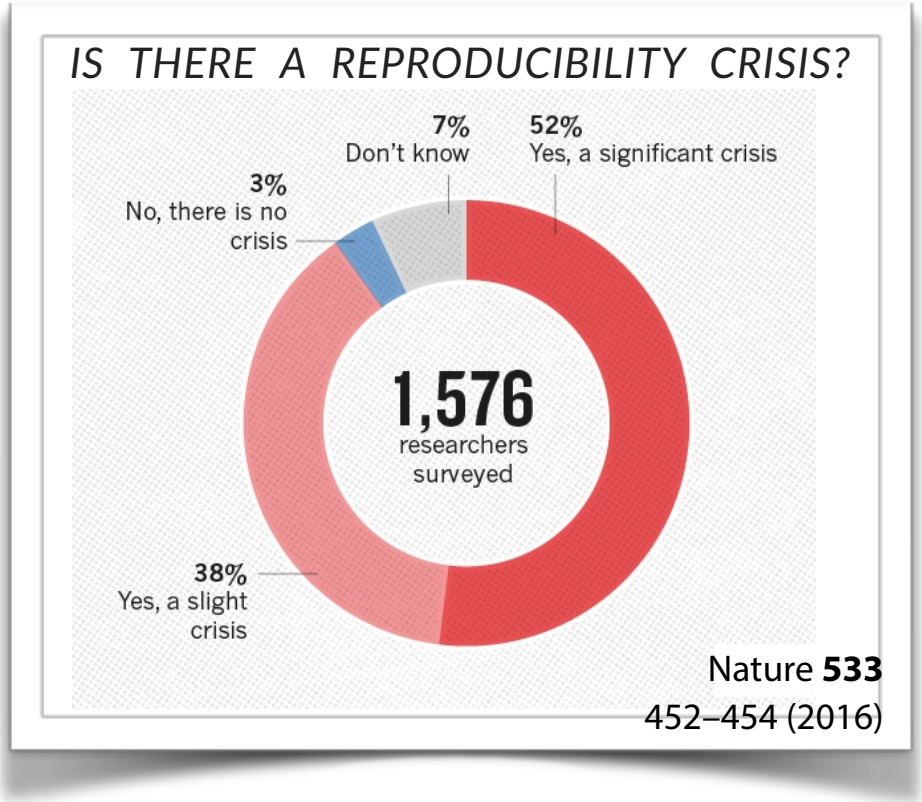
The challenges we address

- Enable **high-throughput** research (10'000+ simulations/day); automate simulations, **automatically track provenance**
- Share simulations **according to the FAIR principles and *beyond*, guaranteeing reproducibility**
- Encode scientists' knowledge in **automated workflows** (scientific know-how, numerical parameters, choice of data to preserve and share)
- Provide advanced **data analytics tools**
- Allow to **create metadata on the simulations a posteriori** using automatically tracked provenance (meta)data

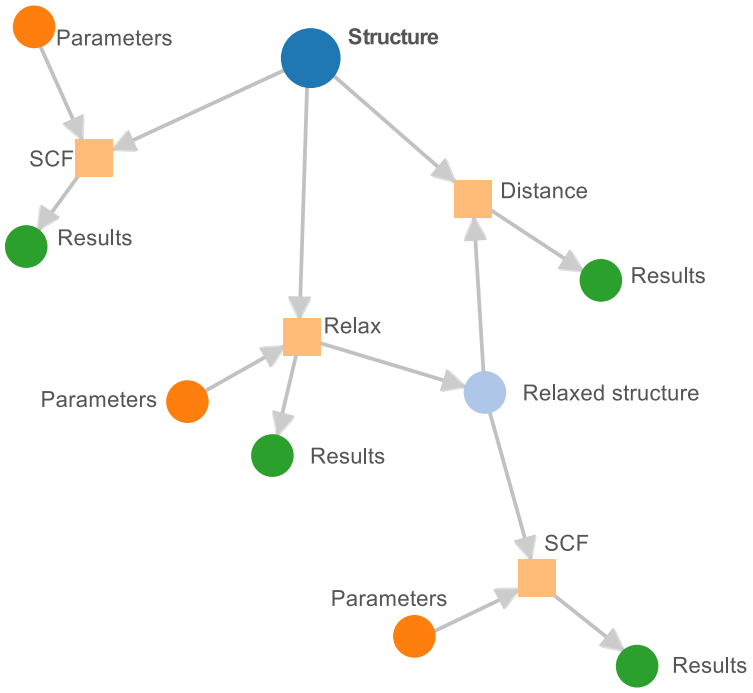
How to manage data, simulations and their provenance?

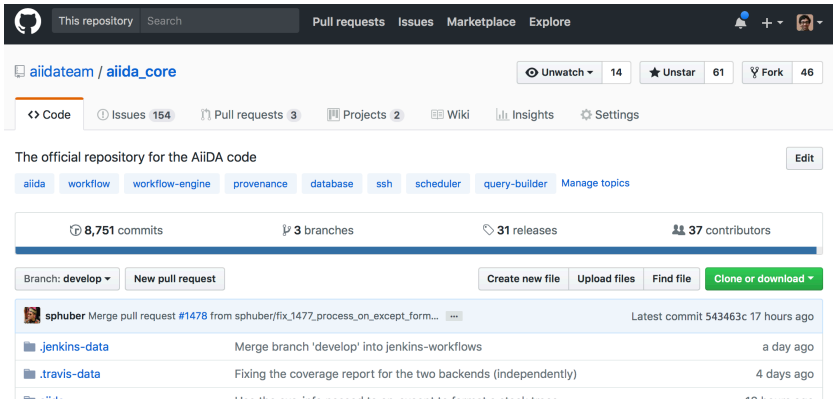


We need a **tool** to help us
automate research, organise it,
store **provenance** guaranteeing **reproducibility**,
then **analyze results**,
and finally **share them** with metadata

Data provenance: Directed Acyclic Graphs

AiiDA





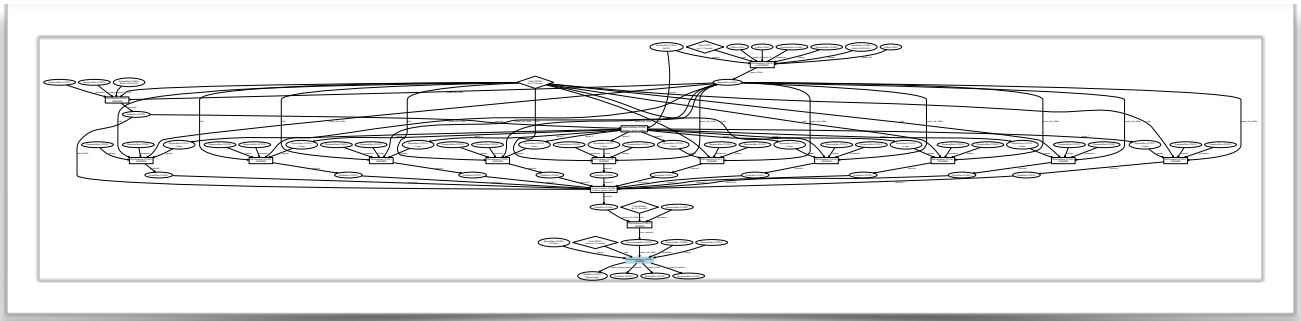
MIT license (open source)

