

*Shared Metadata and Data Formats for Big-Data Driven
Materials Science: a NOMAD/FAIR-DI Workshop*

**Open questions and needs from
hard and soft materials:
discussion on molecular mechanics**

Moderated by **James Kermode**

Warwick Centre for Predictive Modelling
School of Engineering, University of Warwick, United Kingdom

Metadata – using, storing and sharing

- **My background – using and reusing atomistic configurations**
 - Concurrent multiscale simulations (QM/MM)
 - Fitting interatomic potentials, uncertainty quantification
 - Experience of what's needed for a minimal workable code-independent format
- **Drivers for sharing**
 - Reproducibility
 - Validation & verification – e.g. as part of continuous integration testing of software
 - Archiving of research data – may be required by funders, journals, or simply useful!
 - Sharing research data between collaborators (pre-publication) and wider (afterwards)
- **Barriers**
 - Technical – file formats, sharing mechanisms, access control
 - Social – cost/benefit, single or few configurations give small benefit, retain control
- **Consensus on distributed database solution**
 - Emerging consensus on computational framework – Python and/or REST APIs
 - Growing number of large datasets that have high value (e.g. OPTiMaDe)
 - Enable reuse of data by machine learning and data mining methods
- **Need for minimal, discoverable FAIR standard for exchanging data and metadata**

