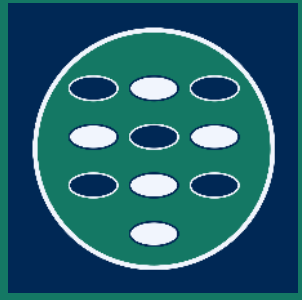
Norges miljø- og biovitenskapelige universitet

Materialteori og -informatikk



Digitalisering på Ås

Universität Stuttgart

Science and Technology Facilities Council

Case study on epistemic metadata in molecular modelling

Martin Horsch,[1,2] Silvia Chiacchiera,[2] and Björn Schembera[3]

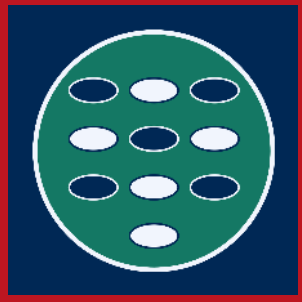
[1]Norwegian Univ. Life Sciences, Faculty of Science and Technology, Ås, Norway
[2]UKRI STFC Daresbury Laboratory, Scientific Computing Department, Daresbury, UK
[3]Univ. Stuttgart, Institute of Applied Analysis and Numerical Simulation, Stuttgart, Germany

Fakultet for realfag og teknologi

Forskergruppe materialteori og -informatikk

Noregs miljø- og biovitskaplege universitet

Materialteori og -informatikk



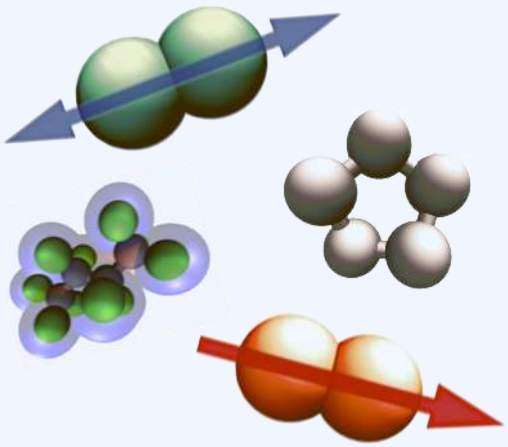
Digitalisering på Ås

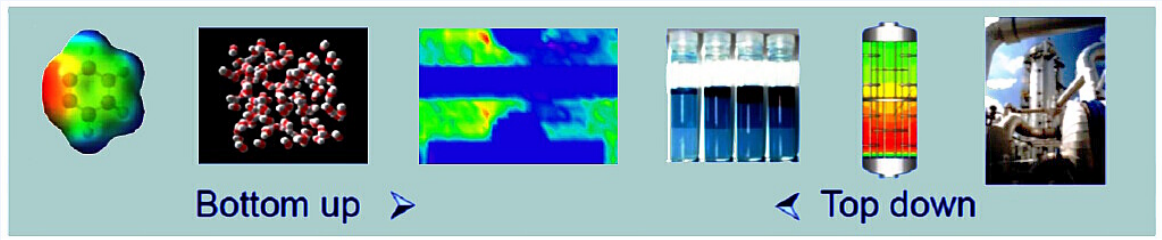
Domain and background

First stage and an ontology

Second stage and grounding

Molecular simulation in engineering





physics-driven aspects (qualitative validity) ⟷ **data-driven aspects (quantitative reliability)**

- Realistic representation of underlying physical features
- Models with parameters that can be adjusted to data
- Reliable interpolation, extrapolation, prediction

RPTU Rheinland-Pfälzische Technische Universität

Helmut Schmidt Universität Universität der Bundeswehr Hamburg

Computational Molecular Engineering

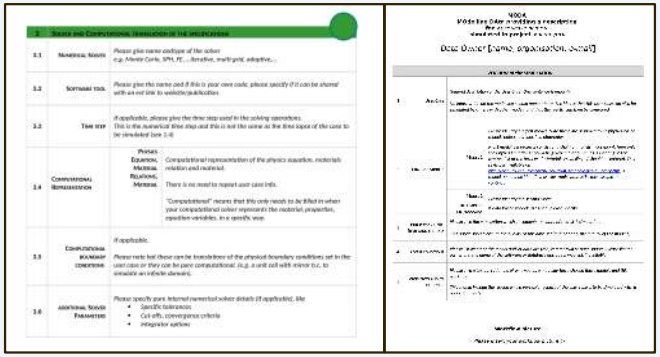
MolMod Database

- Process Engineering
- Thermodynamics
- Scientific Computing

The EMMC's first attempts

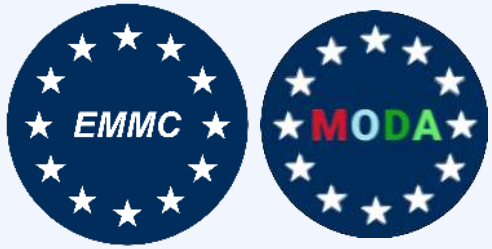
EMMC: European Materials Modelling Council

The EMMC's documentation standard MODA ("model data") **failed to meet researchers' needs** by making too much annotation mandatory, much more than needed in practice.



