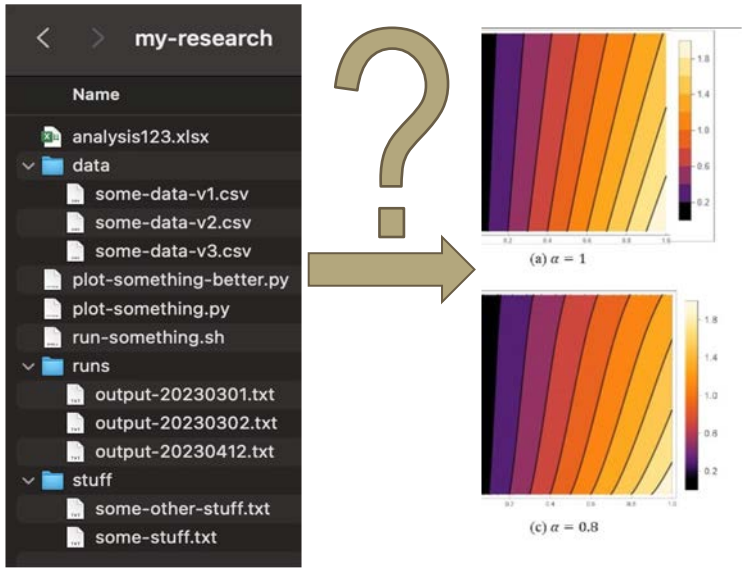
Towards fully reproducible workflows

AiiDA informatics infrastructure

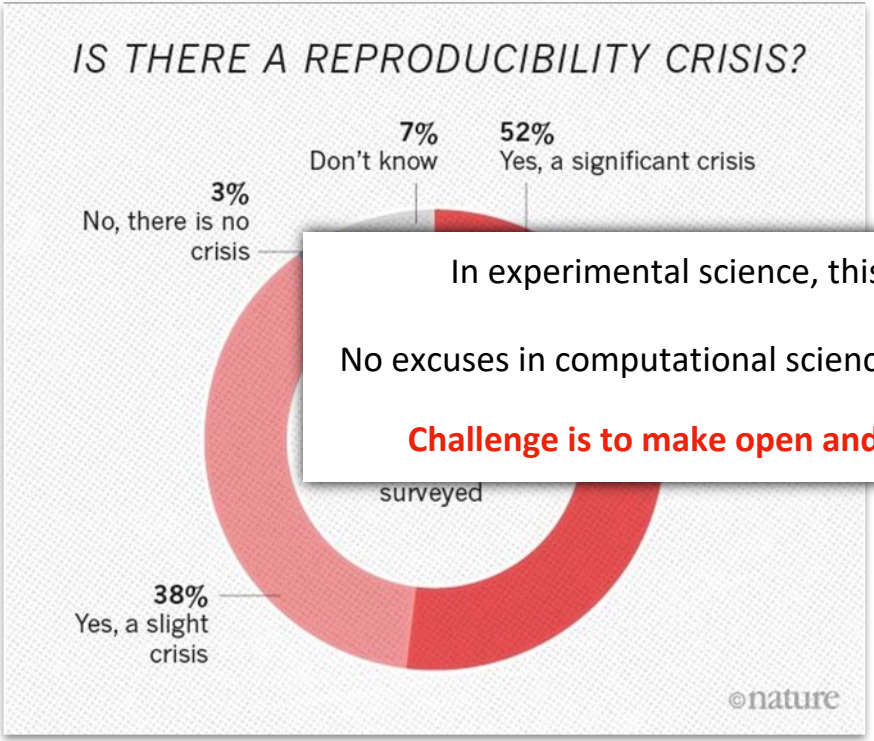
Chris Sewell & Marnik Bercx

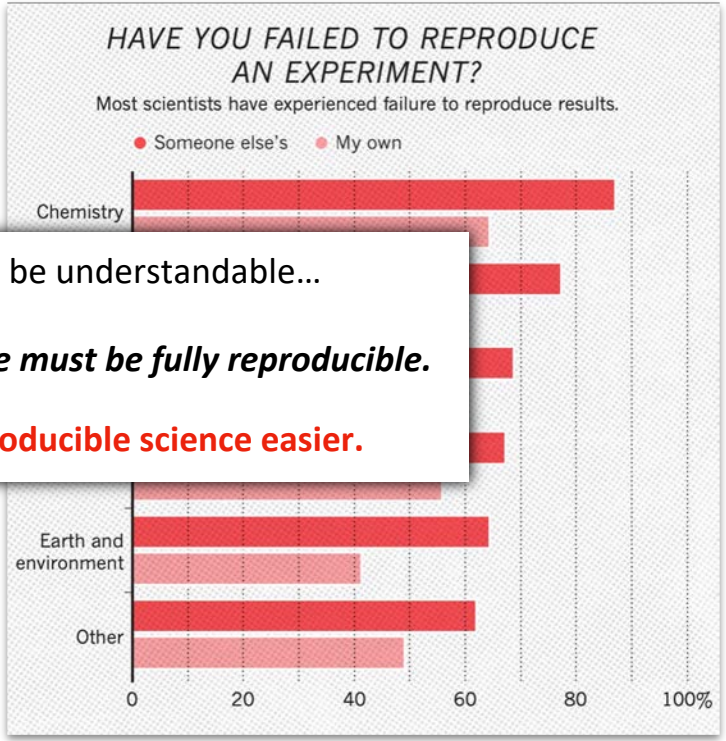
NCCR Ventures - Open Research Data
8 May 2023 - Lausanne

Have you ever been here?