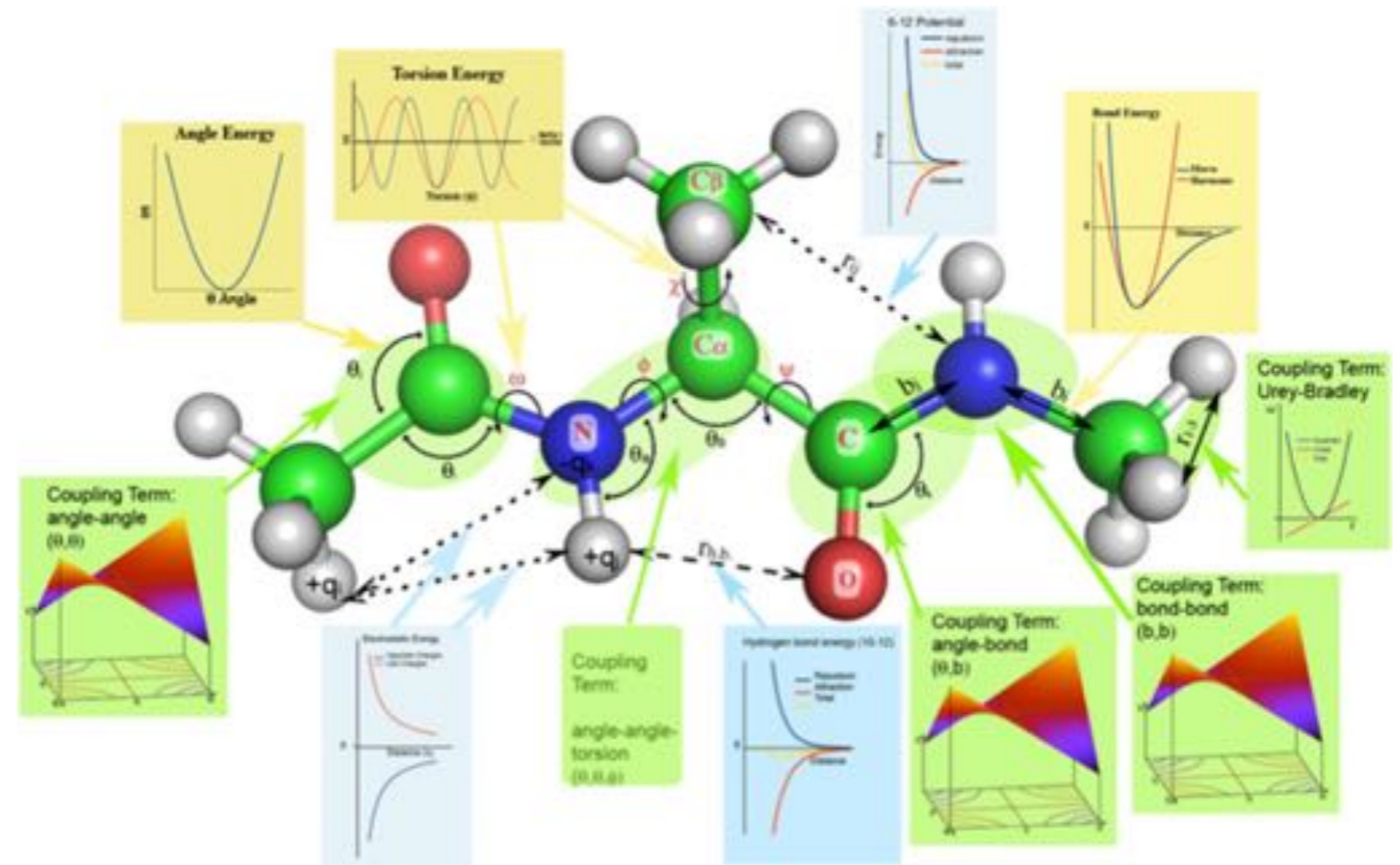


Metadata Challenges for Force Field codes

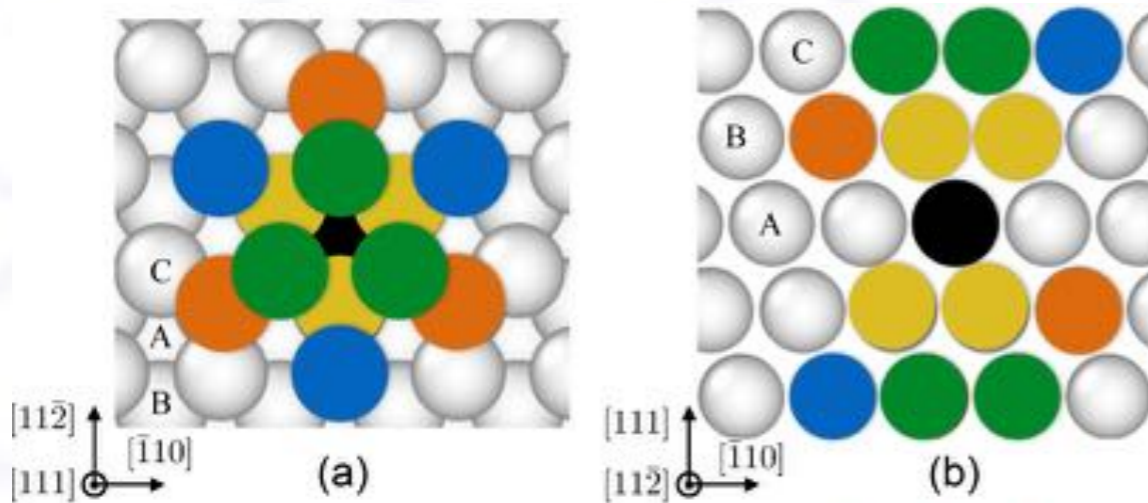
- Representing initial conditions (atoms, cell) **Common to DFT and FF**
- Defining output properties (quantities of interest) and uncertainties
- Representing sampling of phase space (thermostats, constraints, ...)
- Large quantity of data produced **Both – more pronounced in FF**
- Representing interatomic potentials (aka force fields)
- Representing topology (atom types, bonds, angles) **FF specific**