Developed since 2013
Used in production from many scientific research projects

G. Pizzi et al.,
Comp. Mat. Sci. 111, 218-230 (2016)

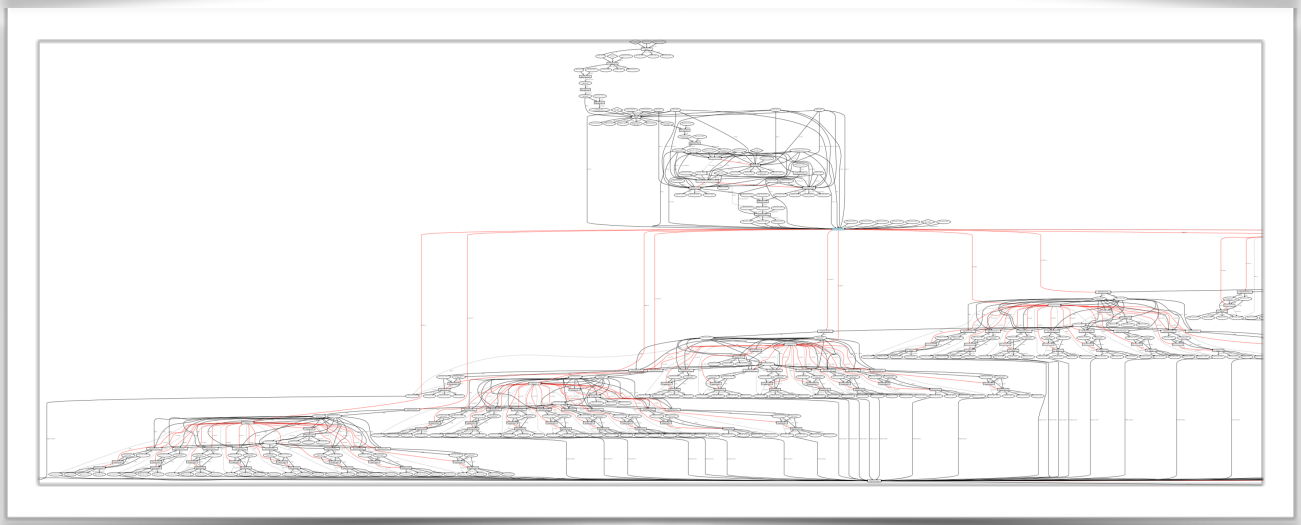
http://www.aiida.net

“Simple” graphs of workflows for a single material



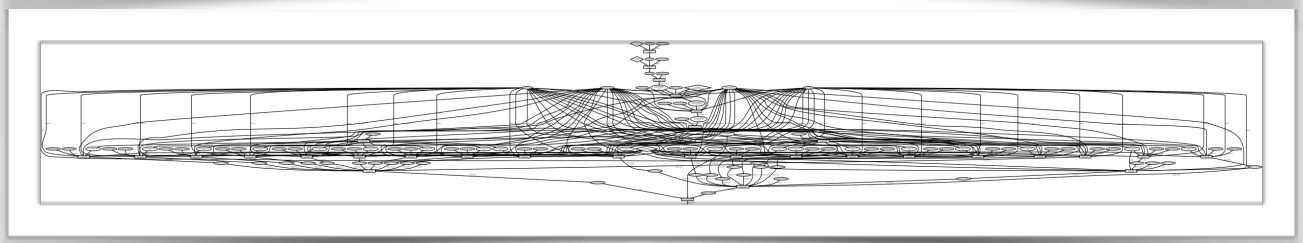
Phonon dispersion

(atom oscillations around equilibrium positions: thermal transport, electronic mobility, …)



Molecular dynamics of Lithium in a solid electrolyte

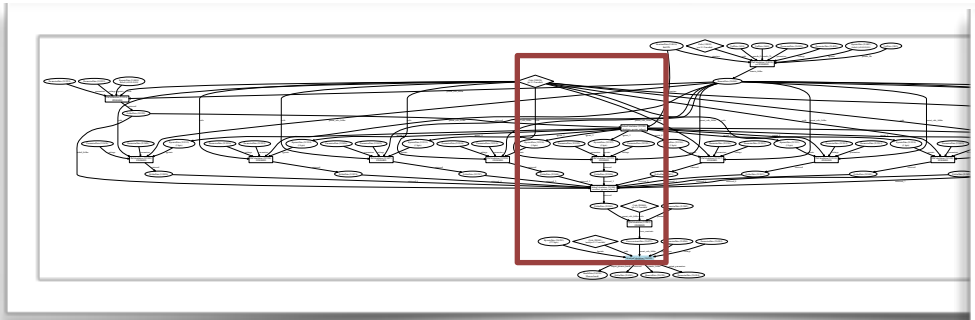
(Discover novel, safe and efficient electrolytes for Li-batteries)

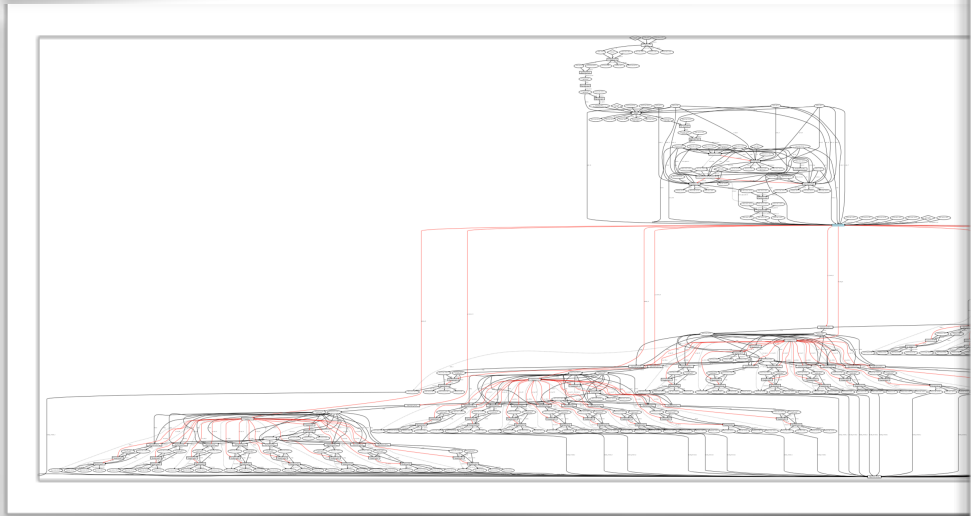


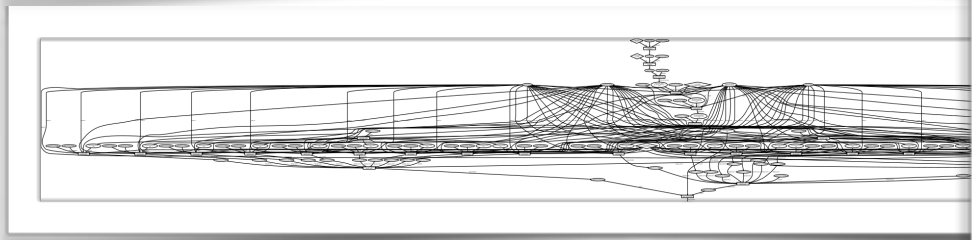
Elastic constants

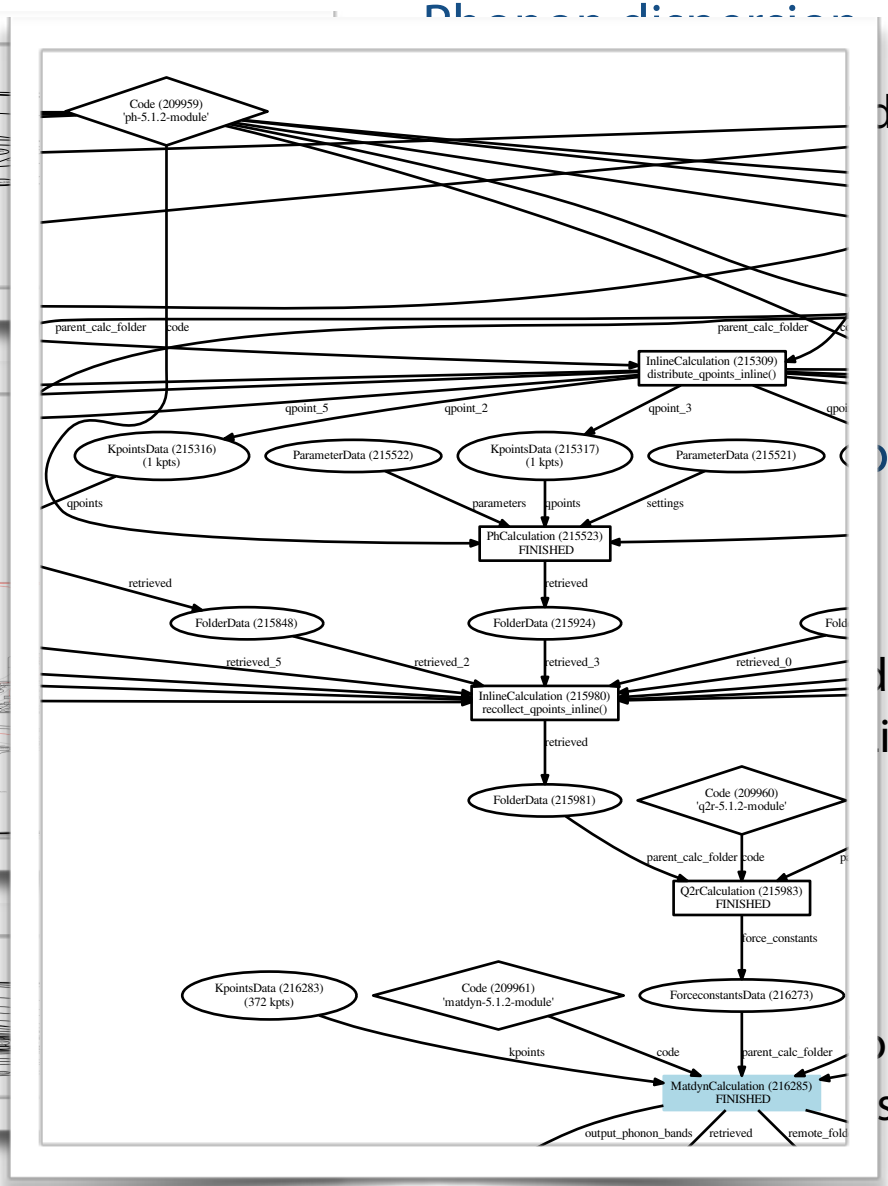
(response of materials to stresses and deformations)