- MODA was a closed semantic and epistemic space: Modelling methods had to be chosen from a small list.[1, 2]
- MODA imposed a **given level of detail** in workflow documentation.[1]
- MODA documentations were **complicated**.[3]

CWA 17284: 2018 E «MODA»



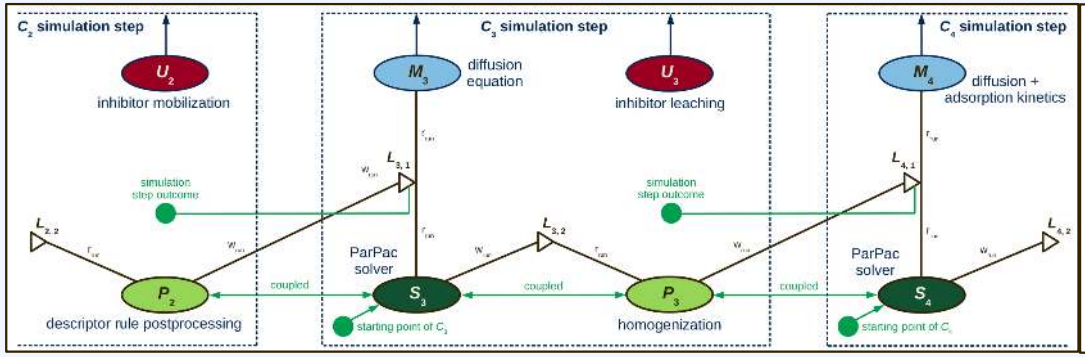
[1]CEN Workshop Agreement 17284:2018 E, «Materials modelling: Terminology, classification and metadata», **2018**.

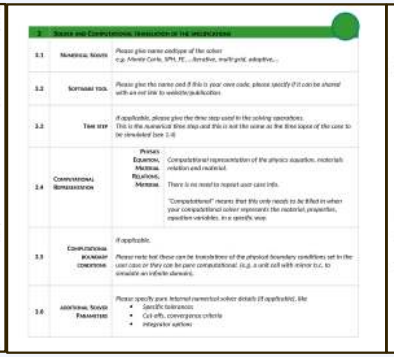
[2]A. F. de Baas, *What Makes a Material Function? Let Me Compute the Ways*, EU Publications, doi:10.2777/417118, **2017**.

[3]ReaxPro project deliverable D2.1, «ReaxPro MODA diagrams», **2020**.

The EMMC's first attempts

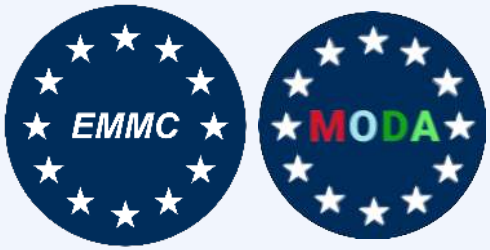
Focus on provenance documentation





- MODA was a **closed semantic and epistemic space**: Modelling methods had to be chosen from a small list.
- MODA imposed a **given level of detail** in workflow documentation; namely, **unrealistically detailed**.
- MODA documentations were **complicated** and **of limited use** to all, including to humans.[1, 2]

CWA 17284: 2018 E «MODA»



[1]ReaxPro project deliverable D2.1, «ReaxPro MODA diagrams», **2020**.

[2]«European standardization efforts from FAIR toward explainable-AI-ready data documentation in materials modelling», in *Proc. ICAPAI 2023*, doi:10.1109/icapai58366.2023.10193944, IEEE, **2023**.

Case study in molecular modelling

Epistemic metadata and their documentation were explored for the domain of molecular modelling and simulation within engineering thermodynamics:

First stage report (10 cases), doi:10.5281/zenodo.7516532, **2023**.

Discussion of *five papers each* from *two research groups* (London, Berlin) without involving the papers' authors. Obtained a tentative **taxonomy for epistemic metadata**, later implemented into the PIMS-II ontology.

Second stage report (12 claims), doi:10.5281/zenodo.7608074, **2023**.

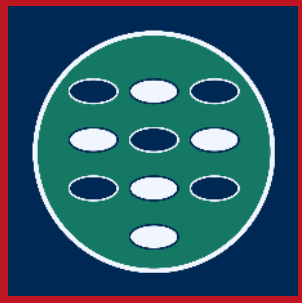
Discussion of *two claims each* from *six papers*, with two papers each from three research groups (London, Berlin, Kaiserslautern), involving the papers' authors. Discussed aspects such as the **grounding of knowledge claims** with authors.

