NOMAD
NOVEL MATERIALS DISCOVERY

FAIR-DI

Metadata towards FAIR data sharing for data-driven materials science: achievements and open challenges

Luca M. Ghiringhelli
FRITZ-HABER-INSTITUT
MAX-PLANCK-GESELLSCHAFT

Shared metadata and data formats for Big-Data Driven Materials Science:
a NOMAD-FAIRDI workshop
IRIS Adlershof, Berlin, Germany, July 8th to 12th, 2019

For data-driven (materials) science...

NOVEL MATERIALS DISCOVERY

Mark D. Wilkinson, Michel Dumontier, *et al.*
The FAIR Guiding Principles for scientific data management and stewardship
Scientific Data **3**, 160018 (2016)

"Data" includes "algorithms, tools, and workflows that led to that data".

For data-driven (materials) science...

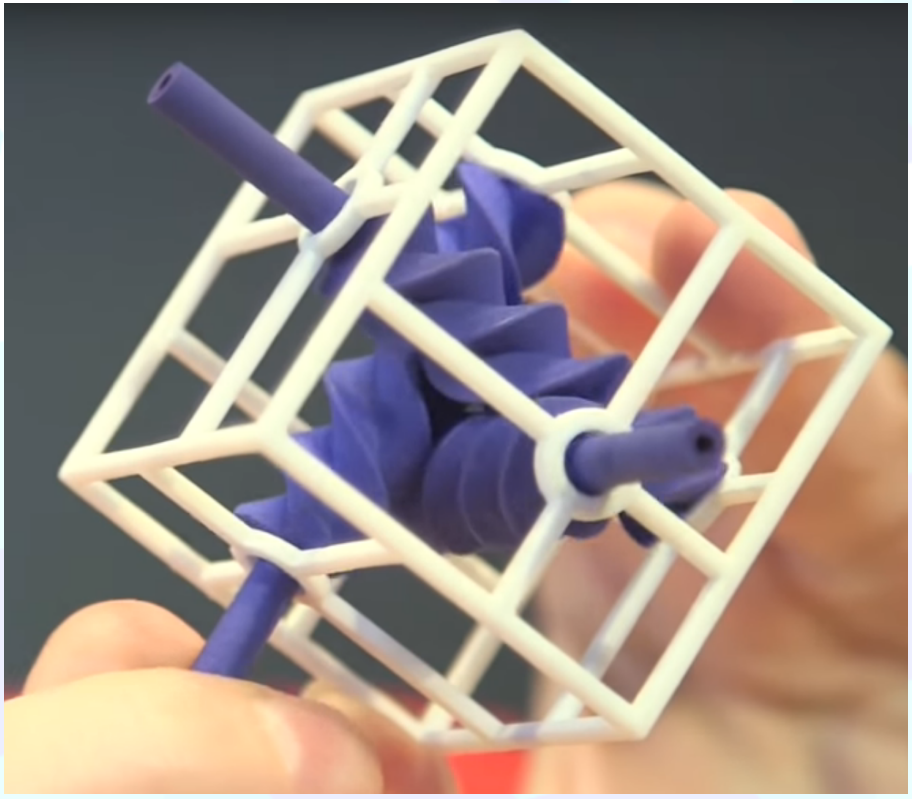
NOVEL MATERIALS DISCOVERY

Mark D. Wilkinson, Michel Dumontier, *et al.*
The FAIR Guiding Principles for scientific data management and stewardship
Scientific Data **3**, 160018 (2016)

Findable **A**ccessible **I**nteroperable

Fun facts on the three-gears contraption

NOVEL MATERIALS DISCOVERY



www.youtube.com/watch?v=5Mf0JpTI_gg

For data-driven (materials) science...

NOVEL MATERIALS DISCOVERY

Mark D. Wilkinson, Michel Dumontier, *et al.*
The FAIR Guiding Principles for scientific data management and stewardship
Scientific Data **3**, 160018 (2016)

Findable **A**ccessible **I**nteroperable **R**eusable

For data-driven (materials) science...

NOVEL MATERIALS DISCOVERY

Mark D. Wilkinson, Michel Dumontier, *et al.*
The FAIR Guiding Principles for scientific data management and stewardship
Scientific Data **3**, 160018 (2016)

Findable
Metadata are unique

Accessible
Data can be located somewhere

Interoperable
Useful ontologies are defined

Reusable
Metadata are "rich"

For data-driven (materials) science...