Reproducibility Crisis





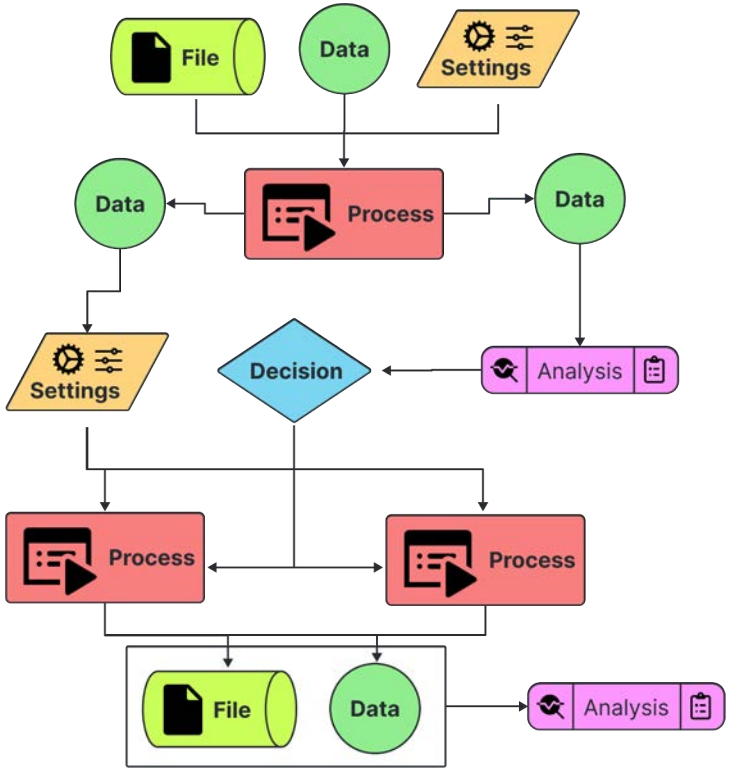
In experimental science, this may be understandable…

No excuses in computational science. ***We must be fully reproducible.***

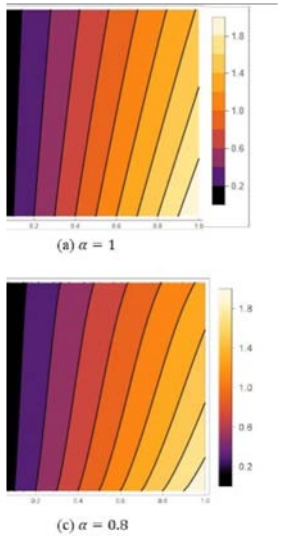
Challenge is to make open and reproducible science easier.

Monya B., *Nature* **533**, 452–454 (2016)

Levels of reproducibility



1. Basic: reproduce analysis from final data.
2. Full reproduction of the whole procedure that created the data.



"Methods section"

Thermodynamic Stability of Sn substitution

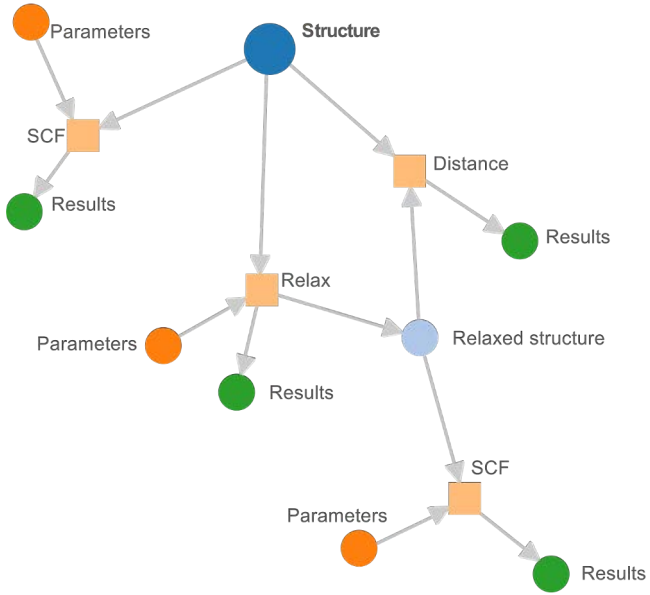
The formation energies of Sn-substituted structures for a range of x-values have been calculated within the Density Functional Theory [1, 2] (DFT) framework, as implemented in the Vienna Ab initio Simulation Package [3–5] (VASP). The projector augmented wave [6, 7] (PAW) method was used to make a distinction between the core and valence electrons, with the standard VASP recommended choice for the number of valence electrons. The exchange-correlation energy was calculated using the SCAN+rVV10 [23, 24] functional to include the van der Waals interaction, which is especially important for a layered structure such as SnO [25]. The wave functions of the valence electrons are expanded in a plane wave basis set, using a high energy cutoff equal to 500 eV, which is advisable for structures containing oxygen. For all 2×2×2 supercell calculations, a 3×3×3 Monkhorst-Pack [8] mesh was used for sampling the Brillouin zone, whereas a 6×6×3 and 9×9×7 mesh were used for Li_2SnO_3 and SnO, respectively. Geometry optimizations were performed with a Gaussian smearing of 0.05 eV, followed by a static calculation using the tetrahedron method [9], for a precise calculation of the total energies. The convergence criterion on the electronic optimization is set at 10^{-4} eV, and 10^{-3} eV for the geometric optimization.

Data Management

- What data do you need to store?
  - Raw ingredients (input parameters, conditions, ...)
  - Process instructions (code, program versions, ...)
  - Results
- How will you store your data?
  - How easy is it for yourself or others to read?
  - How scalable does it need to be?
- How can you search through your data?
  - Have clear data schemas
  - Are there interoperable standards in your field?

AiiDA

- COMPUTATIONAL SCIENCE INFRASTRUCTURE
  - FOR HIGH THROUGHPUT WORKFLOWS
    - WITH FULL DATA PROVENANCE



Language: implemented and API in python

License: MIT open source **http://www.aiida.net/**

Source: https://github.com/aiidateam/aiida-core

MIT LICENSED
open source initiative

Scalable workflow engine

Built-in support for HPC

Automated full data provenance

Flexible plugin system

AiiDA data store

CLI/API allows for fast querying

- All data
- Immuta

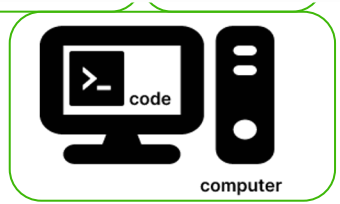
```
%verdi code show pw.x@local_direct
```

▼ Hide cell output

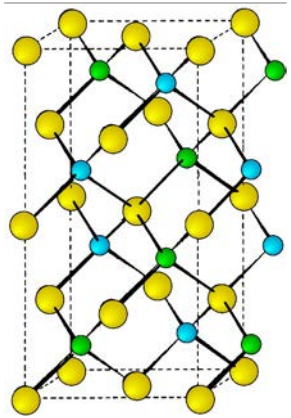
```
--------------------  ------------------------------------------------------
PK                    1
UUID                  43019f16-33cc-468a-99f3-b757d9b3b8bc
Label                 pw.x
Description           pw.x code on local computer
Default plugin        quantumespresso.pw
Type                  remote
Remote machine        local_direct
Remote absolute path  /home/docs/checkouts/readthedocs.org/user_builds/aiid
Prepend text          export OMP_NUM_THREADS=1
Append text
--------------------  ------------------------------------------------------
```