Imperial College London

berlin

RPTU
Rheinland-Pfälzische Technische Universität

Noregs miljø- og biovitskaplege universitet



Domain and background

First stage and an ontology

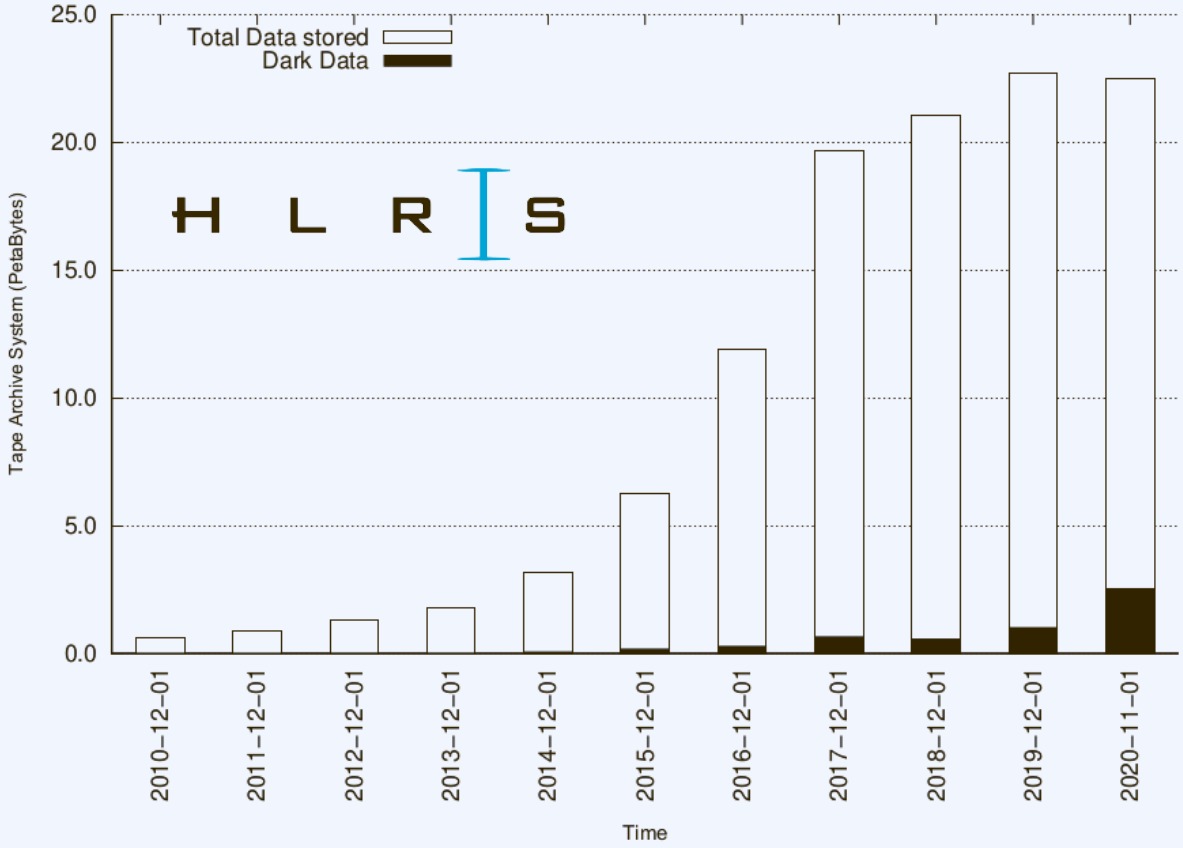
Second stage and grounding

Key epistemic metadata items are the **knowledge claims** made based on data, their **provenance**, **validation** and **reproducibility**, and **epistemic grounding**.

Problem and idea behind epistemic metadata

Dark data are data with an uncharacterized knowledge status.

In other words: We don't know what we know from and about the data.



Flood of dark data:
More and more data are accumulated, but are dark – and useless.

Discussed in work by Björn Schembera and Juan Durán.[1, 2]

[1]Figure from Björn Schembera's doctoral thesis, doi:10.18419/opus-11028, **2019**.

[2]B. Schembera, J. Durán, *Philos. Technol.* **33**: 93–115, doi:10.1007/s13347-019-00346-x, **2019**.

Problem and idea behind epistemic metadata

Epistemic metadata are the information that **establishes the knowledge status** of data or digital objects.[1]

Questions we must answer to establish the knowledge status:

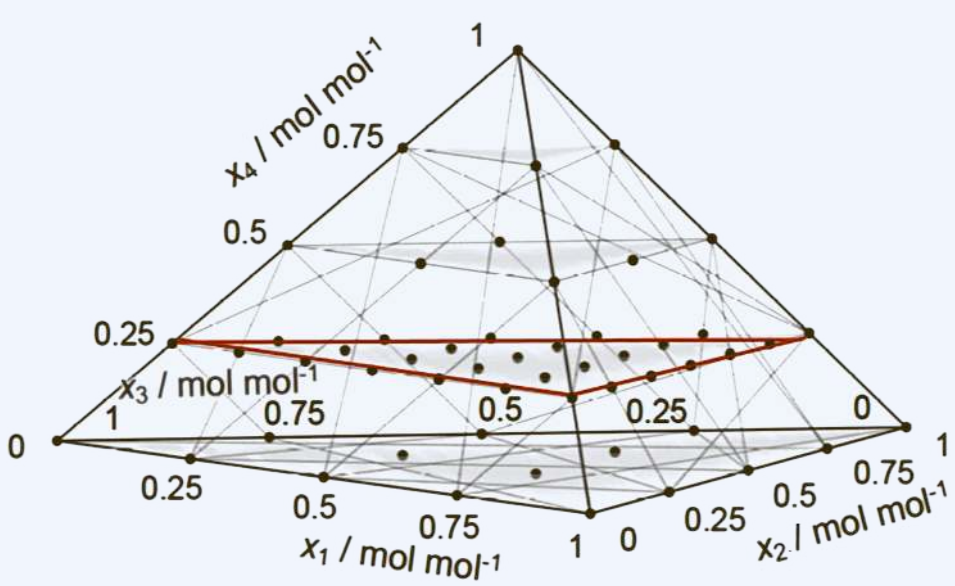
a) "what **knowledge claim** φ has been formulated?,"
b) "where do the data and the claim come from?" (**provenance**),
c) "what **validity claim** was made about φ?,"
d) "why should we accept any of this?" (**grounding**).