Representing Potentials: Materials & Molecules

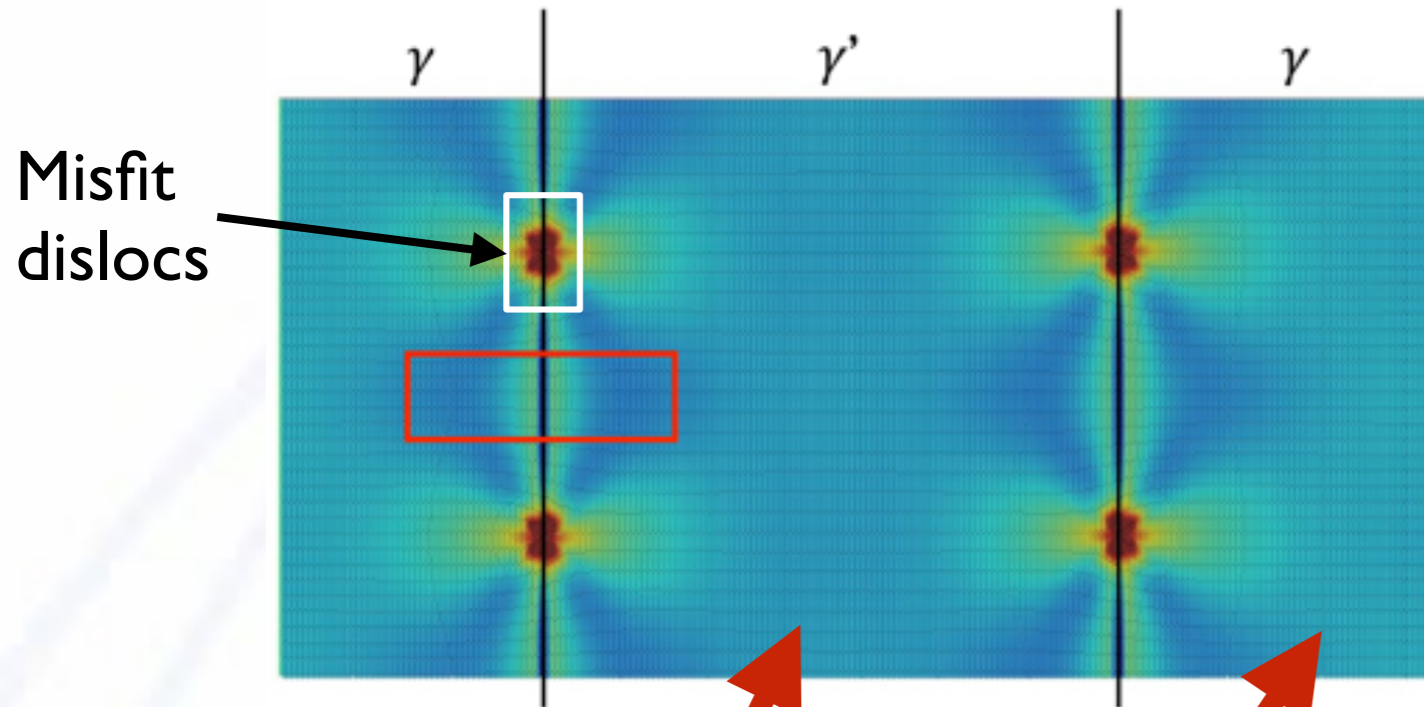
- Potentials used in computational materials science and in biomolecular simulation pose different metadata challenges
- Comp mat sci FFs typically reactive, fully many-body with finite cutoffs
- Functional forms may be analytic, tabulated or implicitly defined from data (e.g. ML potentials)
- Some potentials involve complex algorithms, eg. charge equilibration, polarisation
- Only truly defined by implementation?



- In bio sim, functional forms can be simpler (but not always, CG potentials also fully many-body; long-range electrostatic method must be carefully defined)
- More than one atom type for a given chemical species (e.g. C, C α)
- Typically unreactive – but topology must be built first, often following complex rules
- Often hard to migrate simulations between codes?