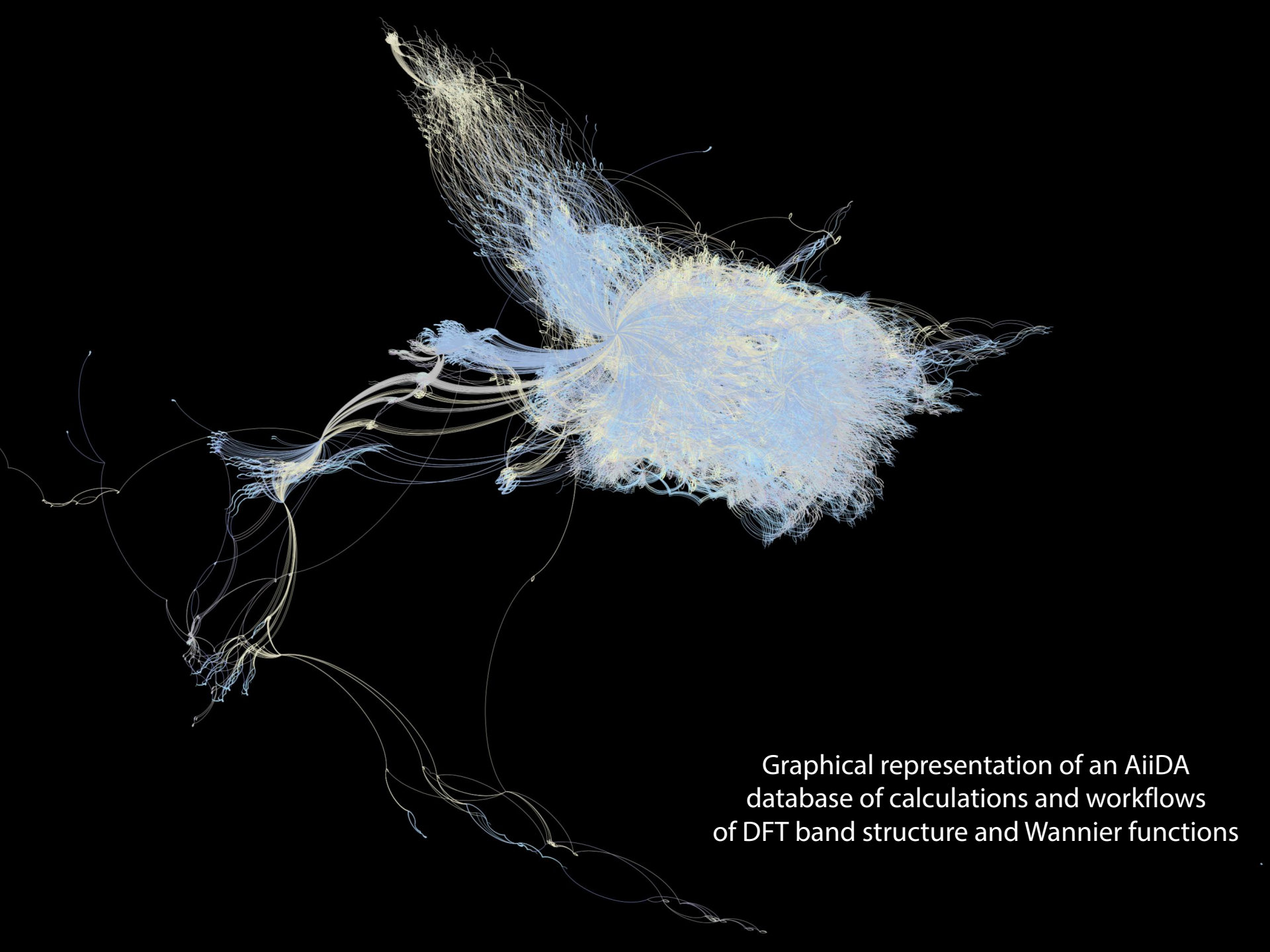
“Simple” graphs of workflows for a single material









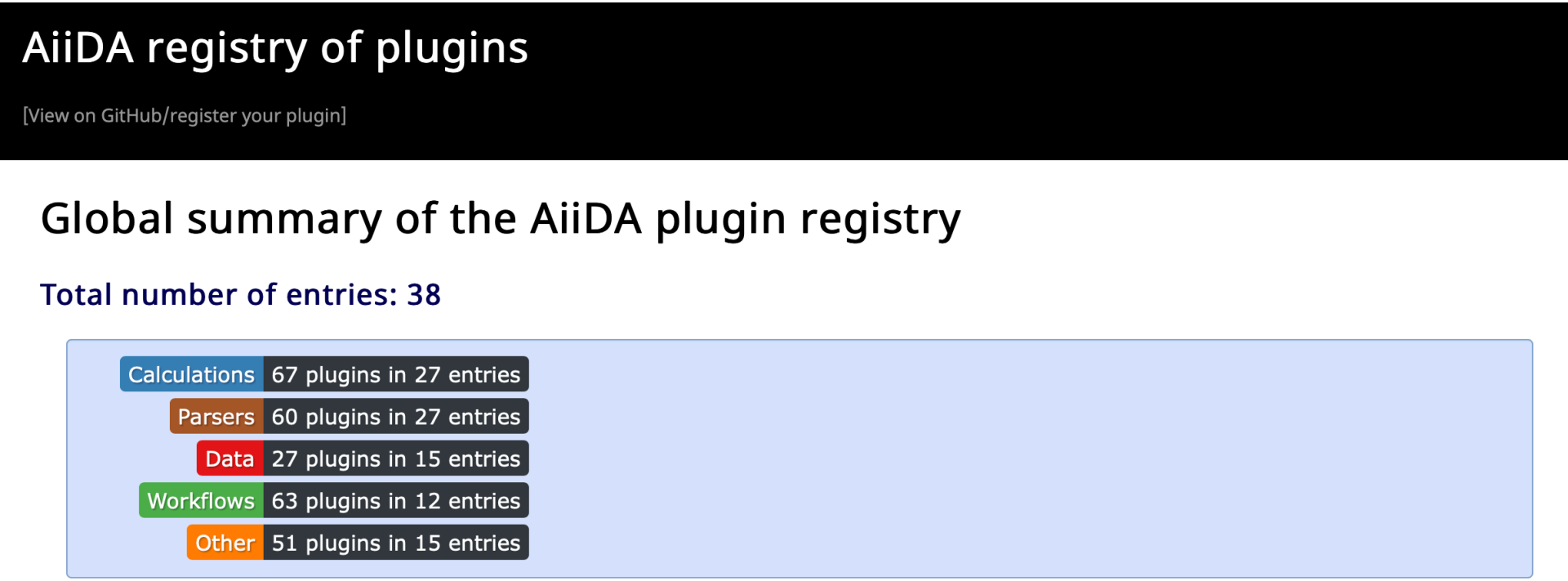


Graphical representation of an AiiDA
database of calculations and workflows
of DFT band structure and Wannier functions

An ecosystem of plugins

https://aiidateam.github.io/aiida-registry/

38 plugin entries for Materials Science,
supporting **67 code executables**, **63 workflows**, …



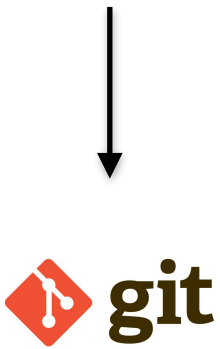
Open Science Platform: AiiDA + Materials Cloud

MATERIALSCLOUD

https://www.materialscloud.org

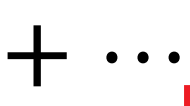
Online since February 2018

Cloud dissemination platform for FAIR data sharing
and **more (cloud simulation and data generation platform)**



