AiiDA: a scalable computational infrastructure for automated reproducible workflows and data provenance

Presenter: Francisco F. Ramirez (THEOS @ EPFL)

MARVEL · MAX DRIVING THE EXASCALE TRANSITION · swissuniversities · FNSNF Swiss National Science Foundation · INTERSECT · MarketPlace · EPFL

CHALLENGES OF (HIGH-THROUGHPUT) COMPUTATIONAL SCIENCE

Francisco Ramirez

Robust workflow automation

- Define complex workflows with built-in error handling
- Automated, high-performant scheduling and execution on local and remote resources

Efficient data management

- Flexible integration of databases and file repositories
- Schemas must be both performant and adaptable
- Data must be stored automatically, but also remain discoverable

Ensuring reproducibility

- Logging of calculations and the computational environment
- Accurate tracking of full data provenance

These challenges will be hardened at exascale!

DATA PROVENANCE

Francisco Ramirez

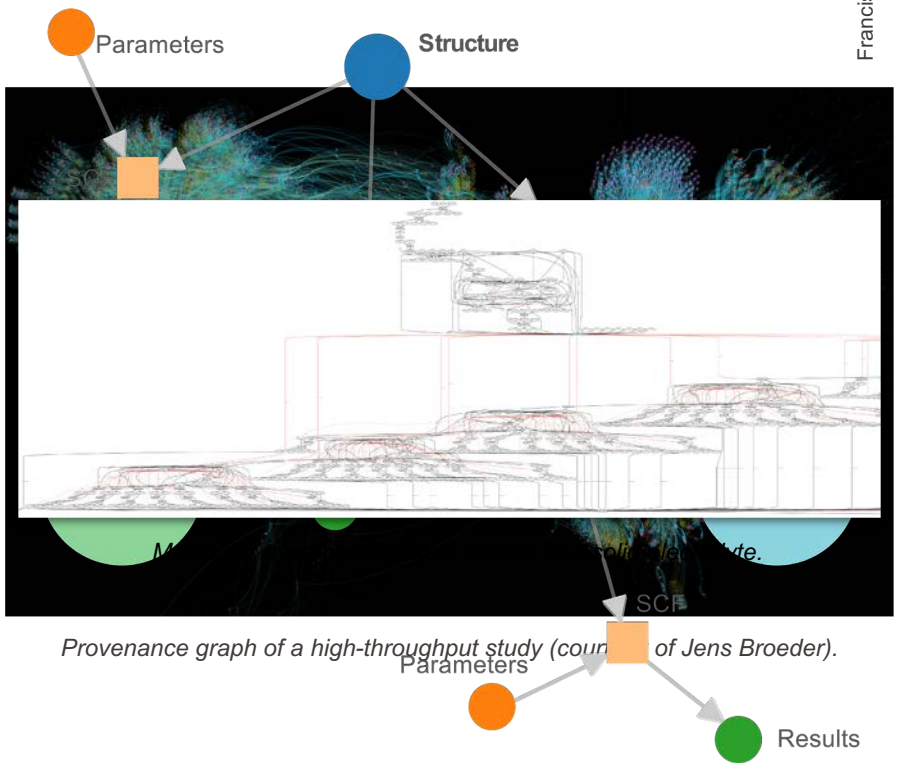
Simple recipe

- Store data transformations or '**calculations**'
- Store its **inputs** and their metadata
- Store its **outputs** and their metadata
- Most **crucially** store the **inter-connections**

Requirements

- Needs to be automated
- Needs to be stored *as data is created*

Complexity grows quickly even for simple workflows and is impossible to reconstruct *a posteriori*



Provenance graph of a high-throughput study (coun of Jens Broeder).