QUANTU

```
&CONTROL
  calculation = 'relax'
  etot_conv_thr =   2.0000000000d-04
  forc_conv_thr =   1.0000000000d-03
  max_seconds =   86400
&SYSTEM
  degauss =   1.0000000000d-02
  ecutrho =   2.4000000000d+02
  ecutwfc =   3.0000000000d+01
/
&ELECTRONS
  conv_thr =   8.0000000000d-10
  electron_maxstep = 80
  mixing_beta =   4.0000000000d-01
/
ATOMIC_SPECIES
Si    28.085 Si.pbe-n-rrkjus_psl.1.0.0.
ATOMIC_POSITIONS angstrom
Si          0.0000000000      0.0000000
Si          1.8940738226      1.093544031
K_POINTS automatic
5 5 5 0 0 0
CELL_PARAMETERS angstrom
      3.7881476452      0.0000000000      0.0000000000
      1.8940738226      3.2806320940      0.0000000000
      1.8940738226      1.0935440313      3.0930096003
```



Interoperable data example - Optimade structures



AFLOW
Automatic-FLOW for Materials Discovery

MATERIALSCLOUD

OPTIMADE
Open Databases Integration for Materials Design

NOMAD

COD

OQMD
The Open Quantum Materials Database

The Materials Project

REST API

Structures

GET /structures Get Structures

GET /structures/{entry_id} Get Single Structure

Python API

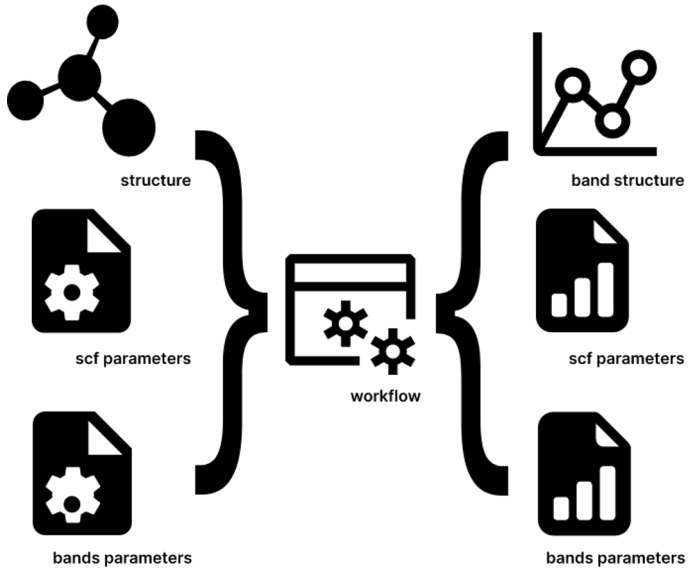
```
from optimade.client import OptimadeClient
client = OptimadeClient()
client.get('elements HAS "Ag" AND nsites < 2')
```

AiiDA

Andersen, C.W., Armiento, R., Blokhin, E. *et al. Sci Data* **8**, 217 (2021)

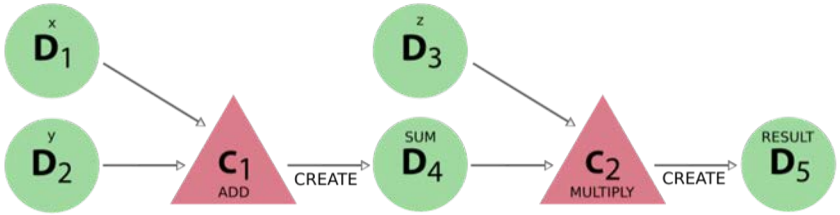
AiiDA workflows

- Python API for plugins to define workflows, to produce new data nodes
  - Workflows can wrap existing computation software
- Plugins encode, re-usable, expert knowledge



Connecting data - Links (a.k.a. Provenance)

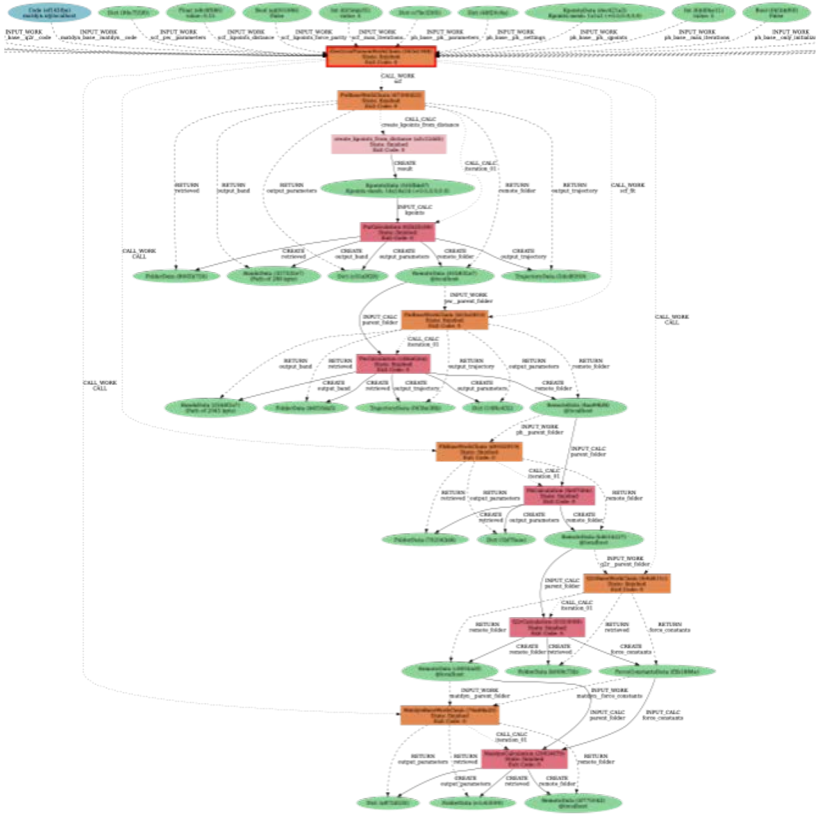
- Multiple workflows can be connected, to form complex Directed Acyclic Graph (DAG)
- Generated automatically during execution
- Link types