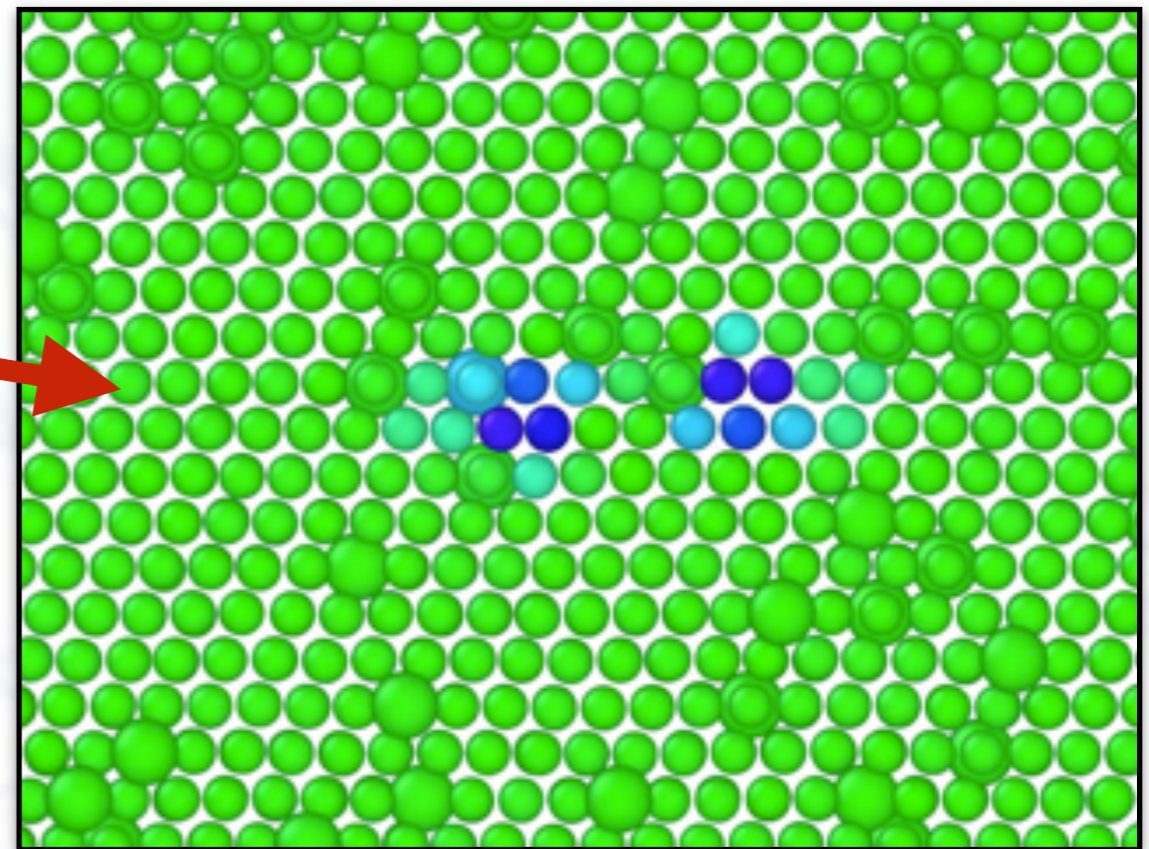
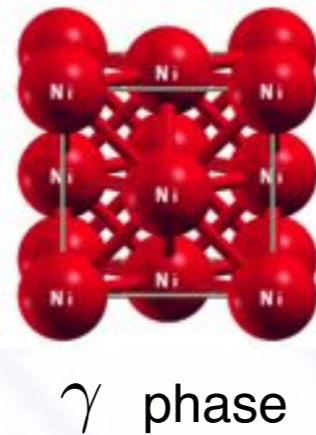
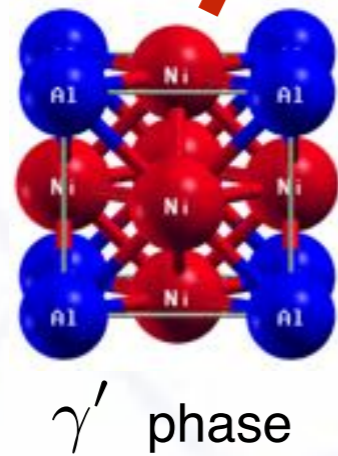