AiiDA

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

Language: Implemented and API in Python

License: MIT open source www.aiida.net

Source: github.com/aiidateam/aiida-core

Numfocus affiliated project as of February 2020

MIT LICENSED
open source initiative

NUMFOCUS
OPEN CODE = BETTER SCIENCE

DESIGN PRINCIPLES

Francisco Ramirez

- runs on your computer, connects to cluster via SSH
- daemon manages calculations
- flexible plugin system (codes, data types, schedulers, …)
- dynamic workflows
- focus on provenance tracking & sharing
- PostgreSQL backend

- easily parallelize existing code
- serverless
- globus integration

signac

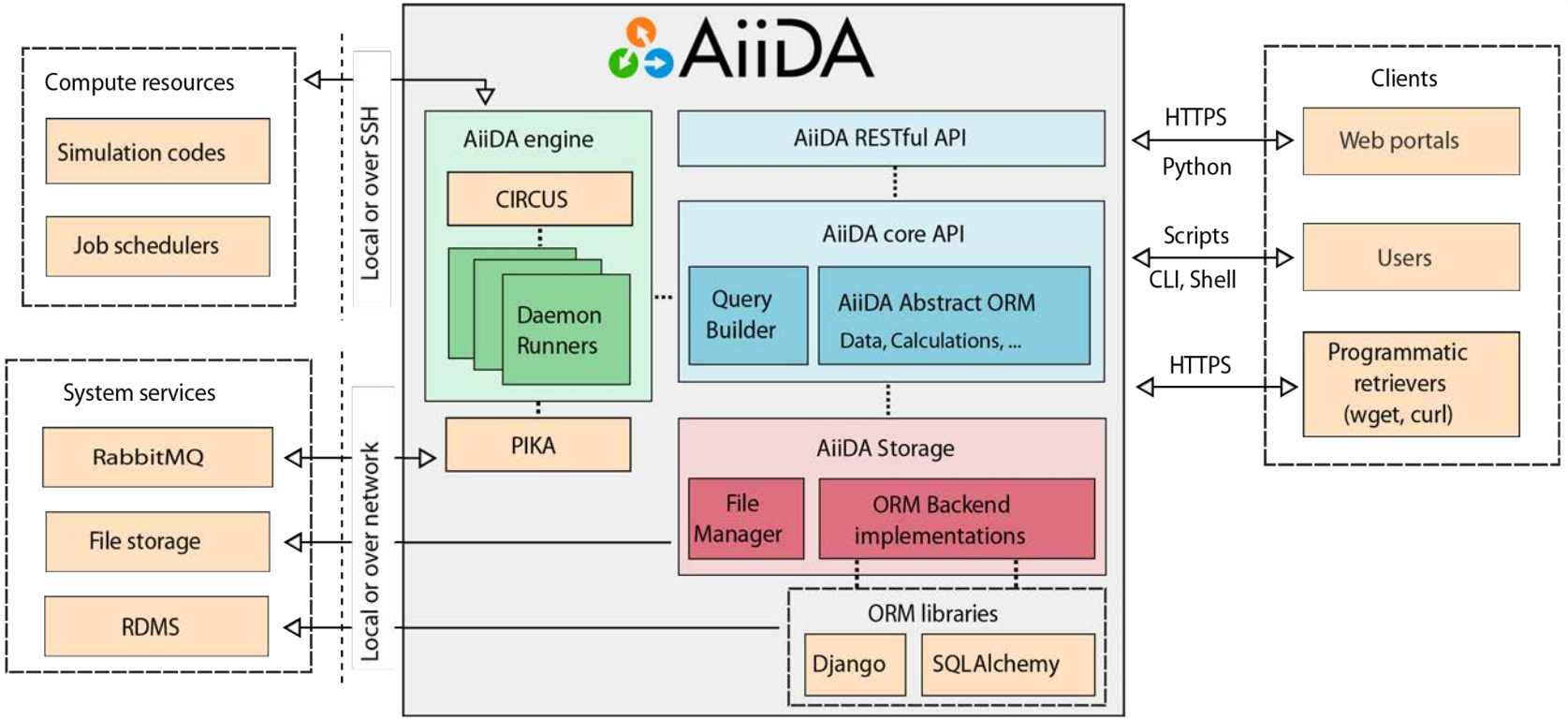
- runs on cluster
- serverless
- fast prototyping, simple data schema