Links can be used to query for data!

```
QueryBuilder().append(
  CalcFunctionNode, filters={'label': 'multiply'}, tag='calc'
).append(
  Float, edge_filters={'label': 'result'}, with_incoming='calc'
).all()
```

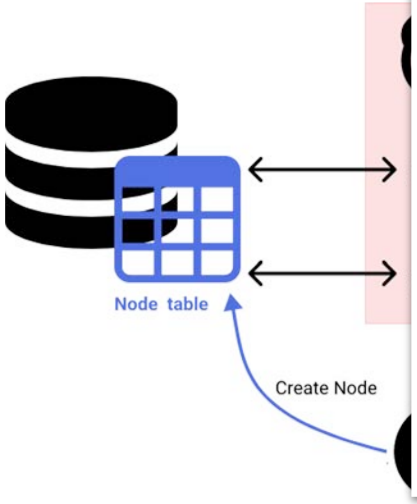
Find me all the "result" outputs of the "multiply" processes.



AiiDA execution engine

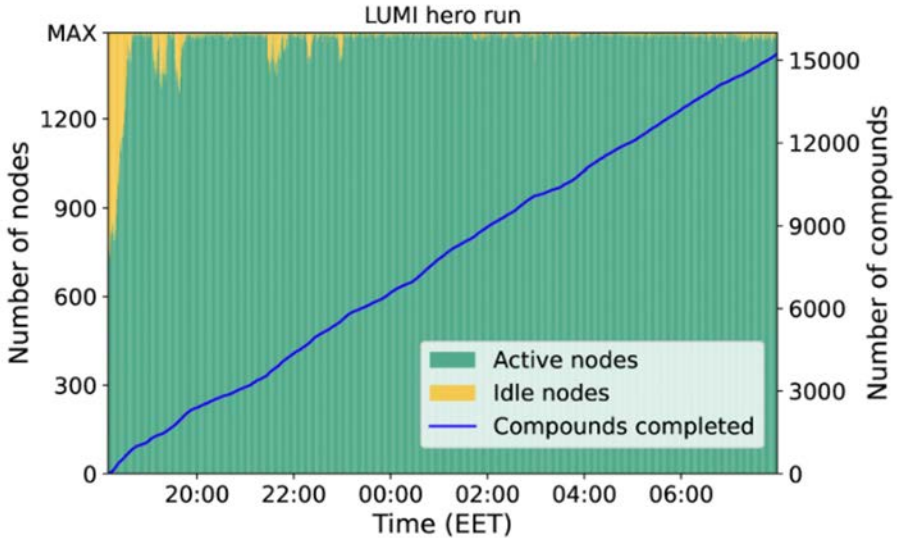
- The AiiDA daem
- Workers can as
  allowing for a h
- Automated inte

workflows
lculations,



LUMI

- Ran 55704 Quantum ESPRESSO calculations in 12 hours
- Handled 7887 errors on the fly
- Fully optimized 15324 geometries

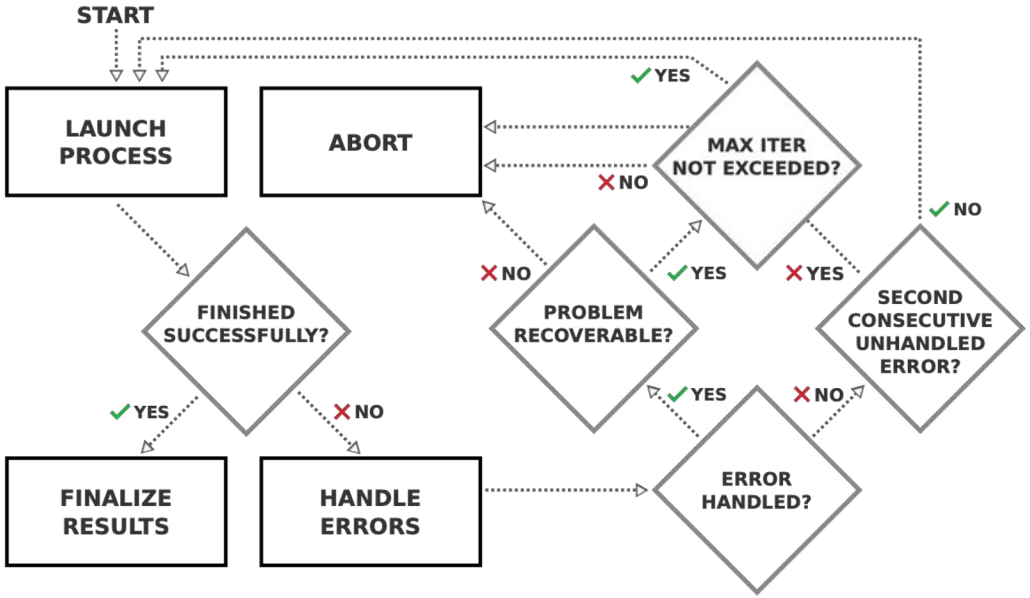


Automated failure and error recovery

Every failure mode is returned as an "exit code"