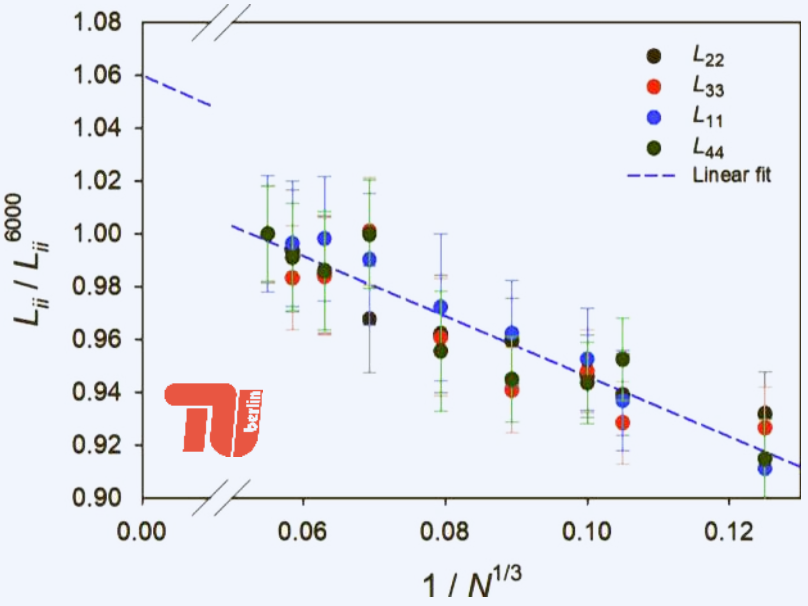
Key epistemic metadata items are the **knowledge claims** made based on data, their **provenance**, **validation** and **reproducibility**, and **epistemic grounding**.

[1]«Documentation of epistemic metadata by a mid-level ontology of cognitive processes», in *Proc. JOWO 2022*, CEUR *vol.* **3249**: *p. 2 (CAOS)*, CEUR-WS, **2022**.

The first stage of the case study

Example: The work by Guevara *et al.*[1] (2020) was considered at both stages.[2, 3]





[1]G. Guevara Carrión, R. Fingerhut, J. Vrabec, «Fick diffusion coefficient matrix of a quaternary liquid mixture by molecular dynamics», *J. Phys. Chem. B* **124**(22): 4527–4535, doi:10.1021/acs.jpcb.0c01625, **2020**.

[2]M. T. Horsch, B. Schembera, «Epistemic metadata in molecular modelling: First-stage case-study report (10 cases)», Inprodat technical report 2023–A, doi:10.5281/zenodo.7516532, **2023**.

[3]M. Horsch, S. Chiacchiera, G. Guevara, M. Kohns, E. Müller, D. Šarić, S. Simon, I. Todorov, J. Vrabec, B. Schembera, «Epistemic metadata in molecular modelling: Second-stage case-study report (12 claims)», Inprodat technical report 2023–B, doi:10.5281/zenodo.7610237, **2023**.

Guevara *et al.* (2020) paper:[1] First-stage analysis[2]

Question: What is a good methodology for obtaining Fick diffusion coefficients in multicomponent mixtures by [equilibrium molecular dynamics] simulation?

Object of research: The object of research is the Fick diffusion coefficient matrix as such.

Knowledge claim: […] methodology […] first, the explicit inclusion of a finite-size correction, where it is specifically novel that this correction is applied to the Onsager coefficients, and second, obtaining the Darken correction from [Kirkwood-Buff] integrals.

Grounding: KB part […] validated against "the Wilson excess Gibbs energy model […]" […] not clear what should make us accept the finite-size methodology […]. It yields a correction of 6% […] whereas the "[…] following Yeh and Hummer would have led to corrections of around 15%." It is based on a linear regression in $N^{-1/3}$ […] ad hoc fit.

[1]G. Guevara Carrión, R. Fingerhut, J. Vrabec, «Fick diffusion coefficient matrix of a quaternary liquid mixture by molecular dynamics», *J. Phys. Chem. B* **124**(22): 4527–4535, doi:10.1021/acs.jpcb.0c01625, **2020**.