MATERIALSCLOUD

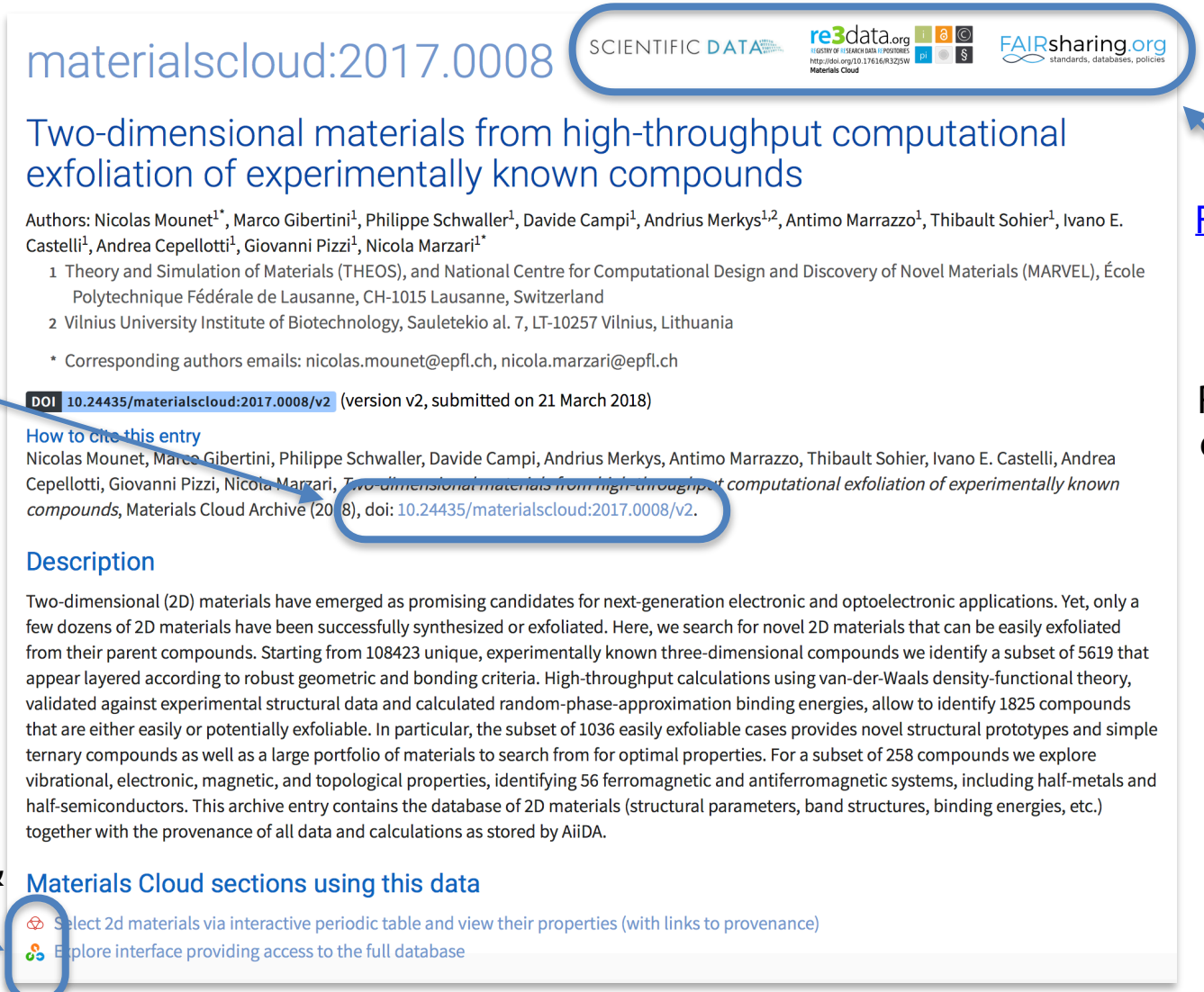
GitHub



swissuniversities

Data and metadata in
AiiDA and Materials Cloud

Open and FAIR data sharing: **Archive, Discover, Explore**



materialscloud:2017.0008

SCIENTIFIC DATA · re3data.org · FAIRsharing.org standards, databases, policies

Two-dimensional materials from high-throughput computational exfoliation of experimentally known compounds

Authors: Nicolas Mounet[1*], Marco Gibertini[1], Philippe Schwaller[1], Davide Campi[1], Andrius Merkys[1,2], Antimo Marrazzo[1], Thibault Sohier[1], Ivano E. Castelli[1], Andrea Cepellotti[1], Giovanni Pizzi[1], Nicola Marzari[1*]

1 Theory and Simulation of Materials (THEOS), and National Centre for Computational Design and Discovery of Novel Materials (MARVEL), École Polytechnique Fédérale de Lausanne, CH-1015 Lausanne, Switzerland

2 Vilnius University Institute of Biotechnology, Sauletekio al. 7, LT-10257 Vilnius, Lithuania

* Corresponding authors emails: nicolas.mounet@epfl.ch, nicola.marzari@epfl.ch

DOI 10.24435/materialscloud:2017.0008/v2 (version v2, submitted on 21 March 2018)

How to cite this entry

Nicolas Mounet, Marco Gibertini, Philippe Schwaller, Davide Campi, Andrius Merkys, Antimo Marrazzo, Thibault Sohier, Ivano E. Castelli, Andrea Cepellotti, Giovanni Pizzi, Nicola Marzari, *Two-dimensional materials from high-throughput computational exfoliation of experimentally known compounds*, Materials Cloud Archive (2018), doi: 10.24435/materialscloud:2017.0008/v2.

Description

Two-dimensional (2D) materials have emerged as promising candidates for next-generation electronic and optoelectronic applications. Yet, only a few dozens of 2D materials have been successfully synthesized or exfoliated. Here, we search for novel 2D materials that can be easily exfoliated from their parent compounds. Starting from 108423 unique, experimentally known three-dimensional compounds we identify a subset of 5619 that appear layered according to robust geometric and bonding criteria. High-throughput calculations using van-der-Waals density-functional theory, validated against experimental structural data and calculated random-phase-approximation binding energies, allow to identify 1825 compounds that are either easily or potentially exfoliable. In particular, the subset of 1036 easily exfoliable cases provides novel structural prototypes and simple ternary compounds as well as a large portfolio of materials to search from for optimal properties. For a subset of 258 compounds we explore vibrational, electronic, magnetic, and topological properties, identifying 56 ferromagnetic and antiferromagnetic systems, including half-metals and half-semiconductors. This archive entry contains the database of 2D materials (structural parameters, band structures, binding energies, etc.) together with the provenance of all data and calculations as stored by AiiDA.

Materials Cloud sections using this data

Select 2d materials via interactive periodic table and view their properties (with links to provenance)

Explore interface providing access to the full database

DOIs assigned

Direct links to Discover & Explore

FAIRsharing.org
re3data.org

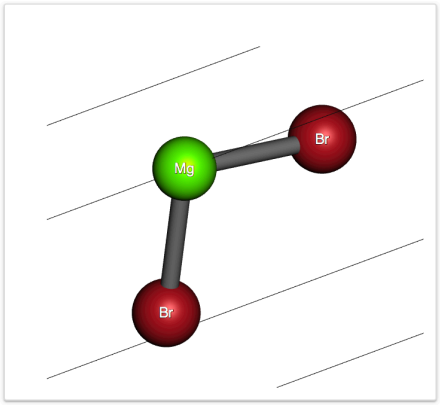
+

Recommended data repository by Nature's journal *Scientific Data*

DISCOVER (CURATED DATA) & EXPLORE (RAW DATA)

DISCOVER

Compound: $MgBr_2$



Info and properties

See definitions...

Formula: $MgBr_2$
Spacegroup: P-3m1
Pointgroup: -3m
Prototype: CdI2
Band gap [eV]: 4.8

Magnetic properties:

Magnetic State: non-magnetic
Tot. Magnetization [μB/cell]: -
Abs. Magnetization [μB/cell]: -

Binding Energies:

DF2-C09 Binding energy [meV/Å^2]: 10.2
(From parent COD 9009107)
rVV10 Binding energy [meV/Å^2]: 15.3
(From parent COD 9009107)

Delta in interlayer distance (vdW vs revPBE):

Δ_{DF2} [%]: 17.1 (From parent COD 9009107)
Δ_{rVV10} [%]: 18.3 (From parent COD 9009107)

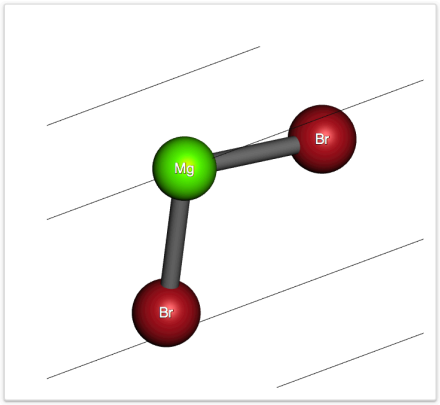
Band structure

UUID links to jump to the provenance graph in the EXPLORE section

DISCOVER (CURATED DATA) & EXPLORE (RAW DATA)

DISCOVER

Compound: $MgBr_2$



Info and properties

See definitions...