Use case: dislocation glide in Ni-based superalloys



MD simulation of dislocations in γ phase Ni

- **QoI:** dislocation core splitting with associated uncertainty (cf. TEM experiments)
- Aleatoric and epistemic uncertainties
- Model uncertainty – both in parameters and functional form
- Random microstructures, limited runtime
- Algorithmic uncertainty in solvers
- How much of this can/should be stored?
- Complex simulation script \approx workflow!

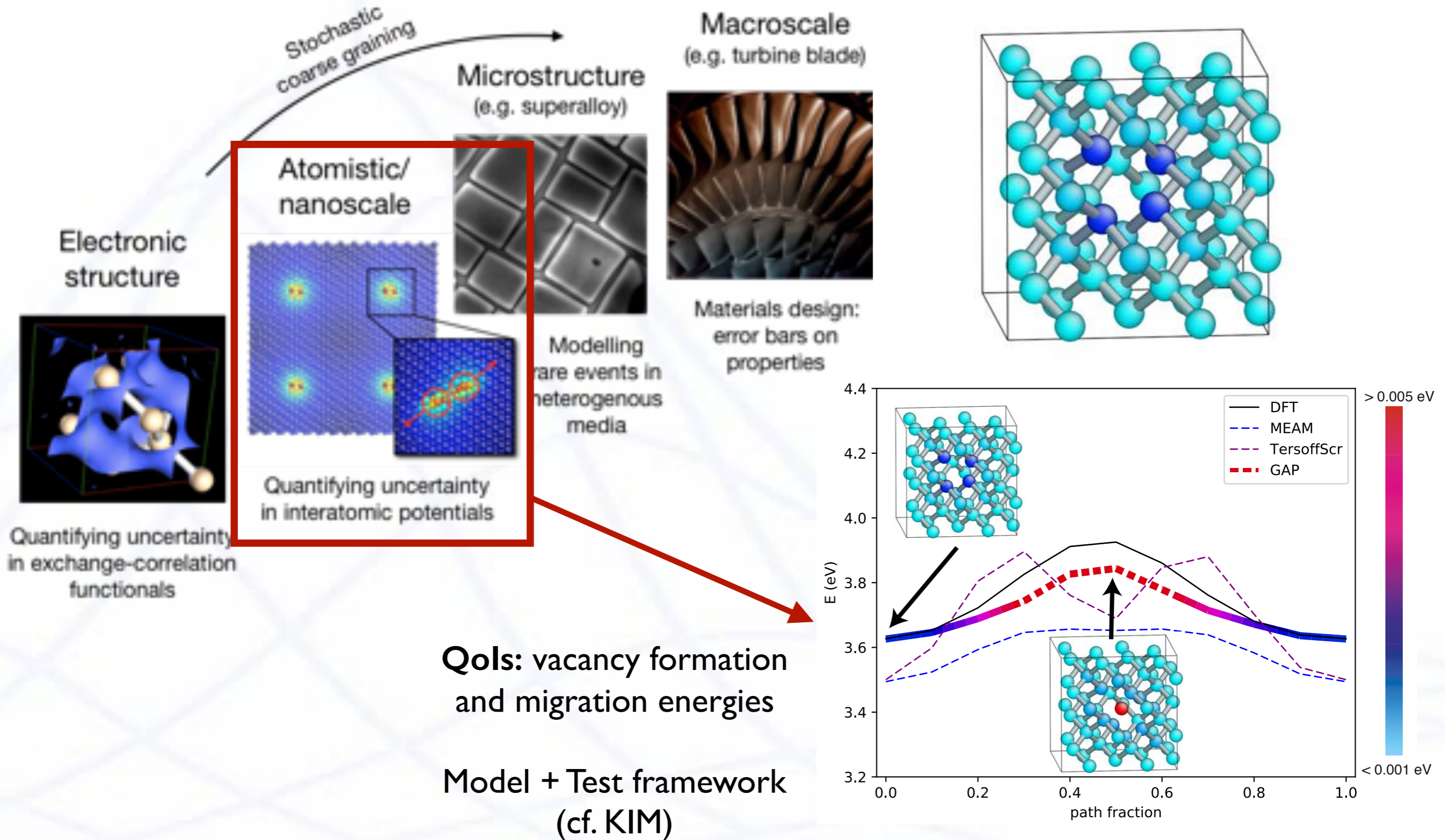


EAM, 5% Al, $T = 300$ K, 100 MPa shear stress

F Bianchini, JRK and A De Vita, *Modell. Simul. Mater. Sci. Eng.* **24** 045012 (2016)

F Bianchini, A Glielmo, JRK and A De Vita **3** 043605 *Phys. Rev. Mat.* (2019)