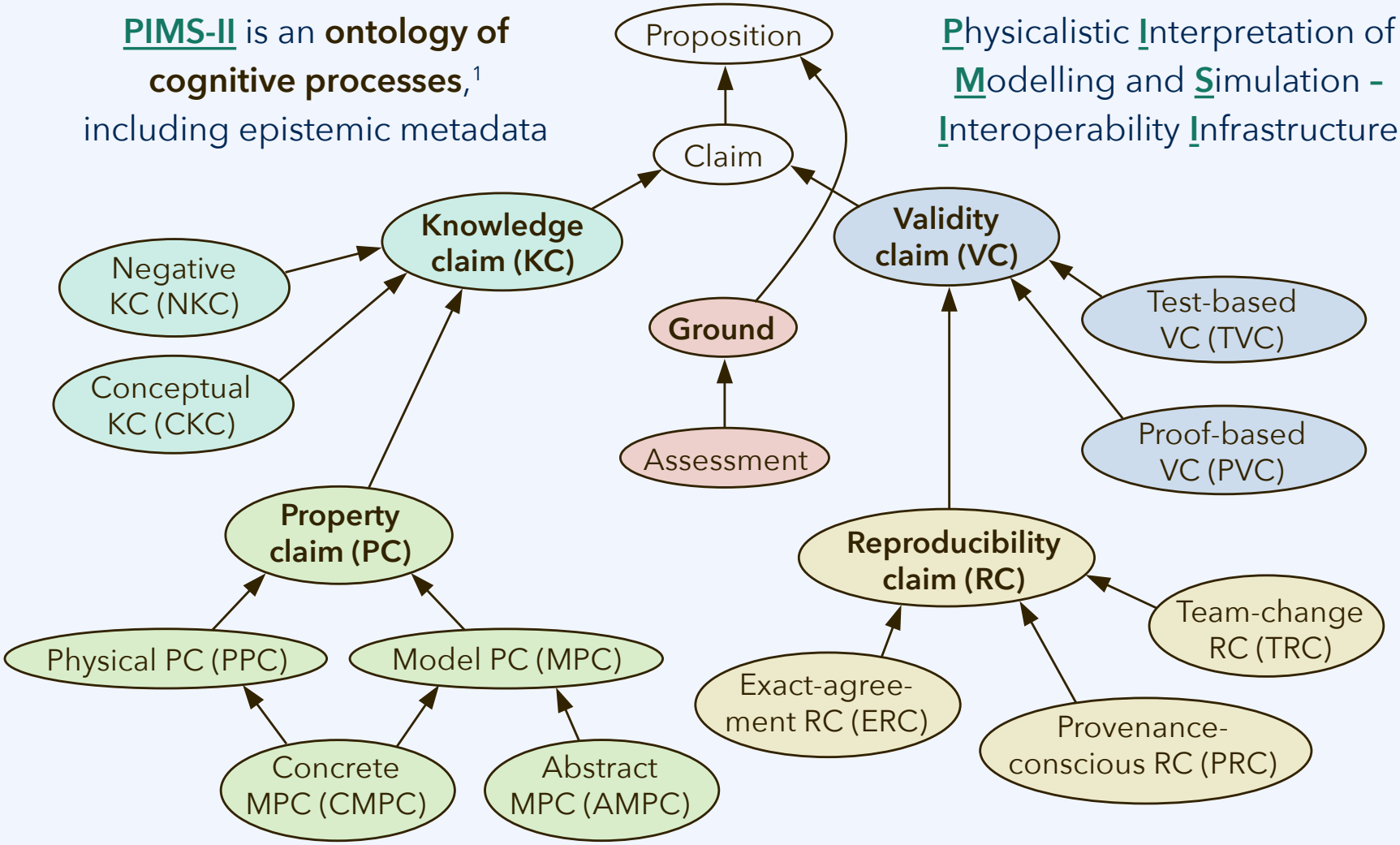
[2]M. T. Horsch, B. Schembera, «Epistemic metadata in molecular modelling: First-stage case-study report (10 cases)», Inprodat technical report 2023–A, doi:10.5281/zenodo.7516532, **2023**.

[3]M. Horsch, S. Chiacchiera, G. Guevara, M. Kohns, E. Müller, D. Šarić, S. Simon, I. Todorov, J. Vrabec, B. Schembera, «Epistemic metadata in molecular modelling: Second-stage case-study report (12 claims)», Inprodat technical report 2023–B, doi:10.5281/zenodo.7610237, **2023**.

Ontology of epistemic metadata[1]

PIMS-II is an **ontology of cognitive processes**,[1] including epistemic metadata

Physicalistic **I**nterpretation of **M**odelling and **S**imulation – **I**nteroperability **I**nfrastructure

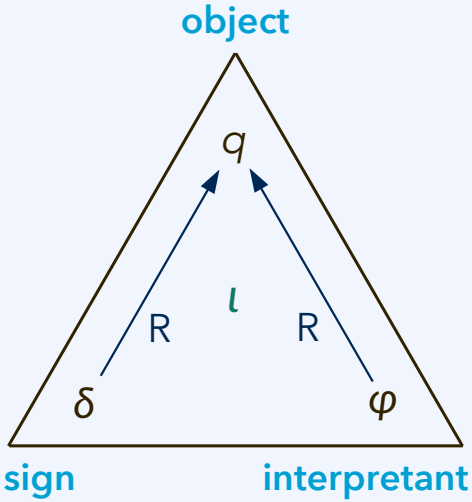


[1]OWL implementation under http://www.molmod.info/semantics/pims-ii.ttl

Ontology of epistemic metadata[1, 2]

Peircean semiotics is applied to the description of cognitive processes, *e.g.*, consider a process in which dataset δ is analysed, yielding knowledge claim φ:

- The data δ are about some research question q.
  So δ is a representamen for q; it has the role of the **sign**.
- The research question q is the **object** of the semiosis.
- As an outcome of the semiosis, claim φ is obtained, which is a new representamen for q, the **interpretant**.



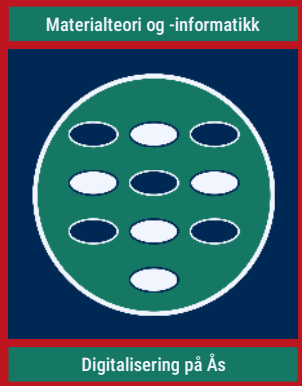
InformationProcessing
(*i.e.*, a Peircean semiosis)

The part of the PIMS-II ontology that deals with Peircean semiotics is also axiomatized in first-order logic,[1] in addition to the OWL implementation.[2]

C. S. Peirce

[1]«Mereosemiotics: Parts and signs», in *Proc. JOWO 2021 (FOUST)*, **2021**.
[2]OWL implementation under http://www.molmod.info/semantics/pims-ii.ttl