Formula: $MgBr_2$
Spacegroup: P-3m1
Pointgroup: -3m
Prototype: CdI2
Band gap [eV]: 4.8

Magnetic properties:

Magnetic State: non-magnetic
Tot. Magnetization [μB/cell]: -
Abs. Magnetization [μB/cell]: -

Binding Energies:

DF2-C09 Binding energy [meV/Å^2]:
(From parent COD 9009107)
rVV10 Binding energy [meV/Å^2]: 15
(From parent COD 9009107)

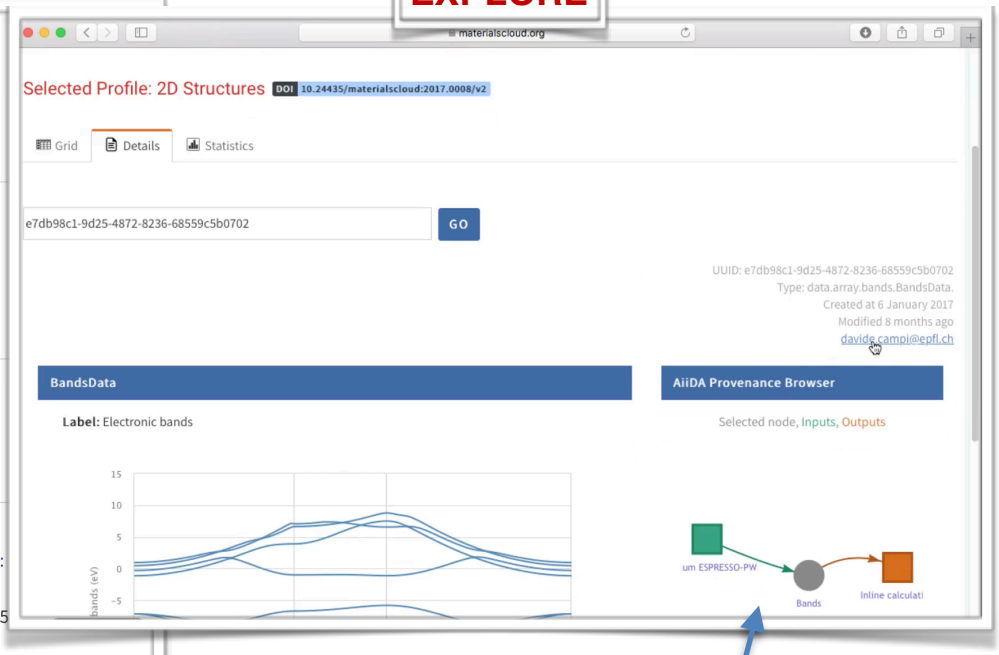
Delta in interlayer distance (vdW vs revPBE):

Δ_{DF2} [%]: 17.1 (From parent COD 9009107)
Δ_{rVV10} [%]: 18.3 (From parent COD 9009107)

Band structure

UUID links to jump to the provenance graph in the EXPLORE section

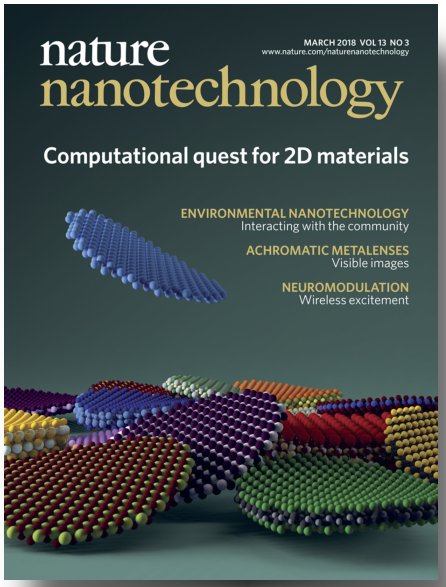
EXPLORE



Browse the full AiiDA provenance graph (inputs, outputs, …) at any level

An example of reuse of open data

High-throughput computational discovery of exfoliable 2D materials

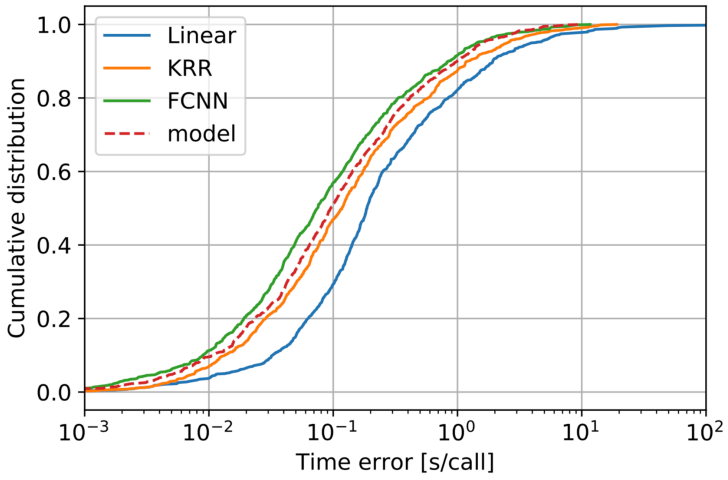


N. Mounet et al.,
Nature Nanotech. 13, 246 (2018)

Data:
Nicolas Mounet et al.,
Materials Cloud Archive (2018),
10.24435/materialscloud:2017.0008/v2

All data published on the *Materials Cloud* with **full provenance** (as stored by AiiDA) and accessible via the AiiDA REST API

Groups at CINECA and University of Bologna: develop a project using the published data, with a **new unforeseen goal:** Prediction of the absolute time per iteration of a Quantum ESPRESSO run



"**Prediction of time-to-solution in material science simulations using deep learning**"
F. Pittino *et al.*, PASC19 Proceedings, 10 (2019)

Need very detailed information ("identikeep") on machines, calculations and provenance

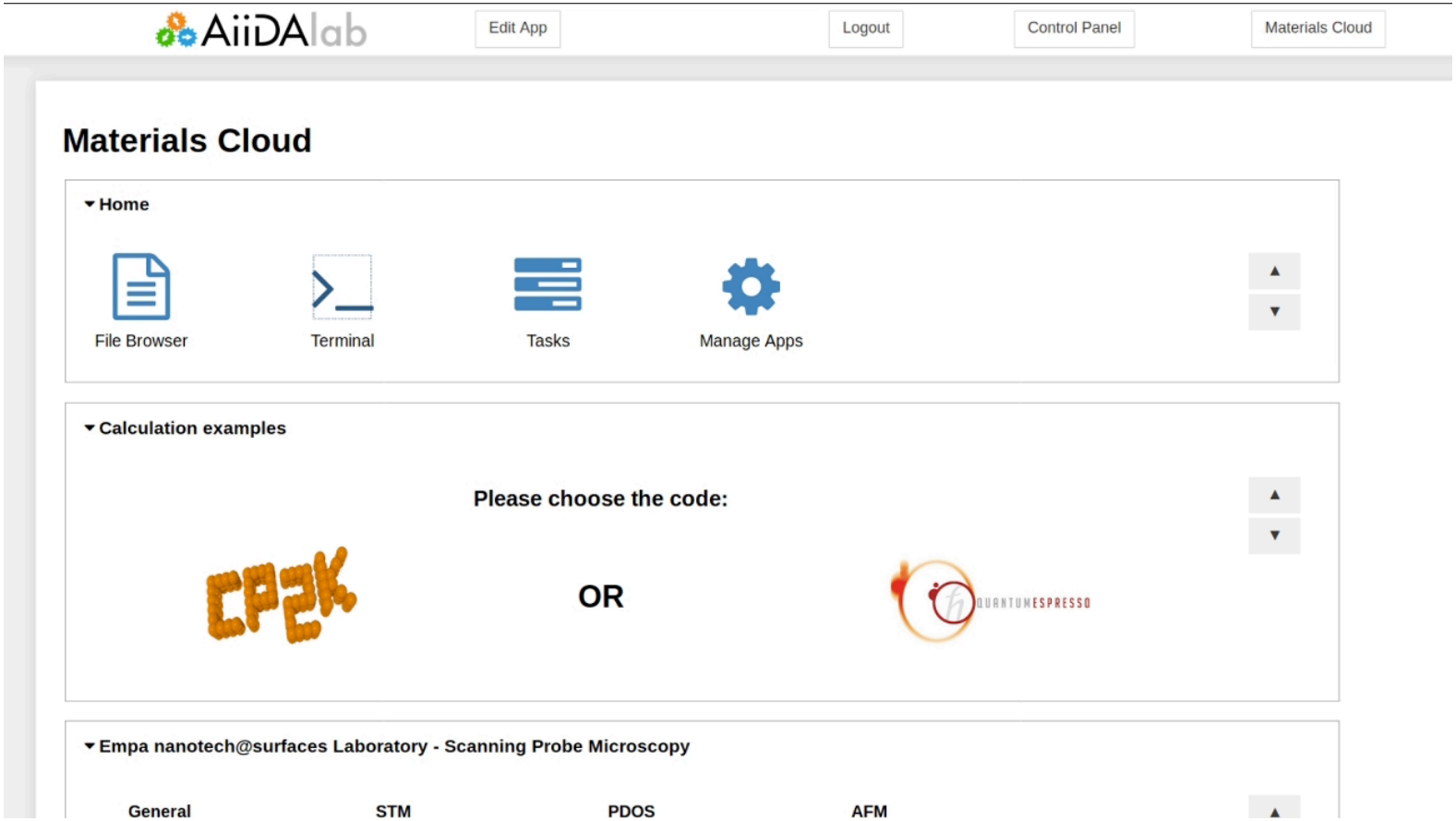
MAX DRIVING THE EXASCALE TRANSITION
MARVEL NATIONAL CENTRE OF COMPETENCE IN RESEARCH