NOVEL MATERIALS DISCOVERY

In this talk:
- NOMAD and FAIR-DI experience

In the next two half days:
- Introduction to ontologies

- Materials Genome Initiative at NIST
- The Molecular Science Software Institute (MolSSI)
- AFLOW
- The Materials Project
- Aiida and Materials Cloud
- OpenKIM
- Simstack at KIT
- OPTiMaDe
- ABCD database at University of Cambridge

The NOMAD Laboratory

A European Centre of Excellence

PROJECT INDUSTRY TEAM RELATED PROJECTS NEWS PRESS KIT CONTACT US

Enter Search...

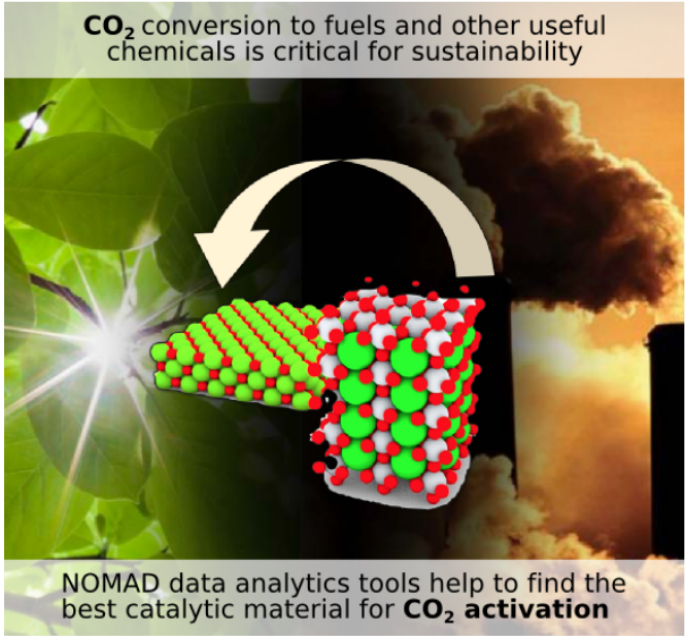
NOMAD REPOSITORY | THE ARCHIVE | ENCYCLOPEDIA | BIG-DATA ANALYTICS | ADVANCED GRAPHICS | HPC INFRASTRUCTURE | OUTREACH

https://nomad-coe.eu/
NOvel **MA**terials **D**iscovery Laboratory

NOMAD Scope and Overview

Data is a crucial raw material of the 21st century.

NOMAD Success Stories



The NOMAD Laboratory

NOVEL MATERIALS DISCOVERY

Live since 12.2014

NOMAD REPOSITORY

CONVERSION LAYER

THE ARCHIVE

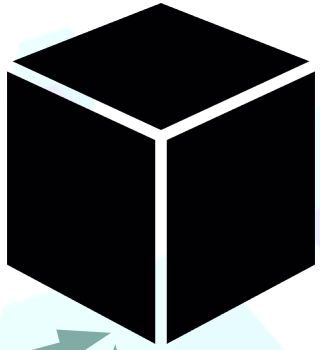
ENCYCLOPEDIA

ADVANCED GRAPHICS

BIG-DATA ANALYTICS

Computational materials science: basics

NOMAD
NOVEL MATERIALS DISCOVERY



- Code name
- Version
- Libraries

Input structure
- Coordinates
  $Z_i \; x_i \, y_i \, z_i$
- Cell vectors
- (Topology)

Output
- Total energy
- Forces
- Electron density
- Electrostatic pot.
- El. band structure
- Self energy
- …

Input model
- xc treatment / force field
- Relativity treatment
- Basis set
- Numerical integr. settings

More complex input model
- Model_1 (scf)→ Model_2 (perturbative)
- QM/MM (space partitioning of models)
- …

Computational materials science: basics

NOMAD
NOVEL MATERIALS DISCOVERY

- Code name
- Version
- Libraries

Input structure
- Coordinates
- Cell vectors
- (Topology)

Input model
- xc treatment / force field
- Relativity treatment
- Basis set
- Numerical integr. settings