Noregs miljø- og biovitskaplege universitet



Domain and background

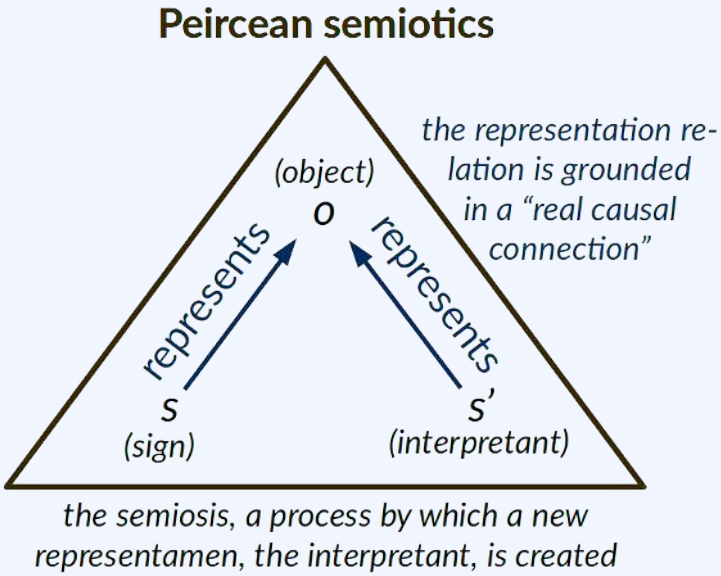
First stage and an ontology

Second stage and grounding

Knowledge claim (KC), including the provenance

«Researcher *a* did *κ* and found *φ* (and thus claims to know *φ*).»
→ Therefore, when research process *κ* is carried out, it **can** lead to outcome *φ*.

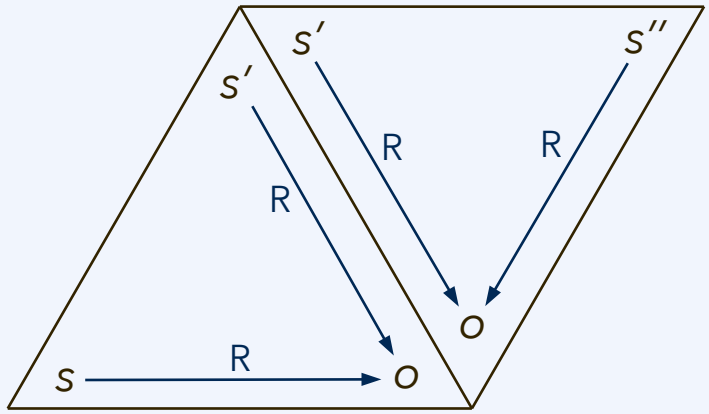
Grounding as implemented in PIMS-II (so far)



Cognitive process (example):

- First, experimental data *s* for material *o* are used to parameterize a model, obtaining model *s′*.
- Then, a simulation is done using model *s′*, yielding the simulation result *s″* (which also represents *o*).

Research workflows as cognitive processes:[1]



cognitive process κ

Each cognitive step starts from one representation relation, *e.g.*, *Rso*, and creates a new one, *Rs′o*.

The successor step reuses *Rs′o* and creates the next relation, *Rs″o*.

[1]«Mereosemiotics: Parts and signs», in *Proc. JOWO 2021 (FOUST)*, **2021**.

Guevara *et al.* (2020) claims:[1] Second-stage analysis[2]

Interviews were done with the authors; *e.g.*, on 24th January 2023, two papers, among them Guevara *et al.*[1] (2020) were discussed in a 70-minutes meeting. Two of the three authors participated (Gabriela Guevara and Jadran Vrabec).

«**Why is it knowledge?**

- Yeh & Hummer instead use a semiempirical correlation, relying on all sorts of properties, working with the end result which D.
- The new method is formally much simpler, relying only on N, and it works with the underlying quantity L which is more fundamental, rather than with the end outcome D.
- Also, linear behaviour of D in $1/N^3$ was already claimed before by others, and not only for D, it is something like "community shared understanding". In particular, Yeh-Hummer also has $1/N^3$ in it.

Validation:

- Is it better than Yeh-Hummer? Really such a validation still needs to be done.»

[1]G. Guevara Carrión, R. Fingerhut, J. Vrabec, «Fick diffusion coefficient matrix of a quaternary liquid mixture by molecular dynamics», *J. Phys. Chem. B* **124**(22): 4527–4535, doi:10.1021/acs.jpcb.0c01625, **2020**.

[2]M. Horsch, S. Chiacchiera, G. Guevara, M. Kohns, E. Müller, D. Šarić, S. Simon, I. Todorov, J. Vrabec, B. Schembera, «Epistemic metadata in molecular modelling: Second-stage case-study report (12 claims)», Inprodat technical report 2023–B, doi:10.5281/zenodo.7610237, **2023**.

Guevara *et al.* (2020) claims:[1] Second-stage analysis[2]

Interviews summarized in the second-stage report,[2] with two claims per paper.

Selected knowledge claims from the paper:

1. A novel finite-size correction methodology for the phenomenological diffusion coefficient matrix **L** based on linear extrapolation over $1/N^3$ to the limit $1/N^3 \to 0$ is proposed and successfully used to calculate **D**.

2. The Fick diffusion coefficient matrix **D** of the considered mixture has the values given in Table 1 of the paper under the conditions specified there.

d The novel method looks preferable or more plausible as it exhibits what is typically seen in the community as theoretical virtues: *First*, Yeh and Hummer [21] use a semiempirical correlation relying on multiple properties, while the novel method is formally much simpler, relying only on N. *Second*, the Yeh-Hummer method operates on the end result **D**, whereas the novel method operates on the intermediate result **L** that directly experiences the finite-size limitation in the molecular simulation.

[1]G. Guevara Carrión, R. Fingerhut, J. Vrabec, «Fick diffusion coefficient matrix of a quaternary liquid mixture by molecular dynamics», *J. Phys. Chem. B* **124**(22): 4527–4535, doi:10.1021/acs.jpcb.0c01625, **2020**.

[2]M. Horsch, S. Chiacchiera, G. Guevara, M. Kohns, E. Müller, D. Šarić, S. Simon, I. Todorov, J. Vrabec, B. Schembera, «Epistemic metadata in molecular modelling: Second-stage case-study report (12 claims)», Inprodat technical report 2023–B, doi:10.5281/zenodo.7610237, **2023**.

