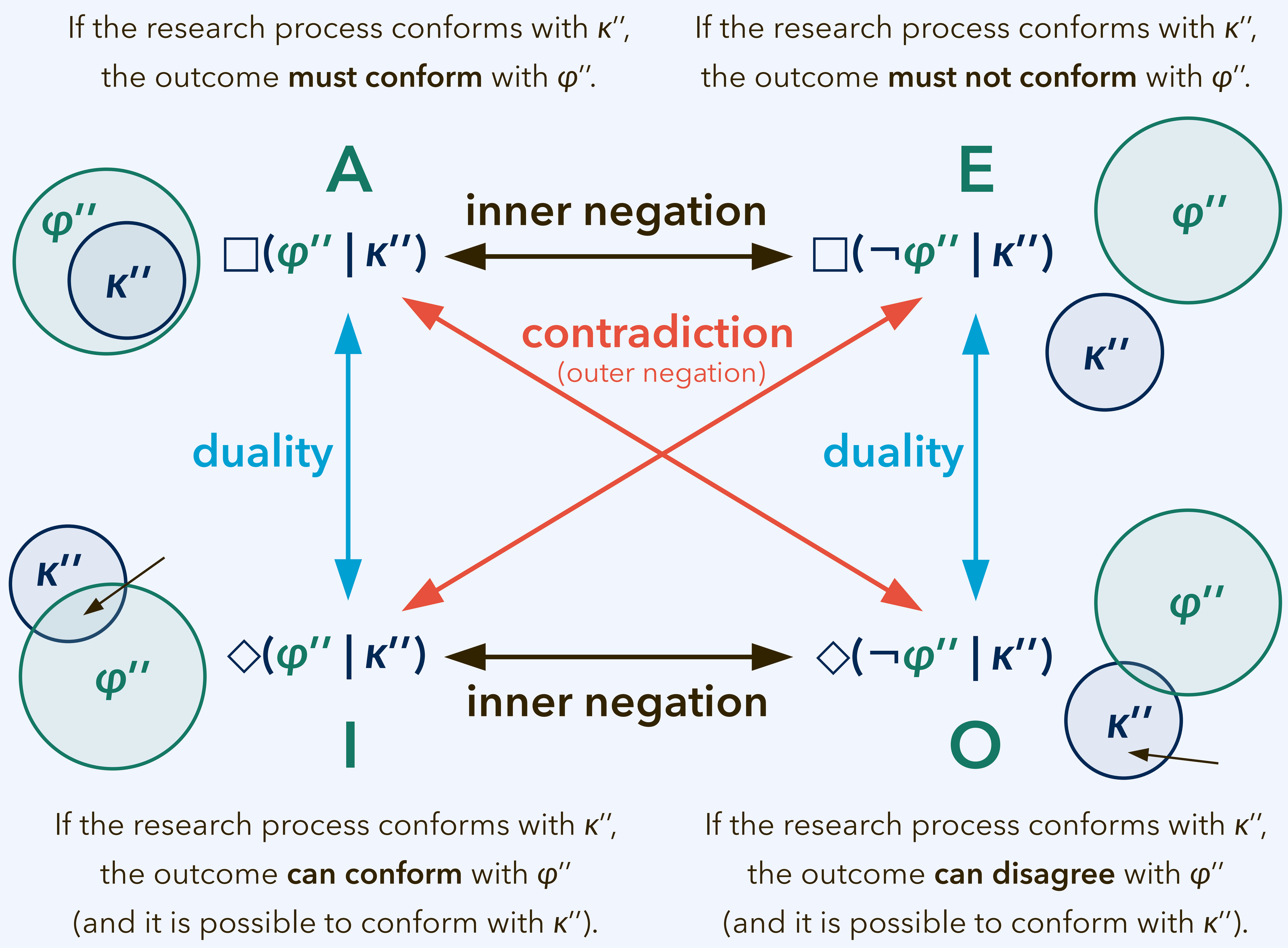

Fig. 2. Square of opposition for conditional necessity and possibility operators, applied to RCs.

Literature

[1] M. T. Horsch, B. Schembera, "Documentation of epistemic metadata by a mid-level ontology of cognitive processes," *p.* 2 in *Proc. JOWO 2022 (CAOS)*, CEUR-WS, **2022**.

[2] B. Schembera, "Like a rainbow in the dark: Metadata annotation for HPC applications in the age of dark data," *J. Supercomput.* **77**: 8946–8966, **2021**.

[3] A. Corallo, A. M. Crespino, V. Del Vecchio, M. Lazoi, M. Marra, "Understanding and defining dark data for the manufacturing industry," *IEEE Transact. Eng. Manag.* **70**(2): 700–712, **2022**.

[4] B. Schembera, J. M. Durán, "Dark data as the new challenge for big data science and the introduction of the scientific data officer," *Philos. Technol.* **33**: 93–115, **2019**.

[5] L. A. Barba, "The hard road to reproducibility," *Science* **354**(6308): 142, **2016**.

[6] J. M. Durán, N. Formanek, "Grounds for trust: Essential epistemic opacity and computational reliabilism," *Minds Machin.* **28**: 645–666, **2018**.

[7] J. Lenhard, "Holism, or the erosion of modularity: A methodological challenge for validation," *Philos. Sci.* **85**: 832–844, **2018**.

[8] M. T. Horsch, "Mereosemiotics: Parts and signs," *p.* 3 in *Proc. JOWO 2021 (FOUST)*, CEUR-WS, **2021**.

[9] H. A. Preisig, T. F. Hagelien, J. Friis, P. Klein, N. Konchakova, "Ontologies in computational engineering," *p.* 262 in *Proc. WCCM-ECCOMAS 2020*, Scipedia, **2021**.

[10] S. Arndt, B. Farnbacher, M. Fuhrmans, S. Hachinger, J. Hickmann, N. Hoppe, *et al.*, "Metadata4Ing: An ontology for describing the generation of research data within a scientific activity," techn. rep., NFDI4Ing, **2022**.

[11] H. E. Plesser, "Reproducibility vs. replicability: A brief history of a confused terminology," *Frontiers Neuroinform.* **11**: 76, **2018**.

[12] M. Schappals, A. Mecklenfeld, L. Kröger, V. Botan, A. Köster, S. Stephan, *et al.*, "Round robin study: Molecular simulation of thermodynamic properties from models with internal degrees of freedom," *J. Chem. Theor. Comput.* **13**(9): 4270–4280, **2017**.

[13] J. Lenhard, U. Küster, "Reproducibility and the concept of numerical solution," *Minds Machin.* **29**: 19–36, **2019**.

[14] S. Stephan, M. Horsch, J. Vrabec, H. Hasse, "MolMod: An open access database of force fields for molecular simulations of fluids," *Mol. Sim.* **45**(10): 806–814, **2019**.

[15] M. T. Horsch, B. Schembera, "Epistemic metadata in molecular modelling: First-stage case-study report (10 cases)," **2023**.

[16] M. T. Horsch, S. Chiacchiera, G. Guevara Carrión, M. Kohns, E. A. Müller, D. Šarić, *et al.*, "Epistemic metadata in molecular modelling: Second-stage case-study report (12 claims)," **2023**.