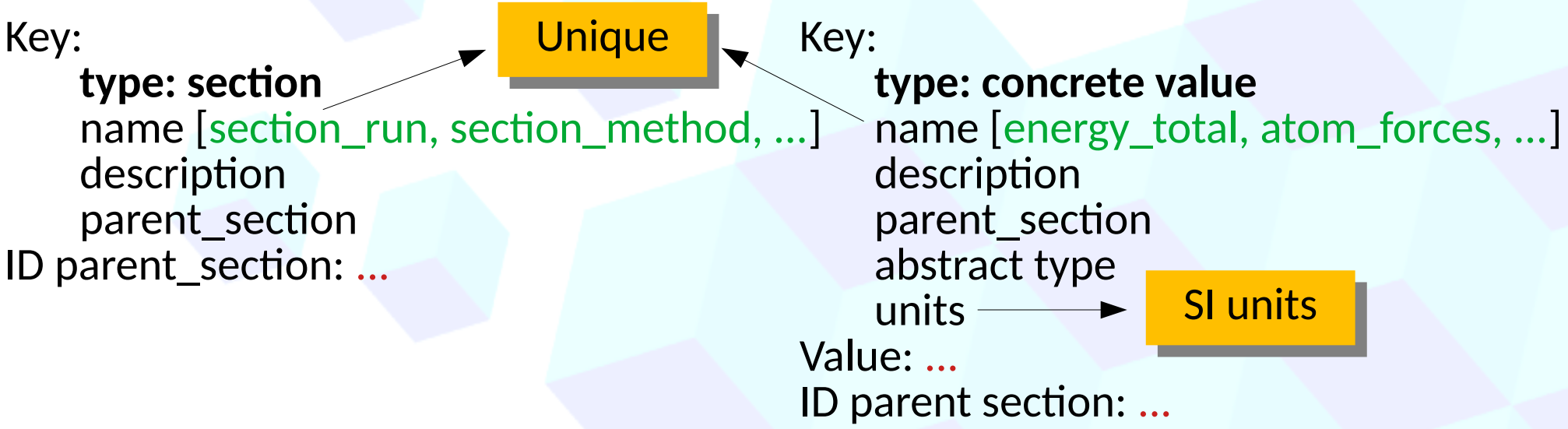
Output
- Total energy
- Forces
- Electron density
- Electrostatic potential
- Electronic band structure
- Self energy

Our solution: NOMAD MetaInfo

NOVEL MATERIALS DISCOVERY

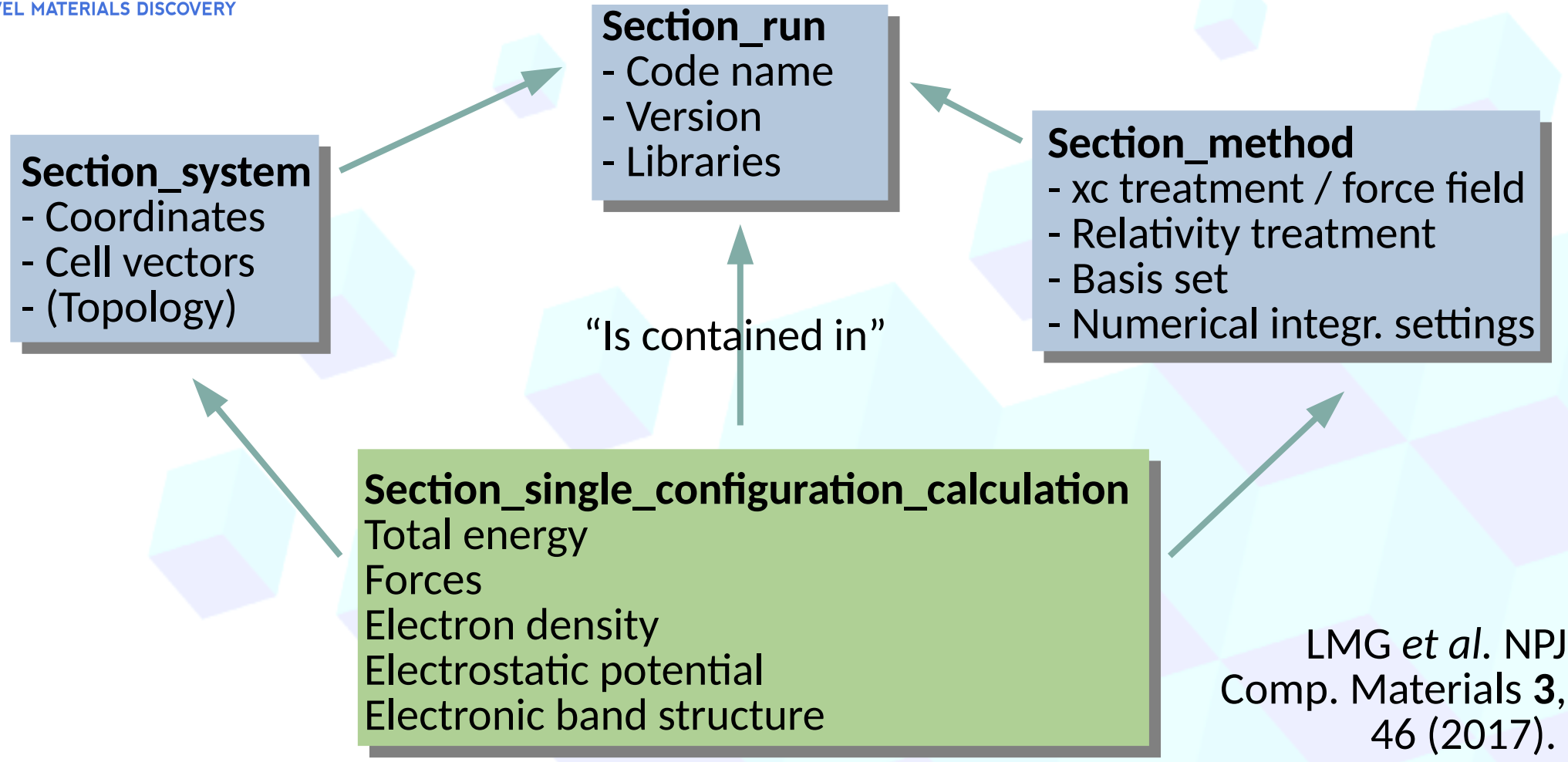
Metadata: in general dictionaries of key:value pairs, e.g.:
Name: ...
Address: ...
Passport N.:...