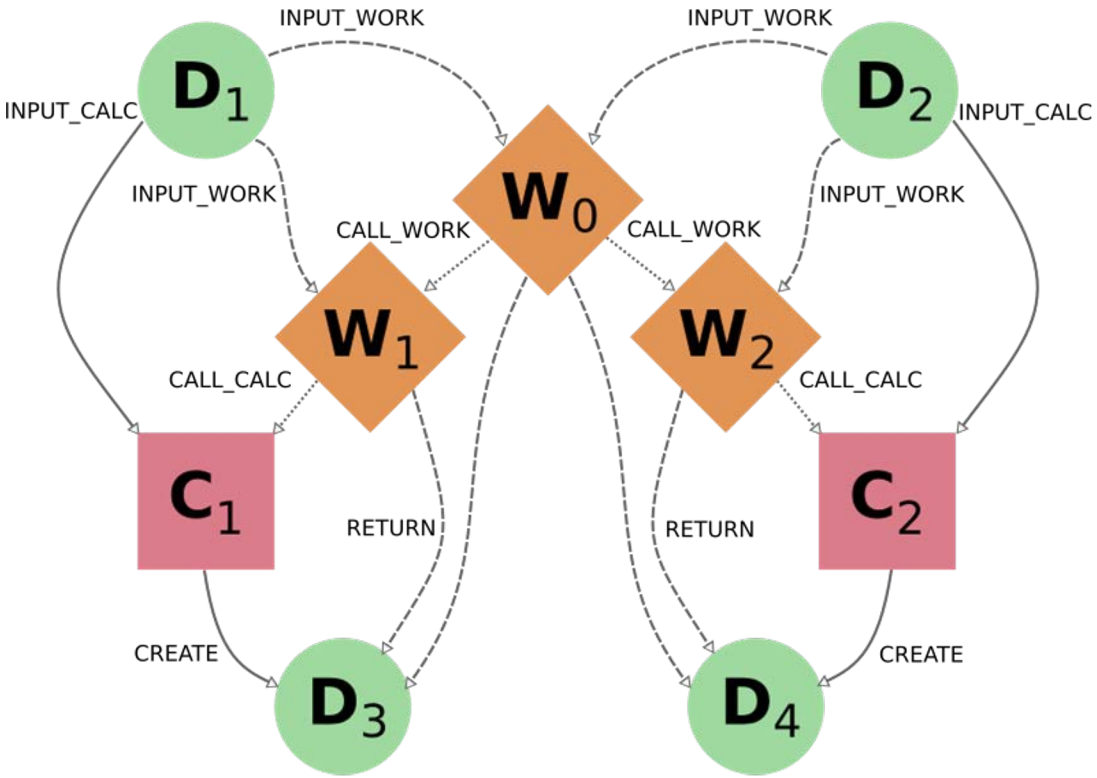
```
Exit codes:

  0  The process finished successfully.
  1  The process has failed with an unspecified error.
  2  The process failed with legacy failure mode.
 10  The process returned an invalid output.
 11  The process did not register a required output.
100  The process did not have the required `retrieved` output.
110  The job ran out of memory.
120  The job ran out of walltime.
131  The specified account is invalid.
140  The node running the job failed.
150  {message}
301  The retrieved temporary folder could not be accessed.
302  The retrieved folder did not contain the required stdout output file.
303  The retrieved folder did not contain the required XML file.
304  The retrieved folder contained multiple XML files.
305  Both the stdout and XML output files could not be read or parsed.
310  The stdout output file could not be read.
311  The stdout output file could not be parsed.
312  The stdout output file was incomplete probably because the calculation got
     interrupted.
320  The XML output file could not be read.
```

BaseRestartWorkChain



How to deal with failures can also be encoded in work chain.

Data consistency



What if you?

- Delete calculation C2?
- Share Data D4?

Data must be kept consistent!

Sharing data on materials cloud

Default records

Engineering frustrated lewis pair active sites in porous organic scaffolds for catalytic CO_2 hydrogenation

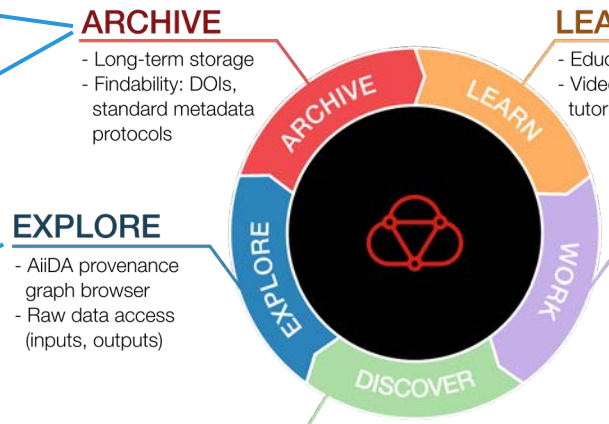
DOI 10.24435/materialscloud:31-wz

AiiDA-generated records

A Standard Solid State Pseudopotentials (SSSP) library optimized for precision and efficiency

DOI 10.24435/materialscloud:f3-ym

MATERIALSCLOUD



ARCHIVE
- Long-term storage
- Findability: DOIs, standard metadata protocols

LEARN
- Educational material
- Videos, courses, tutorials

EXPLORE
- AiiDA provenance graph browser
- Raw data access (inputs, outputs)

WORK
- Simulation tools and services
- AiiDA lab simulation environment (on the cloud & on premises)

DISCOVER
- Curated datasets
- Interactive interfaces and visualizations

GO FAIR

- **F**indable
- **A**ccessible
- **I**nteroperable
- **R**eusable

Materials Cloud 3D section

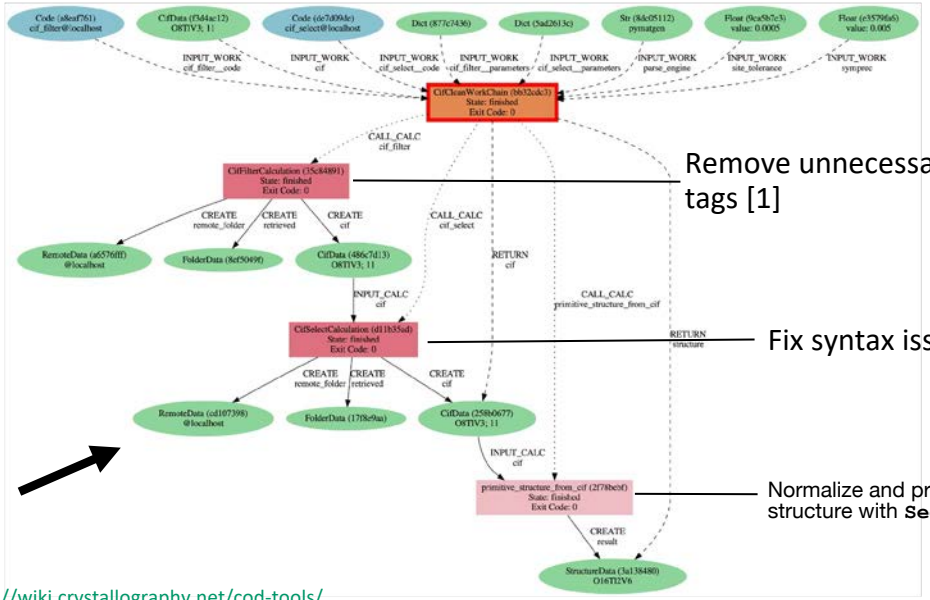
MPDS

ICSD
FIZ Karlsruhe

COD

901,210 Experimental structures

CifCleanWorkChain



Remove unnecessary tags [1]