Reproducibility & Uncertainty Quantification





A. P. Bartok, JRK, N. Bernstein and G. Csanyi, PRX **8**, 041048 (2018)

<https://github.com/libAtoms/silicon-testing-framework>

<https://doi.org/10.5281/zenodo.1250555>

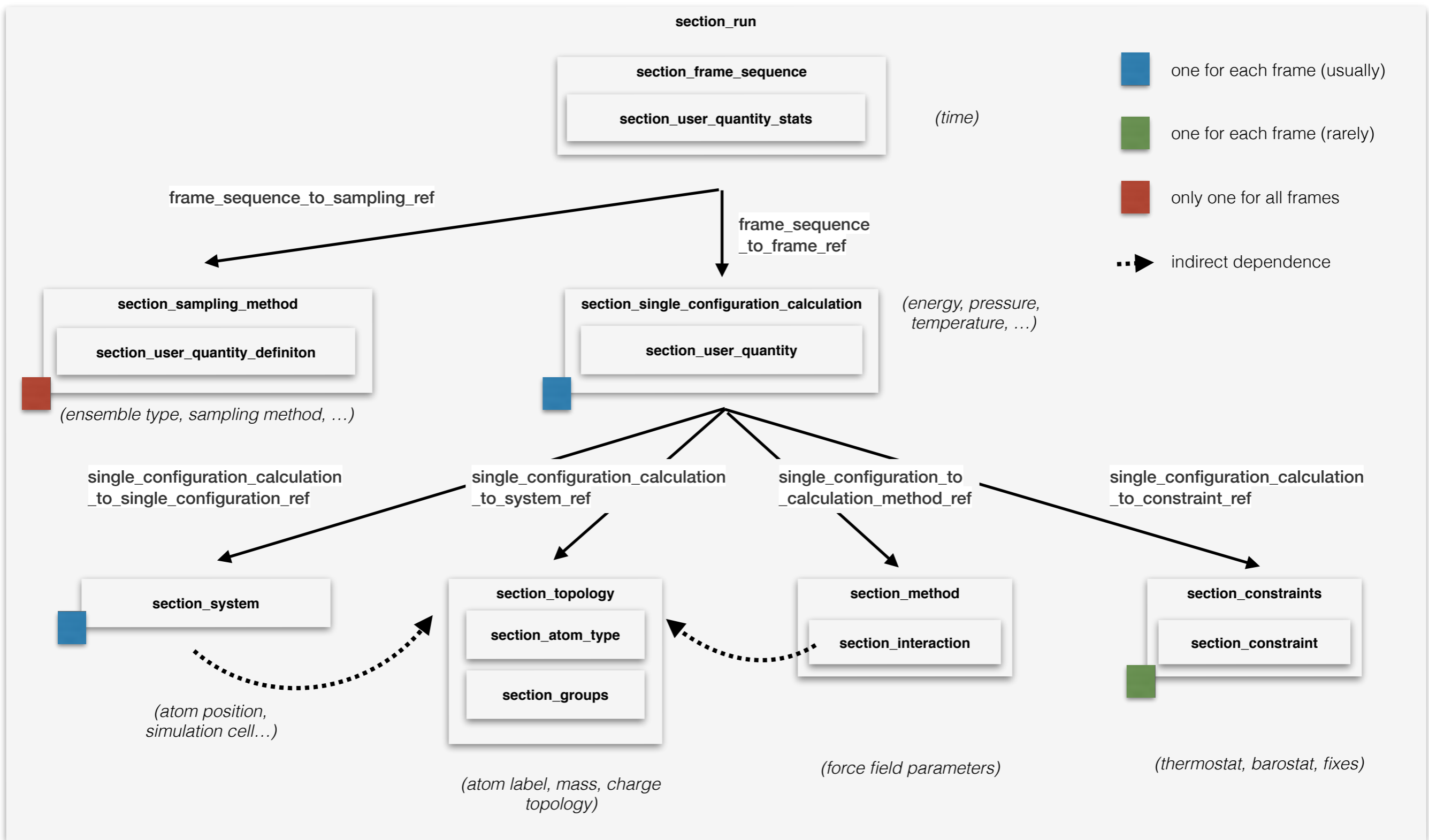
Status of support for MD codes in NOMAD

Code	Citations (2013-17)	Type	Search Name
Gaussian	19100	DFT	Frisch
VASP	17900	DFT	Kresse
Gromacs	11200	FF	Lindahl
LAMMPS	10300	FF	Plimpton
Amber	9440	FF	Kollman
NAMD	7110	FF	Schulten
GROMOS	7080	FF	Van Gunsteren
Quantum Espresso	6960	DFT	Giannozzi
ASE/ASAP	6650	FF	Jacobsen
CHARMM	6250	FF	Karplus
Discovery Studio	6240	DFT, FF	<i>Accelrys</i>
GAMESS	5780	DFT	Gordon
WIEN2k	5570	DFT	Blaha
CASTEP	5330	DFT	Payne
Molpro	4440	DFT	Werner

 Parser codes developed in our group.
 Parsers from other groups in NOMAD.

- Conversion layers for the most popular MD codes have been implemented
- Remaining challenges including fully representing force fields, topology, sampling schemes, etc.
- Prototype interface between NOMAD & OpenKIM:
 - ✓ Enables extraction of reference data from NOMAD to KIM for validation
 - ✗ Matching KIM model IDs to NOMAD data – not implemented; currently model IDs can be output by LAMMPS, but not by other supported codes

NOMAD Metadata for MD simulations



Metadata Challenges for Force Field codes

- Representing initial conditions (atoms, cell) **Common to DFT and FF**
- Defining output properties (quantities of interest) and uncertainties
- Representing sampling of phase space (thermostats, constraints, ...)
- Large quantity of data produced **Both – more pronounced in FF**
- Representing interatomic potentials (aka force fields)
- Representing topology (atom types, bonds, angles) **FF specific**

Questions for Discussion

1. How should we deal with the large quantity of data produced by MD simulations? Leave it up to users? Subsample trajectories? If so, by what metric?
2. How can interatomic potentials be represented?
cf. OpenKIM (comp mat sci), OpenForceField and MolSSI (bio sim)
3. For molecular systems, do we need code independent representations of topology? (atom types, bonds, angles, dihedrals, what else?)
4. Can we represent sampling of phase space in a code independent way? (thermostats, constraints, algorithms, workflows, ...)
5. Is containerisation of codes part of the solution (e.g. Docker, Singularity)?
Is this just an excuse for bad dependency management by developers?
6. How precisely can/should we define output properties? Instantaneous vs. averaged?
Are ontologies needed? (cf. KIM properties framework)
7. Do we need/want a single metadata/ontology approach to achieve critical mass?
8. Do workflows in MD calculations need special treatment?
9. How should changes to metadata be managed?
10. Anything else?