NOMAD MetaInfo, basic elements:



A posteriori, hierarchical scheme

NOVEL MATERIALS DISCOVERY



LMG *et al.* NPJ Comp. Materials **3**, 46 (2017).

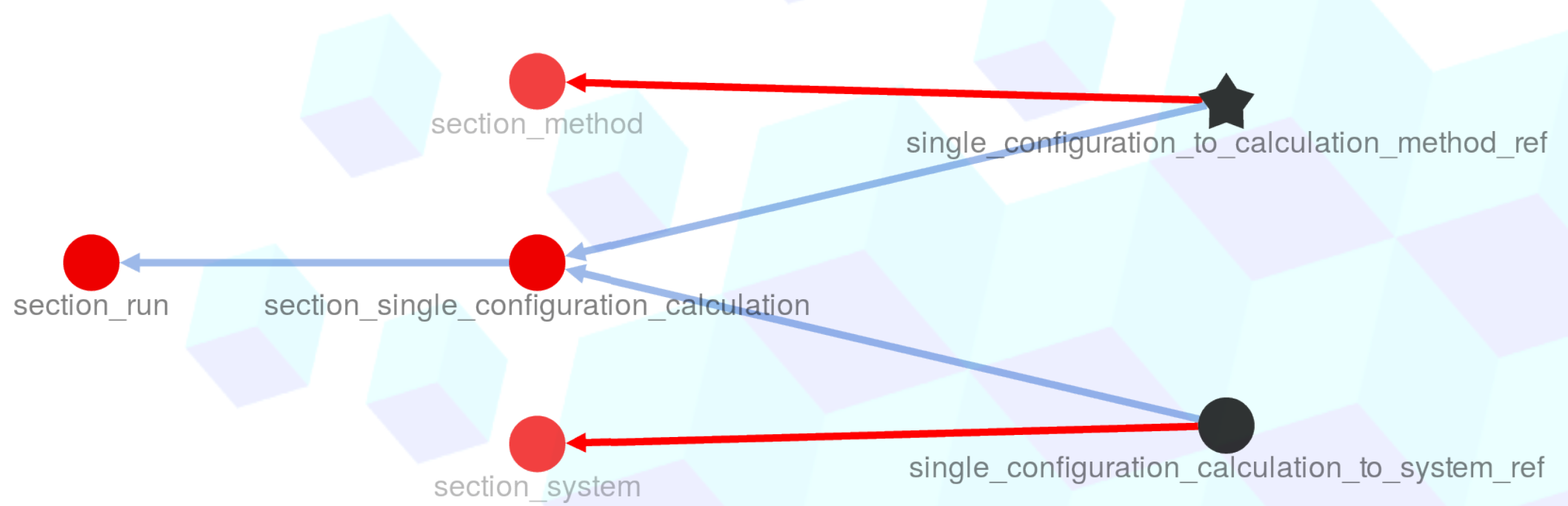
NOMAD

https://metainfo.nomad-coe.eu

NOVEL MATERIALS DISCOVERY

NOMAD