Fix syntax issues [1]

Normalize and primitivize structure with `SeeK-path` [2]

[1] https://wiki.crystallography.net/cod-tools/
[2] https://www.materialscloud.org/work/tools/seekpath
[3] https://pymatgen.org/

858,296 Structures

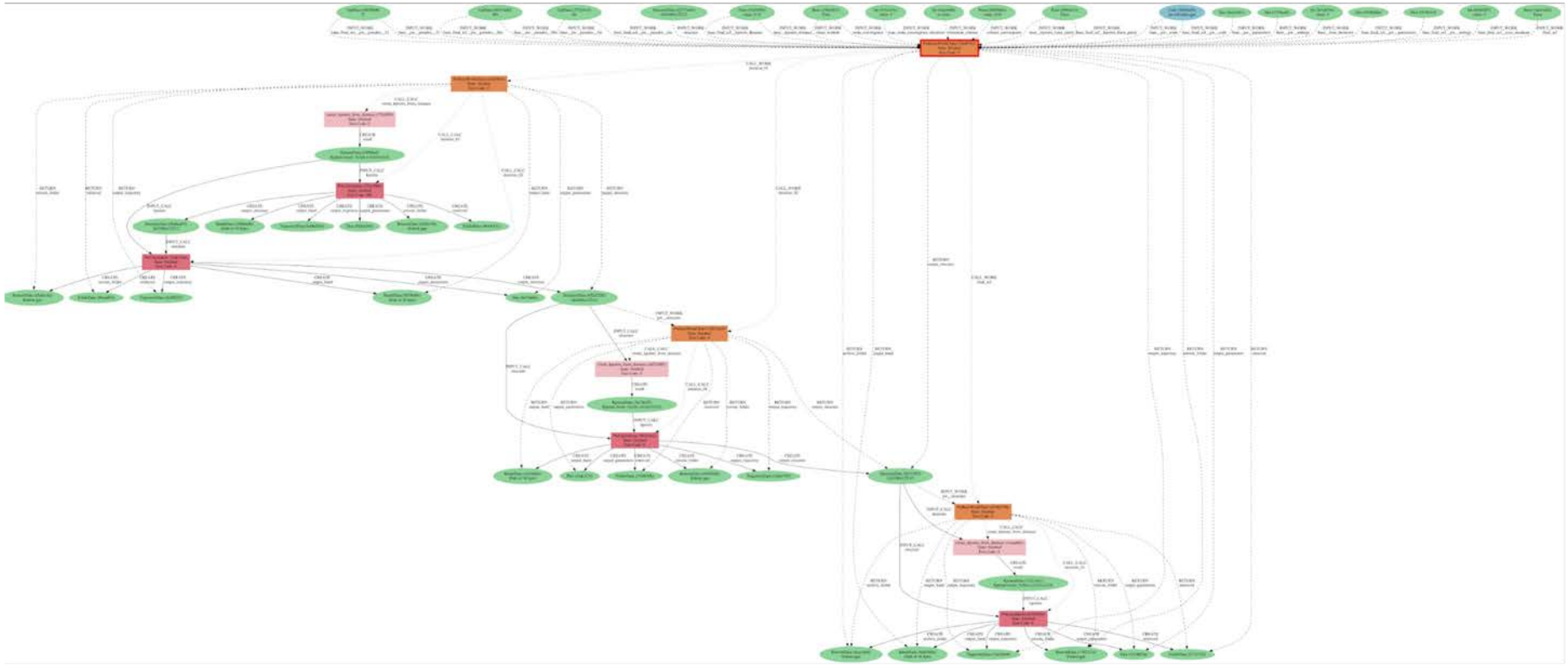
Remove non-stoichiometric compounds

pymatgen[3] `StructureMatcher`

Remove molecular crystals from COD

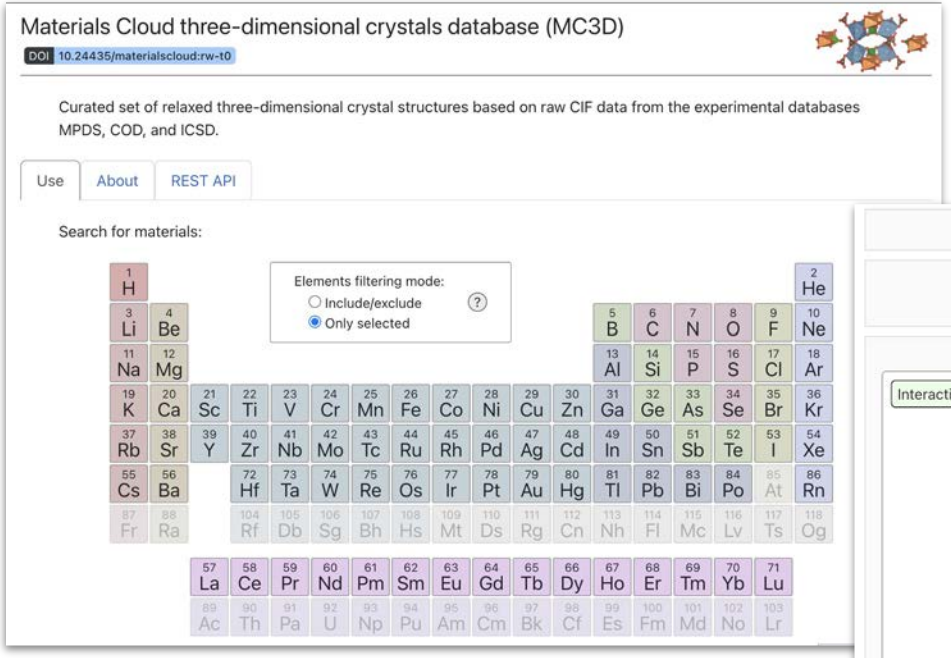