Data and metadata in AiiDA and Materials Cloud

swissuniversities EPFL

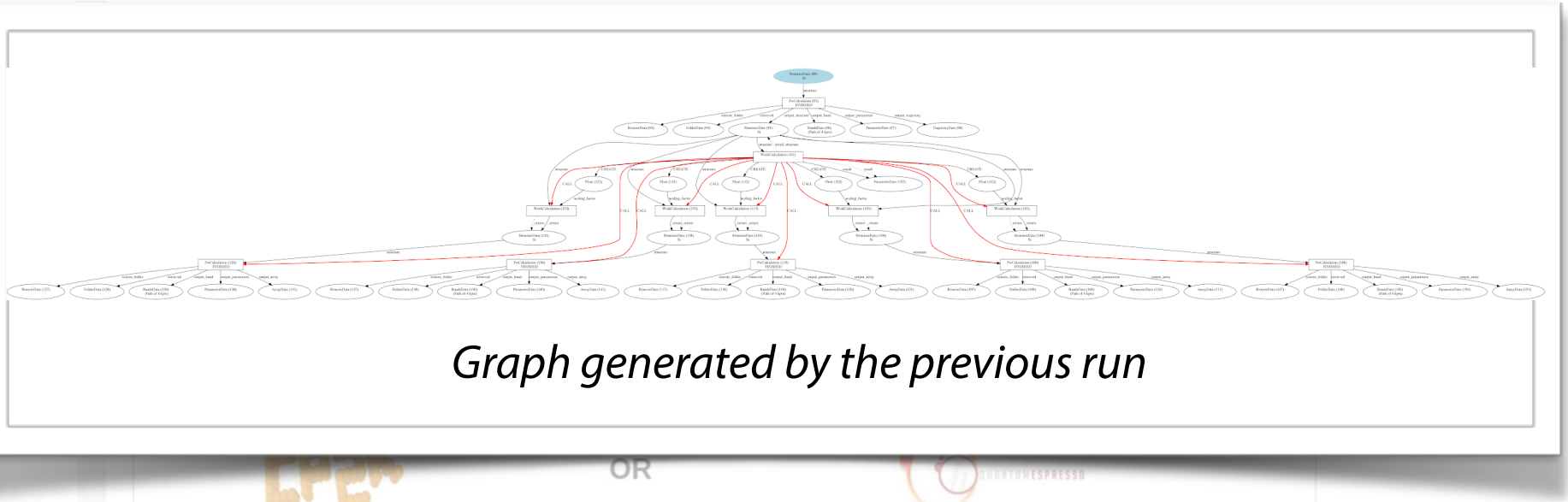
WORK: AiiDA Lab (submission)

- Our **cloud data generation platform** and **data analysis platform**
- Based on AiiDA + Jupyter + App Mode



WORK: AiiDA Lab (submission)

- Our **cloud data generation platform** and **data analysis platform**
- Based on AiiDA + Jupyter + App Mode



Graph generated by the previous run

Data and metadata issues for workflow provenance tracking

Our approach

Coupling automation and storage

```
code = load_code('pw-6.3@daint-mr25')
builder = code.get_builder()

builder.metadata.options = {
    'max_wallclock_seconds': 600,
    'resources': {"num_machines": 2}}

Structure = DataFactory('structure')
structure = Structure(ase = read('TiO2.cif'))

Dict = DataFactory('dict')
parameters = Dict(dict={
    'CONTROL': {
        'calculation': 'scf',
        'restart_mode': 'from_scratch'},
    'SYSTEM': {'ecutwfc': 40.}})

Kpoints = DataFactory('array.kpoints')
kpoints = Kpoints(kpoints_mesh = [4,4,4])

builder.structure = structure
builder.parameters = parameters
builder.kpoints = kpoints
builder.pseudos = get_pseudos_from_family(structure,
'SSSP_efficiency_v1.0')

aiida.engine.submit(builder)
```

Switch computers in one line
supports different schedulers, version of codes, …

Define (only) necessary inputs
Interface designed by plugin

Inputs stored in the DB, and handing over to the daemon

Coupling automation and storage

```
code = load_code('pw-6.3@daint-mr25')
builder = code.get_builder()

builder.metadata.options = {
    'max_wallclock_seconds': 600,
    'resources': {"num
```

```
Structure = DataFacto
structure = Structure

Dict = DataFactory('d
parameters = Dict(dic
    'CONTROL': {
        'calculation'
        'restart_mode
    'SYSTEM': {'ecutw

Kpoints = DataFactory
kpoints = Kpoints(kpo
```

```
builder.structure = s
builder.parameters = p
builder.kpoints = kpoints
builder.pseudos = get_pseudos_from_family(structure,
'SSSP_efficiency_v1.0')
```

```
aiida.engine.submit(builder)
```

All inputs stored before submission

The daemon creates and runs the simulations without human intervention: all inputs are known, machine-readable and encoded in a standard format

Guarantee of reproducibility of single calculations

Switch computers in one line
supports different schedulers, version of codes, …

nly) necessary inputs
designed by plugin

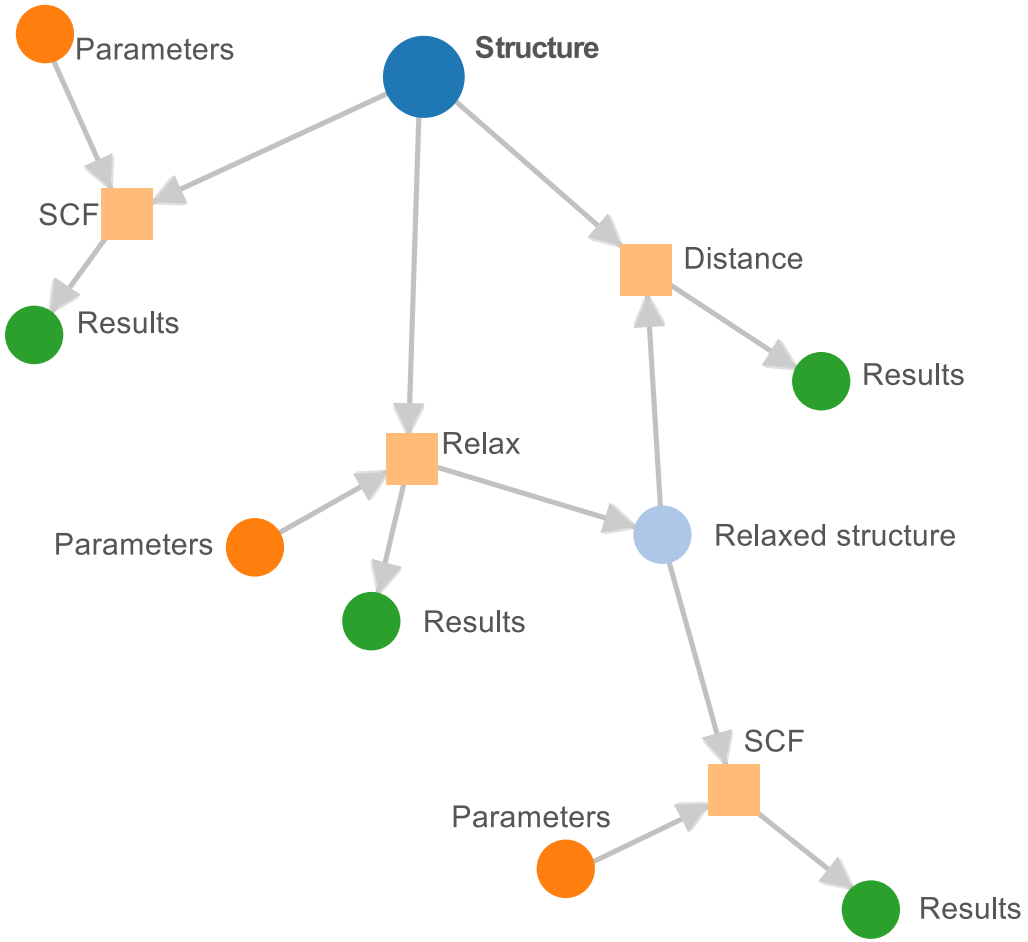
Inputs stored in the DB, and handing over to the daemon

Data formats

- Data stored in AiiDA via custom data classes (via plugins)
  - *Examples*: a crystal structure, a pseudopotential, a grid/list of kpoints, a dictionary of input parameters, a trajectory, …
- Allows **R**euse (the **R** in **FAIR**) of data between different simulations, codes, workflows
- Data formats, and input data formats, are defined and documented by the plugins

Multiple simulations and workflows

- Even just running multiple simulations preserves the connected structure and the provenance



An example of a submission (workflow)

- Encoding the knowledge of the scientist also beyond the original simulation

```python
class BandStructureWorkChain(WorkChain):
    @classmethod
    def define(cls, spec):
        super(BandStructureWorkChain, cls).define(spec)
        spec.input('code', valid_type=orm.Code)
        spec.input('structure', valid_type=orm.StructureData)
        spec.input('protocol', valid_type=orm.Dict)
        spec.input('scf_options', valid_type=orm.Dict)
        spec.outline(
            cls.setup_protocol,
            cls.setup_kpoints,
            cls.setup_parameters,
            cls.run_bands,
            cls.run_results,
        )
    # [...]
    def run_bands(self):
        # [...]
        bands_inputs = get_common_inputs()
        inputs = {
            'structure': self.inputs.structure,
            'bands': bands_inputs,
        }

        future = self.submit(PwBandsWorkChain, **inputs)
        return ToContext(workchain_bands=future)
```