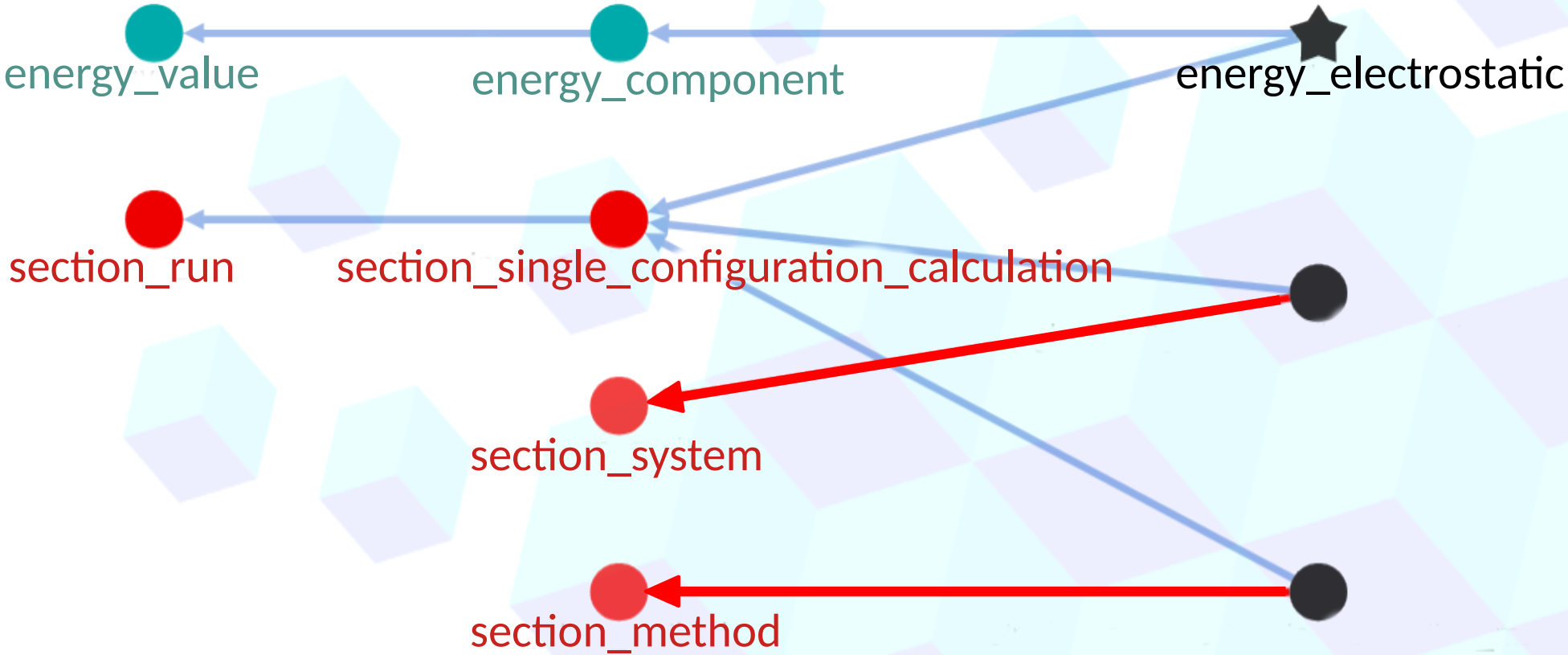
NOMAD Meta Info



energy_electrostatic

Type: Concrete Value
Description: Total electrostatic energy (nuclei + electrons), defined consistently with calculation_method
Data Type: floating point value
Shape: []
Units: J