Type-1 and Type-2 grounding

Type-1 and Type-2 notions inspired by Marr.[1, 2]

Type-1	
Type-2 The **provenance** of the results tells that they are valid.	*Case study example: Šarić et al.* argue: – We are using ion models that worked accurately before. – It was shown before that ion models designed for one water model still perform accurately for another water model. – Therefore we can carry over the ion models designed for SPC/E water to another water model, namely, TIP4P/ε.

[1]D. Marr, *Artificial Intelligence* **9**(1): 37–48, doi:10.1016/0004-3702(77)90013-3, **1977**.

[2]«Documentation of epistemic metadata by a mid-level ontology of cognitive processes», in *Proc. JOWO 2022*, CEUR *vol.* **3249**: *p. 2 (CAOS)*, CEUR-WS, **2022**.

Type-1 and Type-2 grounding

Type-1 and Type-2 notions inspired by Marr.[1, 2]

Type-1 The results' validity is not grounded in the way the results were obtained.	*Typical:* **Mathematical proof** in statistical mechanics for a theoretical framework with widely accepted definitions and axioms.	*Case study example:* Fingerhut *et al.* introduce a method based on Kirkwood-Buff integration (building on previous theoretical work by Ben Naim)
Type-2 The **provenance** of the results tells that they are valid.	*Case study example: Šarić et al.* argue: – We are using ion models that worked accurately before. – It was shown before that ion models designed for one water model still perform accurately for another water model. – Therefore we can carry over the ion models designed for SPC/E water to another water model, namely, TIP4P/ε.	

[1]D. Marr, *Artificial Intelligence* **9**(1): 37–48, doi:10.1016/0004-3702(77)90013-3, **1977**.

[2]«Documentation of epistemic metadata by a mid-level ontology of cognitive processes», in *Proc. JOWO 2022*, CEUR *vol.* **3249**: *p. 2 (CAOS)*, CEUR-WS, **2022**.

Normative grounds and reliabilism

Type-1 The results' validity is not grounded in the way the results were obtained.	*Typical:* **Mathematical proof** in statistical mechanics for a theoretical framework with widely accepted definitions and axioms.
Type-2 The **provenance** of the results tells that they are valid.	**Reliability** of process *m* means that «If *S*'s believing *p* at *t* results from *m*, then *S*'s belief in *p* at *t* is **justified**».[1] → *Typical:* We used a **model**, **method**, and **simulation code** validated in the past and – usually – very accurate. (**process reliabilism**)

[1] J. M. Durán, N. Formanek, *Minds and Machines* **28**(4): 645–666, doi:10.1007/s11023-018-9481-6, **2018**.

Normative grounds and reliabilism

	«**evidence** supporting trustworthiness **cannot be complete**»[1] **trust**	«**reliance** is compatible with – ideally – the **complete evidence**»[1] **reliance**
Type-1 The results' validity is not grounded in the way the results were obtained.	*Typical:* Mathematical proof in statistical mechanics for a theoretical framework with widely accepted definitions and axioms.	***Case study example:*** Guevara *et al.* argue: Our new finite-size correction is better because it is more simple and because it is applied to the directly computed quantity **L**.
Type-2 The **provenance** of the results tells that they are valid.	***Case study example:*** Chatwell and Vrabec argue: It is OK to use a cutoff radius of 5.5σ for the LJ potential, since this was done in three cited works from the literature.	*Typical:* We used a model, method, and simulation code validated in the past and – usually – very accurate. (**process reliabilism**)

See **reference ontology of trust** (ROT) by Baratella *et al.*[1]

[1]Baratella *et al.*, «The many facets of trust», to appear in *Proceedings of FOIS 2023*.

Normative grounds and reliabilism

	trust	reliance
Type-1 The results' validity is not grounded in the way the results were obtained.	*Typical:* Mathematical proof in statistical mechanics for a theoretical framework with widely accepted definitions and axioms.	*Schema:* A new theory is more reliable because it is simpler, covers more phenomena, or represents underlying physics. (**theoretical virtues**)
Type-2 The **provenance** of the results tells that they are valid.	*Case study example:* Chatwell and Vrabec argue: It is OK to use a cutoff radius of 5.5σ for the LJ potential, since this was done in three cited works from the literature.	*Typical:* We used a model, method, and simulation code validated in the past and – usually – very accurate. (**process reliabilism**)

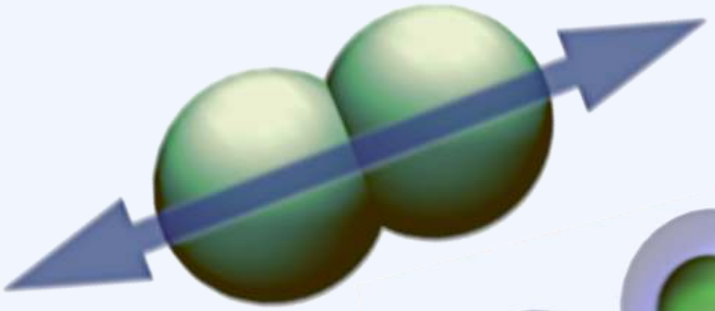
Schema: Grounding of knowledge claims

	trust	reliance
Type-1	rare 	very frequent 
Type-2	very frequent 	frequent 

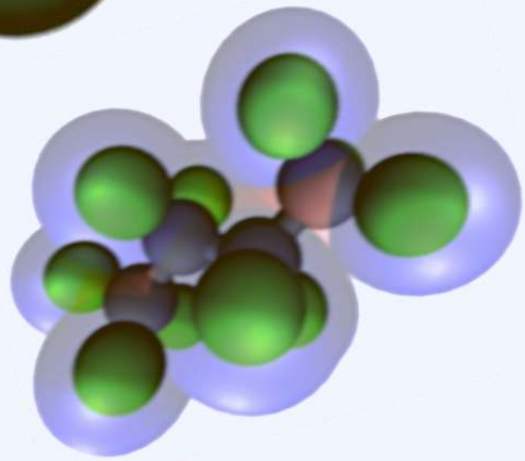
Sample: Bowskill *et al.*, Chatwell & Vrabec, Fingerhut *et al.*, Guevara *et al.*, Stephan & Hasse (long-range).