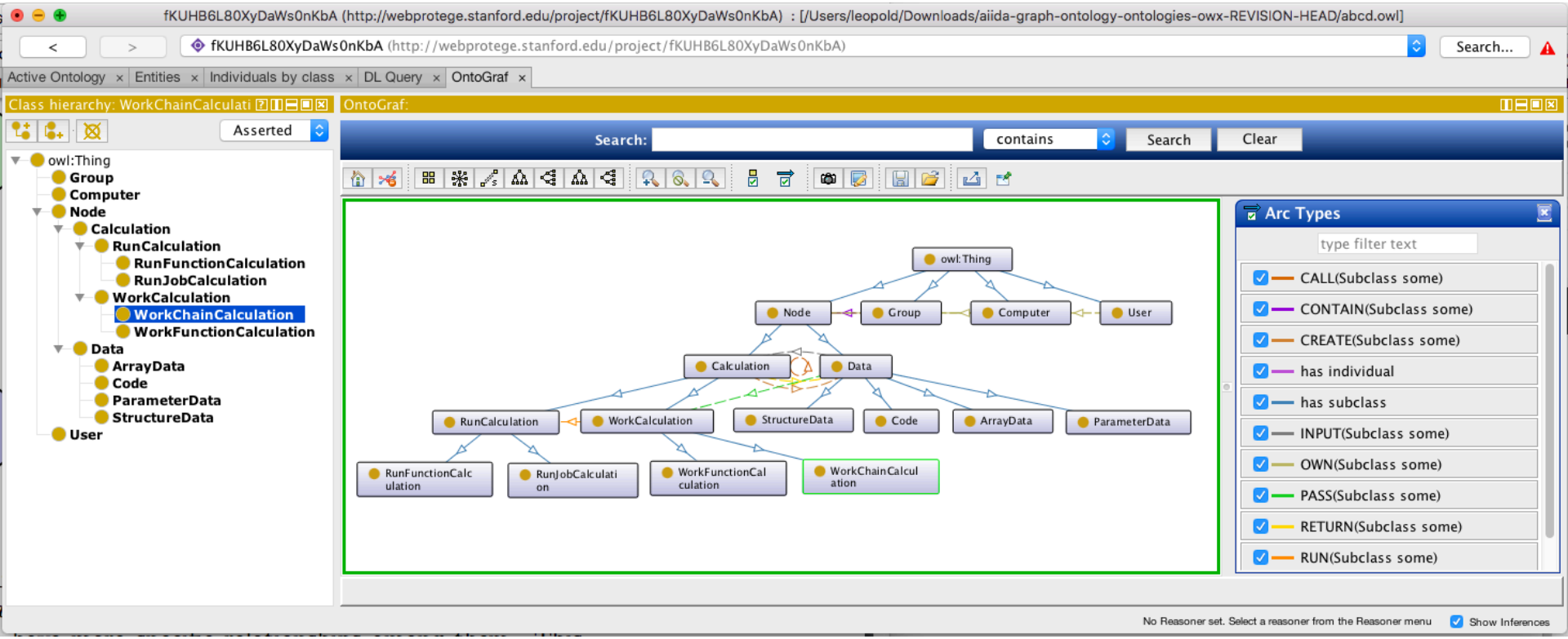
Self-documenting code:
define at the beginning the allowed inputs, outputs and the *outline*

Define the actual steps
creating dynamically the inputs and analysing the parsed outputs

Submit the calculation and make it available to the next steps (that are run upon completion)

OWL description for the provenance graph

- We are working on a OWL description of the provenance graph



Coupling of automation and storage: *a posteriori* metadata

Run first, collect and parse later

- Run all simulations **first**
- **Afterwards**, parse and reconstruct run information, metadata, provenance from calculation outputs

Pros:

- No need to use specific provenance tools
- Can reuse already-run simulations

Cons:

- Hard to guarantee that all inputs are described and correct
- Hard to reconstruct and represent data provenance and workflows

Coupling of automation and storage: *a posteriori* metadata

- Use a tool (e.g. AiiDA) to manage simulations and keep track of data and logic provenance **at the same time**
- Create metadata from automatically tracked provenance information

Pros:

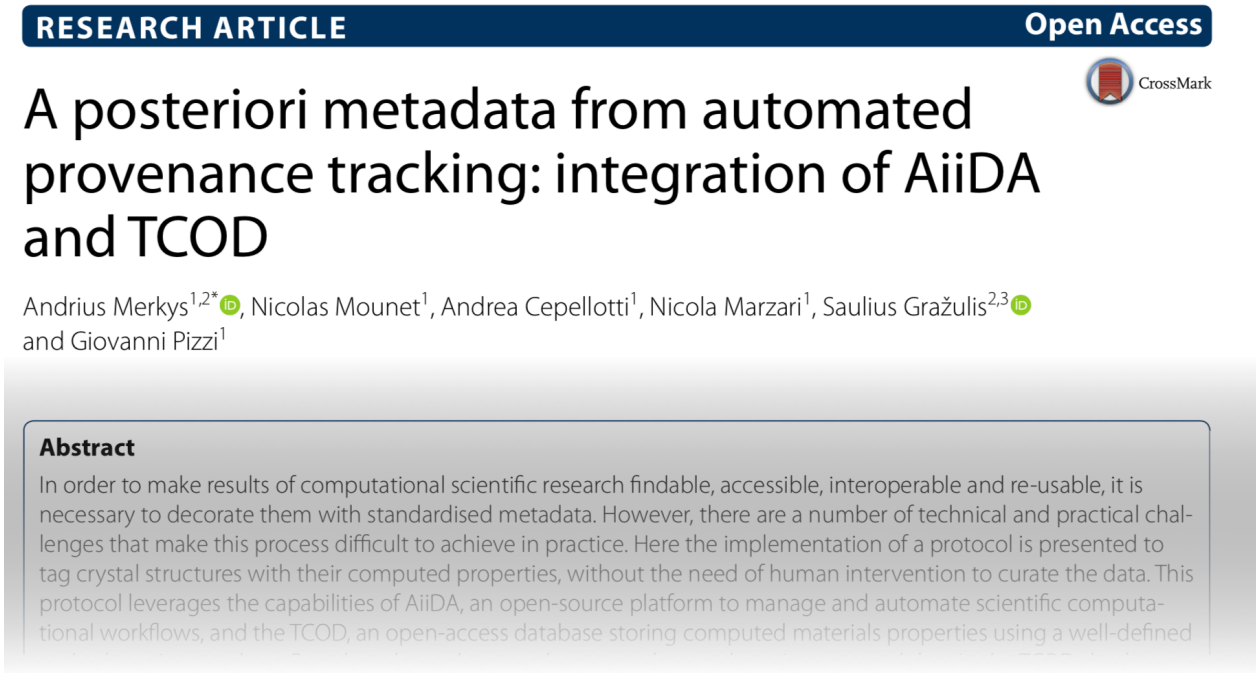
- Automatic guarantee of single calculation reproducibility
- Automatic description of the provenance and reproducibility of entire workflows
- While not the main goal, it is possible to import already-run simulations

Cons:

- Need to use a tool to track provenance (that, however, helps with automation)

A posteriori metadata: AiiDA+TCOD integration

- Coupling to external DBs: implemented via a *plugin interface*
- One current implementation: **creation of TCOD metadata** and deposition into TCOD (Theoretical Crystallography Open Database)
- TCOD: simulation counterpart to the COD (www.crystallography.net)
- TCOD defines its own ontology and dictionaries, in standard CIF format



RESEARCH ARTICLE Open Access

A posteriori metadata from automated provenance tracking: integration of AiiDA and TCOD

Andrius Merkys[1,2*], Nicolas Mounet[1], Andrea Cepellotti[1], Nicola Marzari[1], Saulius Gražulis[2,3] and Giovanni Pizzi[1]

Abstract

In order to make results of computational scientific research findable, accessible, interoperable and re-usable, it is necessary to decorate them with standardised metadata. However, there are a number of technical and practical challenges that make this process difficult to achieve in practice. Here the implementation of a protocol is presented to tag crystal structures with their computed properties, without the need of human intervention to curate the data. This protocol leverages the capabilities of AiiDA, an open-source platform to manage and automate scientific computational workflows, and the TCOD, an open-access database storing computed materials properties using a well-defined

Merkys *et al.*, J Cheminform. *9, 56 (2017),* doi:10.1186/s13321-017-0242-y

A posteriori metadata: AiiDA+TCOD integration

Metadata as defined in the CIF dictionaries

- http://www.crystallography.net/tcod/cif/dictionaries/cif_dft.dic
- http://www.crystallography.net/tcod/cif/dictionaries/cif_tcod.dic

```
data_dft_bulk_modulus
    _name '_dft_bulk_modulus'
    _category dft_calc_property
    _type numb
    _units GPa
    _units_detail gigapascals
    _definition
; Ratio of the infinitesimal pressure increase to
; the resulting relative decrease of the volume.
;

data_tcod_software_package
    _name '_tcod_software_package'
    _category tcod_software
    _type char
    _definition
; Software package used to compute and produce
; the DFT-computed structure file. Only package
; or program name should be used, e.g. 'VASP',
; 'psi3', 'Abinit', etc.
;
```