Abstract type → Ontology



NOMAD MetaInfo, implementation

NOVEL MATERIALS DISCOVERY

The NOMAD team wrote 40+ parsers.
Raw input+output → NOMAD MetaInfo → json + hdf5 file format → Archive

json

```
section_run
    program_name                                  FHI-aims
    program_version                               081912
  section_system
      simulation_cell                             [[1.4e-9 ...]]
      atom_positions                              [[0.0,....]...]
      atom_labels                                 ["Cu",...]
  section_method
      basis_set                                   fhi_aims_tight
      XC_method                                   DFT_GGA_PBE
  section_single_configuration_calculation
    section_scf_iteration
        energy_total_scf_iteration                -1.326e-20
    section_scf_iteration
        energy_total_scf_iteration                -1.344e-20
  energy_total                                    -1.344e-20
```

hdf5
- Binary file format
- Storing of vectors, matrices,
- Efficient for
  non-sequential reading

NOMAD MetaInfo, beyond raw data

NOVEL MATERIALS DISCOVERY

The NOMAD team wrote 40+ parsers.
Raw input+output → NOMAD MetaInfo → json + hdf5 file format → Archive

Standardization
Raw input+output → NOMAD MetaInfo "as they are" (except unit conversion)

Normalization
Standardized NOMAD MetaInfo → derived MetaInfo

e.g., number density = #atoms / volume
Actual examples:
- band structure along path defined in W. Setyawan and S. Curtarolo, Comput. Mater. Sci. **49**, 299-312 (2010).
- space group calculated from structure via *spglib* library

Useful for:
storing of
"good" descriptors

NOMAD MetaInfo, community-driven

NOVEL MATERIALS DISCOVERY

Perspective paper:
L.M. Ghiringhelli, C. Carbogno, S. Levchenko, F. Mohamed, G. Huhs, M. Lueders, M. Oliveira, and M. Scheffler
Towards efficient data exchange and sharing for big-data driven materials science: Metadata and data formats.
NPJ Computational Materials **3**, 46 (2017). DOI: 10.1038/s41524-017-0048-5.

After CECAM-Psi-k workshop:
Towards a Common Format for Computational Materials Science Data
Lausanne, Switzerland, January 25 to January 27 2016

NOMAD MetaInfo, community-driven

NOVEL MATERIALS DISCOVERY

Perspective paper:
L.M. Ghiringhelli, C. Carbogno, S. Levchenko, F. Mohamed, G. Huhs, M. Lueders, M. Oliveira, and M. Scheffler
Towards efficient data exchange and sharing for big-data driven materials science: Metadata and data formats.
NPJ Computational Materials **3**, 46 (2017). DOI: 10.1038/s41524-017-0048-5.

- A common energy zero for total energies
- Electronic and vibrational properties of solids
- Compact representation of scalar fields: density, wavefunction, xc potentials, etc.
  S.V. Levchenko and M. Scheffler, *Compact representation of one-particle wavefunctions and scalar fields obtained from electronic-structure calculations*. Comput. Phys. Comm. **237**, 42-46 (2019).
- Quantities related to excited-state calculations
  many-body perturbation theory (MBPT) calculations (GW, Bethe–Salpeter equation, etc.)

NOMAD MetaInfo, accessibility

NOVEL MATERIALS DISCOVERY

The NOMAD team wrote 40+ parsers.
Raw input+output → NOMAD MetaInfo → json + hdf5 file format → Archive

Access: via URL

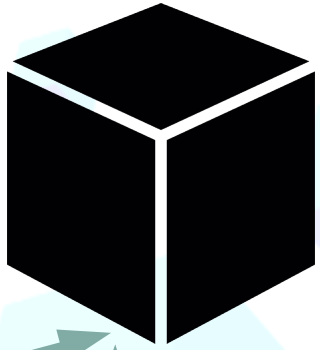
https://analytics-toolkit.nomad-coe.eu/api/resolve/
N-TULHlZnc9cnbg7ihzUALIlSdyww/
C-DWGvyqvK2g_1yLyJf8nN3j_M-xd/
section_run/0c/
section_single_configuration_calculation/0c/
energy_total/0c


```
uri:       "nmd://N-TULHlZnc9cnbg7ihzUALIlSdyww/C-DWGvyqvK2g_1yLyJf8nN3j_M-xd/section_run/0c/section_single_configuration_calculation/0c/energy_total/0c"
type:      "value"
dtypeStr:  "f"
value:     -4.535646058456239e-18
```

Computational MatSci: from basics...