84,064 Structures

Materials Cloud 3D section



Materials Cloud 3D section

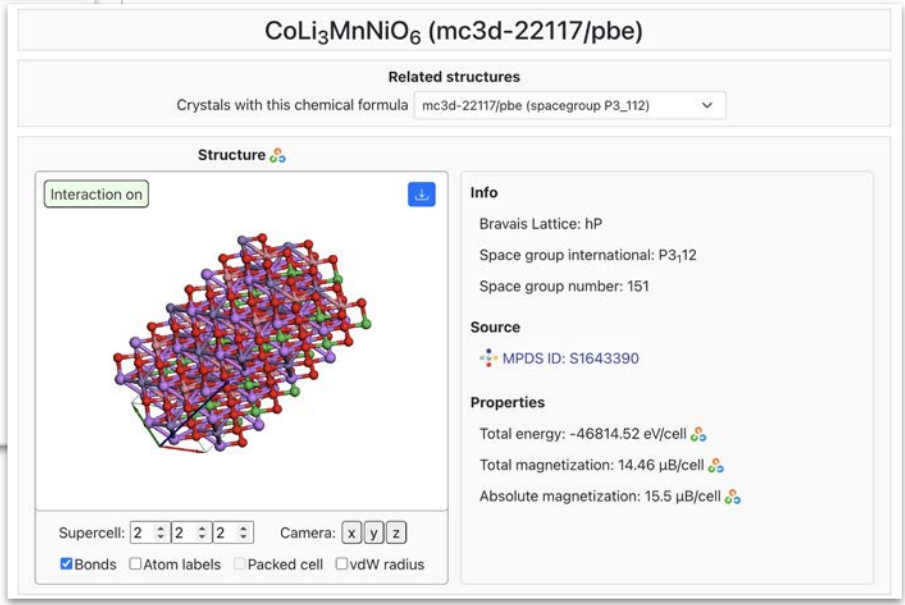
https://mc3d.materialscloud.org/



New React frontend published!

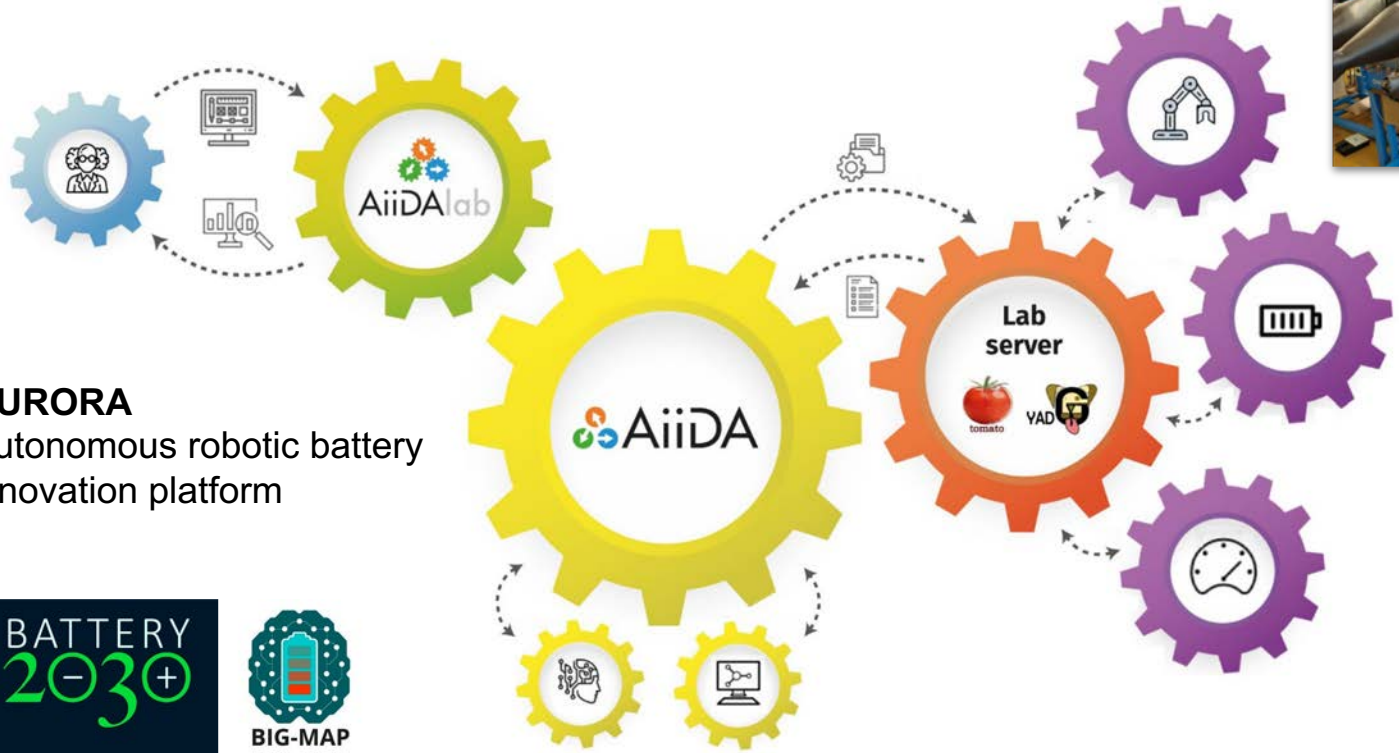
React

Dr. Kristjan Eimre



- All structures up to 60 atoms in unit cell run
- 35,000 structures published
- Completed over 31,000 band structures

What about experimental work?