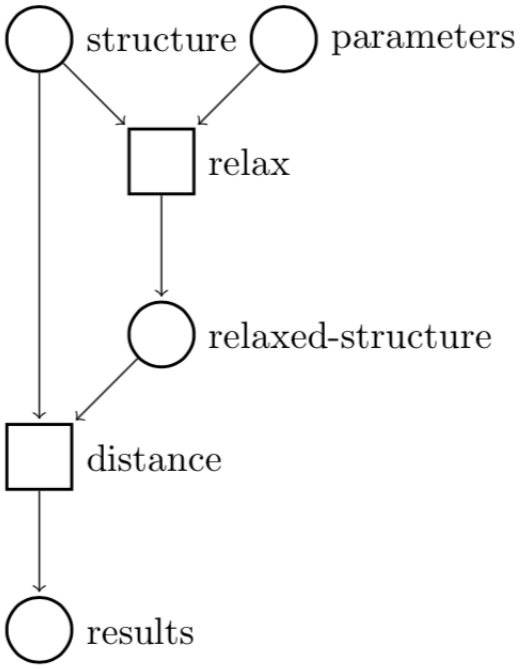
Data defined for each entry in CIF format

```
_dft_bulk_modulus                   5.95
_dft_cell_energy_conv               0.00000001
_dft_BZ_integration_method          Monkhorst-Pack
_dft_kinetic_energy_cutoff_wavefunctions 500
_dft_pseudopotential_type           PAW
_dft_XC_functional_type             GGA
_tcod_database_code                 20000419
_tcod_model                         DFT
_tcod_software_package              VASP
```

Merkys *et al.*, J Cheminform. *9, 56 (2017),* doi:10.1186/s13321-017-0242-y

A posteriori metadata: AiiDA+TCOD integration

- Metadata and data defined also to store the AiiDA graph provenance in CIF format




```
loop_
    _tcod_computation_step
    _tcod_computation_command
1 'relax struct.cif -p param.dat > relaxed.cif'
2 'dist -1 struct.cif -2 relaxed.cif > results.log'

loop_
    _tcod_file_id
    _tcod_file_name
structure           struct.cif
parameters          param.dat
relaxed-structure   relaxed.cif
results             results.log
```

+ JSON-serialized graph

Merkys *et al.*, J Cheminform. *9, 56 (2017),* doi:10.1186/s13321-017-0242-y

A posteriori metadata: open to any metadata format

- The AiiDA+TCOD implementation shows coupling with a CIF-format-based ontology
- In the *a posteriori* metadata approach, **any ontology is allowed**

Table 1 Comparison of a selection of TCOD CIF data items with respect to the corresponding ETSF variables

ETSF variable	TCOD CIF data item(s)	comments
valence_charges	_dft_atom_type_valence_electrons	
pseudopotential_types	_dft_pseudopotential_type	
basis_set	_dft_basisset_type	
exchange_functional	_dft_XC_exchange_functional	
correlation_functional	_dft_XC_correlation_functional	
fermi_energy (E_h)	_dft_fermi_energy (eV)	
smearing_scheme	_dft_BZ_integration_smearing_method _dft_BZ_integration_MP_order	ETSF appends M-P order to the scheme, TCOD CIF has a separate data item
smearing_width	_dft_BZ_integration_smearing_width	
kinetic_energy_cutoff (E_h)	_dft_kinetic_energy_cutoff_wavefunctions (eV)	in ETSF it is not clear whether the variable applies to wavefunctions or charge densities
kpoint_grid_shift	_dft_BZ_integration_grid_shift_[XYZ]	
primitive_vectors (a_0)	_cell_length_[abc] (Å) _cell_angle_[alpha,beta,gamma]	
reduced_symmetry_matrices reduced_symmetry_translations	_space_group_symop_operation_xyz	ETSF provides matrices, TCOD CIF uses string notation
space_group	_space_group_IT_number	CIF has 230 spacegroups, ETSF allows for a range from 1 to 232
reduced_atom_positions	_atom_site_fract_[xyz]	
atom_species	_atom_site_type_symbol	
atom_species_names atomic_numbers chemical_symbols	_atom_site_type_symbol	
reduced_coordinates_of_kpoints	_dft_BZ_integration_grid_IBZ_point_[XYZ]	
kpoint_weights	_dft_BZ_integration_grid_IBZ_point_weight	

Comparison of TCOD CIF dictionary with ETSF variables

Table 2 Comparison of a selection of TCOD CIF data items with respect to the corresponding NOMAD metadata

NOMAD metadata	TCOD CIF data item(s)	comments
atom_labels	_atom_site_label	
atom_positions	_tcod_atom_site_initial_fract_[xyz]	NOMAD uses Cartesian coordinates, TCOD uses fractional coordinates
basis_set_planewave_cutoff (J)	_dft_kinetic_energy_cutoff_wavefunctions (eV)	
configuration_periodic_dimensions	_dft_cell_periodic_BC_[XYZ]	
simulation_cell	_cell_length_[abc] _cell_angle_[alpha,beta,gamma]	NOMAD provides vectors
program_compilation_datetime	_tcod_software_package_compilation_timestamp	NOMAD in Unix timestamp, TCOD CIF in ISO 8601
program_name	_tcod_software_package	
program_version	_tcod_software_package_version	
atom_forces (N)	_tcod_atom_site_resid_force_Cartn_[xyz] (eV/Å)	
energy_total (J/atom)	_tcod_total_energy (eV)	
source_references	_tcod_source_*	NOMAD seems to give a free-text field to identify the source of data
time_calculation	_tcod_computation_wallclock_time	

Comparison of TCOD CIF dictionary with NOMAD metadata

Merkys *et al.*, J Cheminform. *9, 56 (2017),* doi:10.1186/s13321-017-0242-y

AiiDA & Materials Cloud Teams

Marco Borelli (EPFL)

Valeria Granata (EPFL)

Sebastiaan P. Huber (EPFL)

Leonid Kahle (EPFL)

Boris Kozinsky (BOSCH)

Snehal P. Kumbhar (EPFL)

Nicola Marzari (EPFL)

Elsa Passaro (EPFL)

Giovanni Pizzi (EPFL)

Thomas Schulthess (ETHZ,CSCS)

Berend Smit (EPFL)

Leopold Talirz (EPFL)

Daniele Tomerini (EPFL)

Joost VandeVondele (ETHZ,CSCS)

Casper Welzel (EPFL)

Aliaksandr Yakutovich (EPFL)

Contributors for the 35+ plugins: Quantum ESPRESSO, Wannier90, CP2K, FLEUR, YAMBO, SIESTA, VASP, …

Contributors to aiida_core and former AiiDA team members — Valentin Bersier, Jocelyn Boullier, Jens Broeder, Andrea Cepellotti, Fernando Gargiulo, Dominik Gresch, Conrad Johnston, Rico Häuselmann, Eric Hontz, Christoph Koch, Espen Flage-Larsen, Antimo Marrazzo, Andrius Merkys, Nicolas Mounet, Tiziano Müller, Riccardo Sabatini, Ole Schütt, Phillippe Schwaller, Andreas Stamminger, Martin Uhrin, Spyros Zoupanos

The CSCS support teams

Acknowledgements and funding

H2020 Centre of Excellence "MaX"

Scaling towards exascale machines and high-throughput efficiency

SNSF NCCR "MARVEL"

Discovery of new materials via simulations and dissemination of curated data

Swissuniversities P-5 "Materials Cloud"

Scaling the web platform, extending to more disciplines

Moreover:

H2020 Marketplace

Providing data and simulation services in a EU Marketplace platform for industry

H2020 Intersect

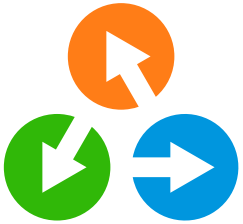
Develop AiiDA workflows to compute transport properties of materials

Conclusions

+

- **Open Science Platform** for computational materials science
- **Guarantee reproducibility** + **FAIR sharing of data** with their provenance
- Define and run **scientific workflows** for materials and provide **data analytics tools**
- Allow to create **metadata a posteriori** without loss of information

Contacts



Website: http://www.aiida.net

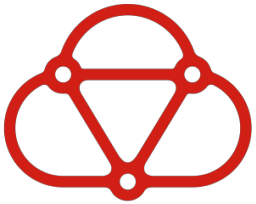
Docs: http://aiida-core.readthedocs.io

Git repo: https://github.com/aiidateam/aiida_core/

Plugin registry: http://aiidateam.github.io/aiida-registry

https://www.facebook.com/aiidateam

@aiidateam



Materials Cloud: http://www.materialscloud.org

- **AiiDA lab**: http://aiidalab.materialscloud.org
- **Archive**: http://archive.materialscloud.org

Quantum Mobile: http://www.materialscloud.org/work/quantum-mobile