NOMAD
NOVEL MATERIALS DISCOVERY

Atomistic-simulation code



- Code name
- Version
- Libraries

Input structure
- Coordinates
  $Z_i \;\; x_i \, y_i \, z_i$
- Cell vectors
- (Topology)

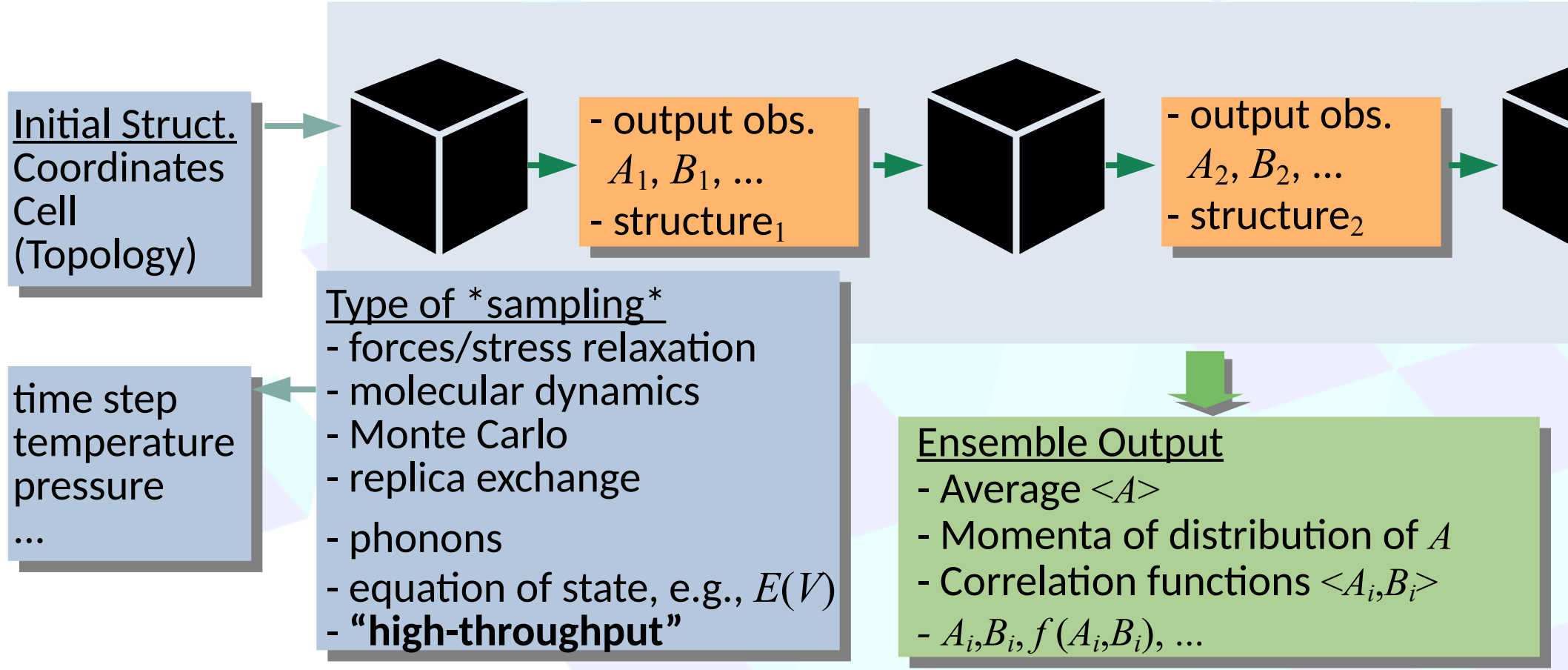
Output
- Total energy
- Forces
- Electron density
- Electrostatic pot.
- El. band structure
- Self energy
- ...

Input model
- xc treatment / force field
- Relativity treatment
- Basis set
- Numerical integr. settings

More complex input model
- Model_1 (scf)→ Model_2 (perturbative)
- QM/MM (space partitioning of models)
- ...