Theoretical virtues

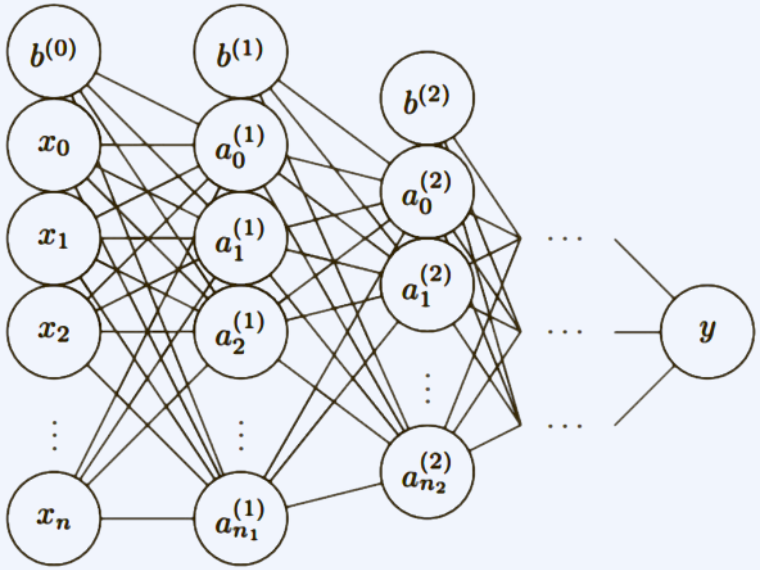
Theoretical virtues often oppose each other.



Simplicity favours few elements and few model parameters.



Representing all the physics **realistically** requires the opposite.



Neural networks are lacking in virtue, yet some people use them.

[1]Figure sources: MolMod DB (link) and Anna Jenul's doctoral thesis (link).

Theoretical virtues

Theoretical virtues found in the case study:

- **Alignment** of representations, qualitative reflection of underlying physics
  - Chatwell & Vrabec: Relaxation model based on an exponential decay as deduced theoretically; functional form of the EOS made plausible by theory.
  - Zhu and Müller note that ML models are devoid of any physics-based insights.
- **Coverage** or good sampling of the phenomenon's state space
  - Bowskill *et al.*: Test cases are representative of real world problems.
  - Stephan and Hasse: Selected mixtures and conditions are representative.
- **Mechanism** or explainability of dependencies
  - Guevara *et al.*: Finite-size effect on immediate result (**L**), not end result (**D**).
- **Simplicity**
  - Guevara *et al.*: Finite-size correction depends on one quantity only (the size).

Challenges at documenting grounding

It is hard to differentiate between grounding in virtues or in process reliabilism.

- Model *s'* is better than *s*. It is equally accurate with fewer parameters,
  - … so we should prefer it because it is **more simple**.
  - … so, **from experience**, its extrapolations are **more reliable**.
    - » Nobody writes either of the above explicitly.

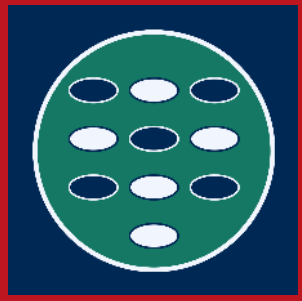
When digitalizing research data, we should respect that:

- Research is a **social process** among humans;
- scientific communication is **human communication**;
- it can rely on **pragmatics** – no need say every small thing explicitly;
- **epistemic grounding** is **usually not spelled out** in detail (or at all); we would often need to impute an interpretation onto the authors.

The ontology's aim is: Help people make more explicit statements if they want.

Noregs miljø- og biovitskaplege universitet

Materialteori og -informatikk

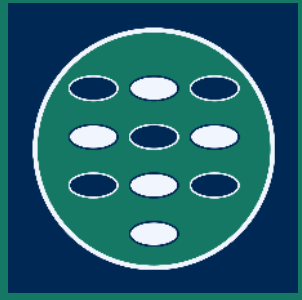


Digitalisering på Ås

Domain and background
First stage and an ontology
Second stage and grounding

Norges miljø- og biovitenskapelige universitet

Materialteori og -informatikk



Digitalisering på Ås

Universität Stuttgart

Science and Technology Facilities Council

Case study on epistemic metadata in molecular modelling

Martin Horsch,[1,2] Silvia Chiacchiera,[2] and Björn Schembera[3]

[1]Norwegian Univ. Life Sciences, Faculty of Science and Technology, Ås, Norway
[2]UKRI STFC Daresbury Laboratory, Scientific Computing Department, Daresbury, UK
[3]Univ. Stuttgart, Institute of Applied Analysis and Numerical Simulation, Stuttgart, Germany

Fakultet for realfag og teknologi

Forskergruppe materialteori og -informatikk