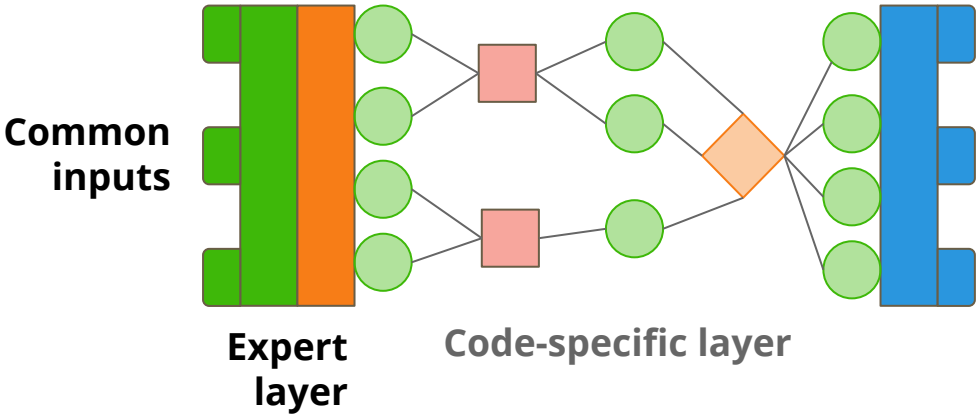
AURORA
Autonomous robotic battery innovation platform

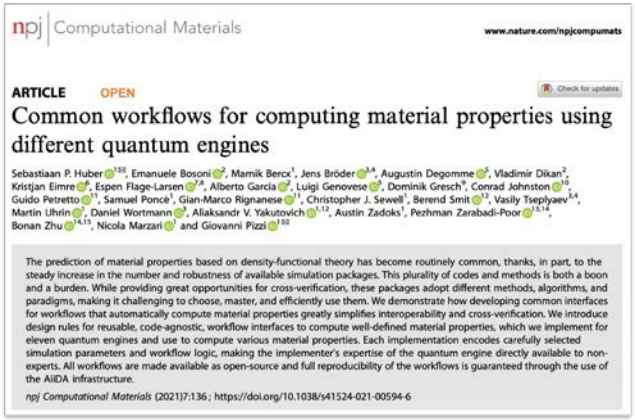
BATTERY 2030+

BIG-MAP

Beyond data: FAIR simulation access

- Open-source, robust turn-key workflows for materials properties
- Common interface among 11 quantum engines
- Easy to use, but flexible for experts, designed to be interfaced to GUIs





npj | Computational Materials

www.nature.com/npjcompumats

ARTICLE OPEN

Common workflows for computing material properties using different quantum engines

Sebastiaan P. Huber[1✉], Emanuele Bosoni[2], Marnik Bercx[3], Jens Bröder[3,4], Augustin Degomme[5], Vladimir Dikan[2], Kristjan Eimre[6], Espen Flage-Larsen[7,8], Alberto Garcia[2], Luigi Genovese[9], Dominik Gresch[6], Conrad Johnston[10], Guido Petretto[11], Samuel Poncé[11], Gian-Marco Rignanese[11], Christopher J. Sewell[3], Berend Smit[12], Vasily Tseplyaev[3,4], Martin Uhrin[3], Daniel Wortmann[6], Aliaksandr V. Yakutovich[1,13], Austin Zadoks[1], Pezhman Zarabadi-Poor[13,14], Bonan Zhu[14,15], Nicola Marzari[6] and Giovanni Pizzi[1✉]

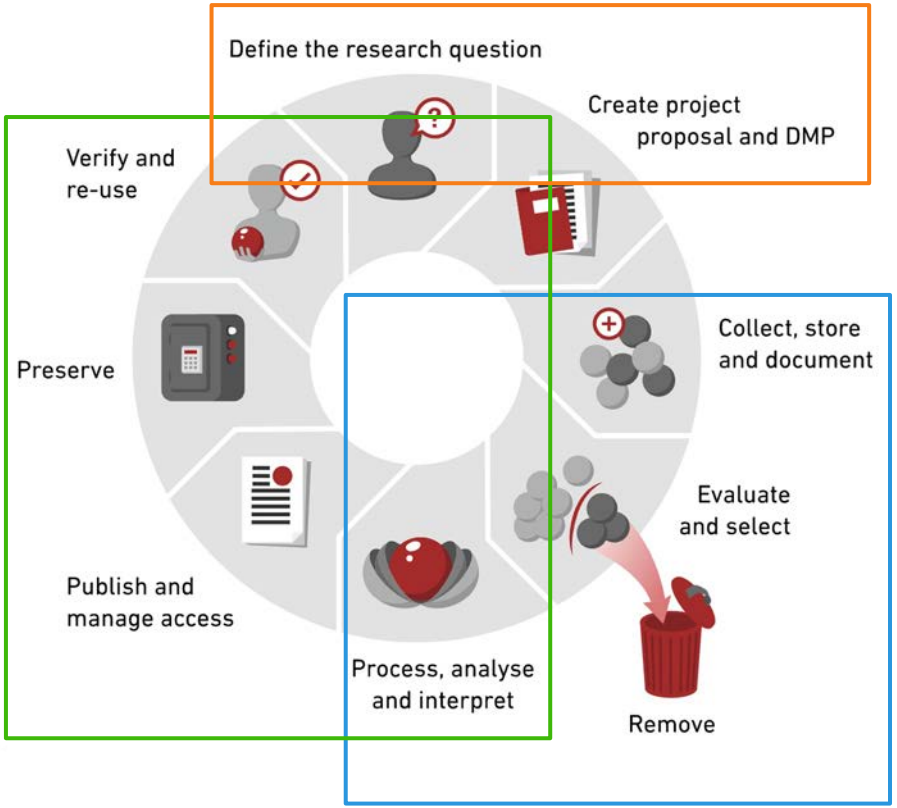
The prediction of material properties based on density-functional theory has become routinely common, thanks, in part, to the steady increase in the number and robustness of available simulation packages. This plurality of codes and methods is both a boon and a burden. While providing great opportunities for cross-verification, these packages adopt different methods, algorithms, and paradigms, making it challenging to choose, master, and efficiently use them. We demonstrate how developing common interfaces for workflows that automatically compute material properties greatly simplifies interoperability and cross-verification. We introduce design rules for reusable, code-agnostic, workflow interfaces to compute well-defined material properties, which we implement for eleven quantum engines and use to compute various material properties. Each implementation encodes carefully selected simulation parameters and workflow logic, making the implementer's expertise of the quantum engine directly available to non-experts. All workflows are made available as open-source and full reproducibility of the workflows is guaranteed through the use of the AiiDA infrastructure.

npj Computational Materials (2021)7:136 ; https://doi.org/10.1038/s41524-021-00594-6

AiiDA and the data life cycle

MATERIALSCLOUD

- Data can be exported into AiiDA "archive".
- Easily shared on Materials Cloud.
- Can be imported: data is not duplicated!
- Can be opened on the fly for analysis.



- Demonstrating (reproducible) plan.
- Data/IO plan becomes more concrete.

AiiDA

- Workflows are (mostly) self-documenting.
- Plugin decided which data is stored.
- Cleaning up data is easy and consistent.

Thank you for your attention!