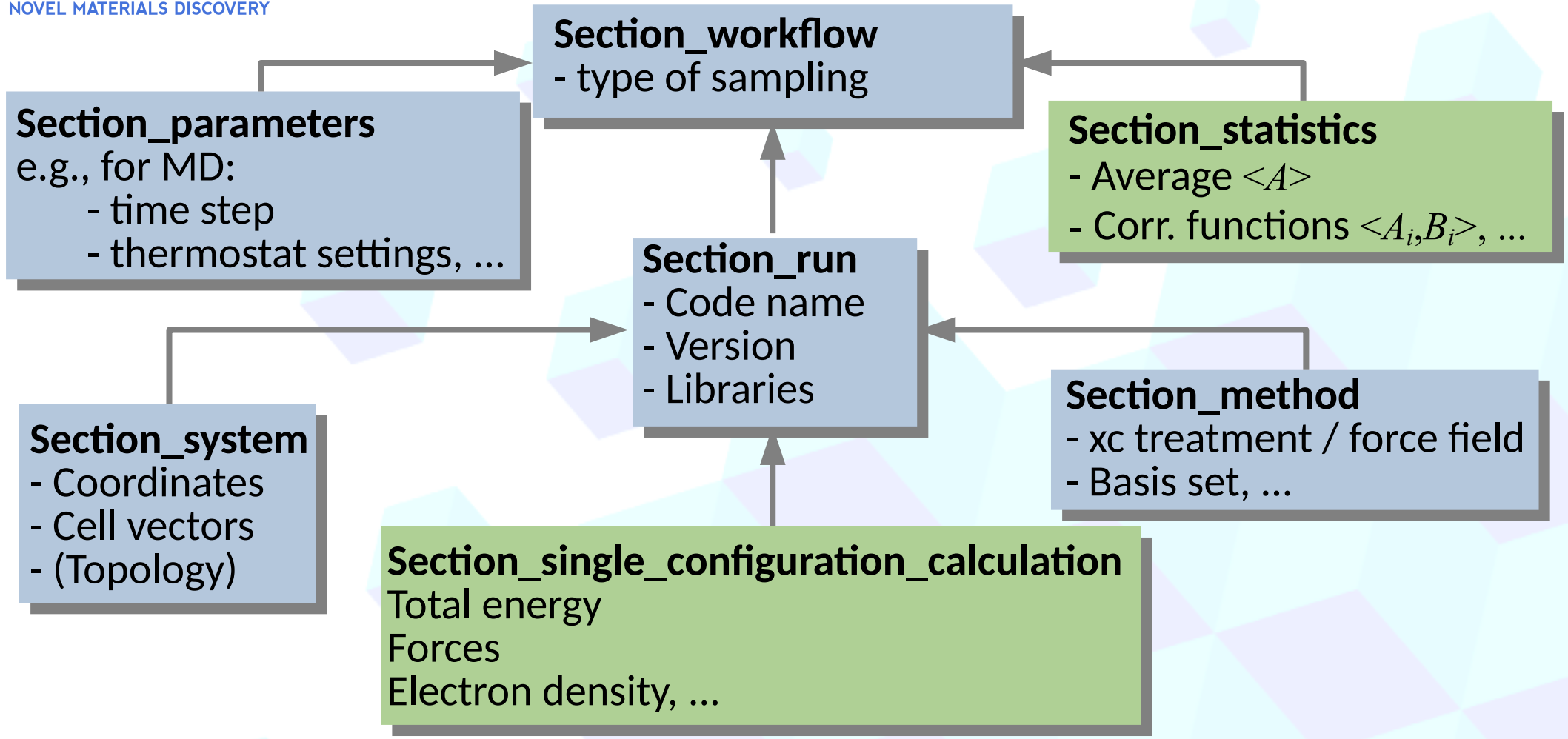
NOMAD
NOVEL MATERIALS DISCOVERY

… to a more complex *workflow* …



… and the related hierarchical scheme

NOVEL MATERIALS DISCOVERY



Keywords

NOVEL MATERIALS DISCOVERY

Metadata for FAIR scientific-data management and stewardship:

- Hierarchical (sections, concrete values, ...)
- Structured (name, description, ...)
- Extensible

Acknowledgments:

Fawzi Mohamed, Pasquale Pavone, Henning Glawe, Micael Olivera, Benjamin Regler, Bryan Goldsmith, Ádám Fekete.
NOMAD PIs: Matthias Scheffler, Claudia Draxl, Angel Rubio, Risto Nieminen, Francesc Illas, Daan Frenkel, Alessandro De Vita, Kristian Thygesen; Kimmo Koski, Stefan Heinzel, Jose Maria Cela, Dieter Kranzlmüller; Ciaran Clissman.

NOMAD continues via FAIR – Data Infrastructure (https://fairdi.eu/)