- runs on cluster
- MongoDB backend
- dynamic workflows

• • •

See e.g. github.com/tilde-lab/awesome-materials-informatics

AiiDA ARCHITECTURE



- AiiDA 0.x used in production since 2016
  - Most recent release: AiiDA 1.4.x in September 2020 (new features, fully backward-compatible)

Francisco Ramirez

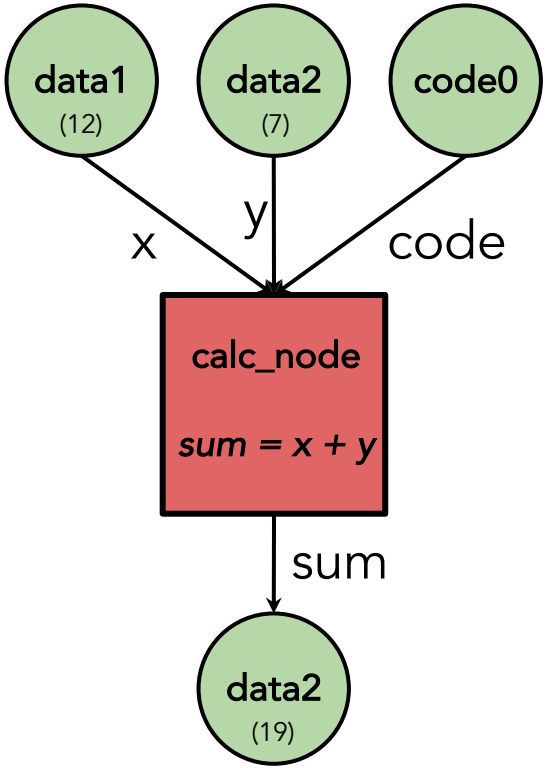
USAGE EXAMPLE

Francisco Ramirez

Verdi (Python) Shell:

```
In [1]: CalcClass = ArithmeticAddCalculation
In [2]: code0 = load_code(label='add')
In [3]: data1 = orm.Int(12)
In [4]: data2 = orm.Int(7)
In [5]: inputs = {
   ...:     'x': data1,
   ...:     'y': data2,
   ...:     'code': code,
   ...: }
In [6]: calc_node = submit(CalcClass, **inputs)
In [7]: data3 = calc_node.outputs.sum
In [8]: ...
```

DATABASE

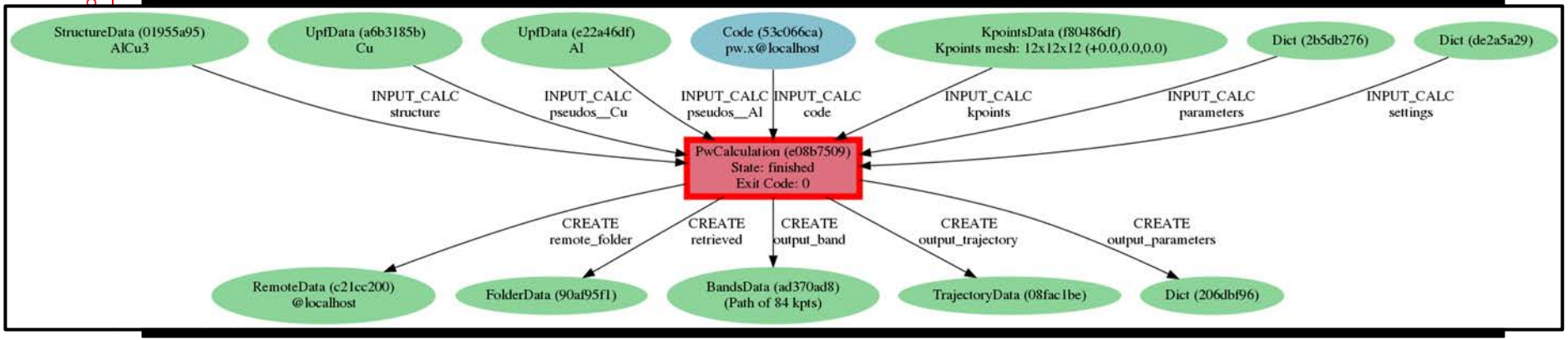


USAGE EXAMPLE

Francisco Ramirez

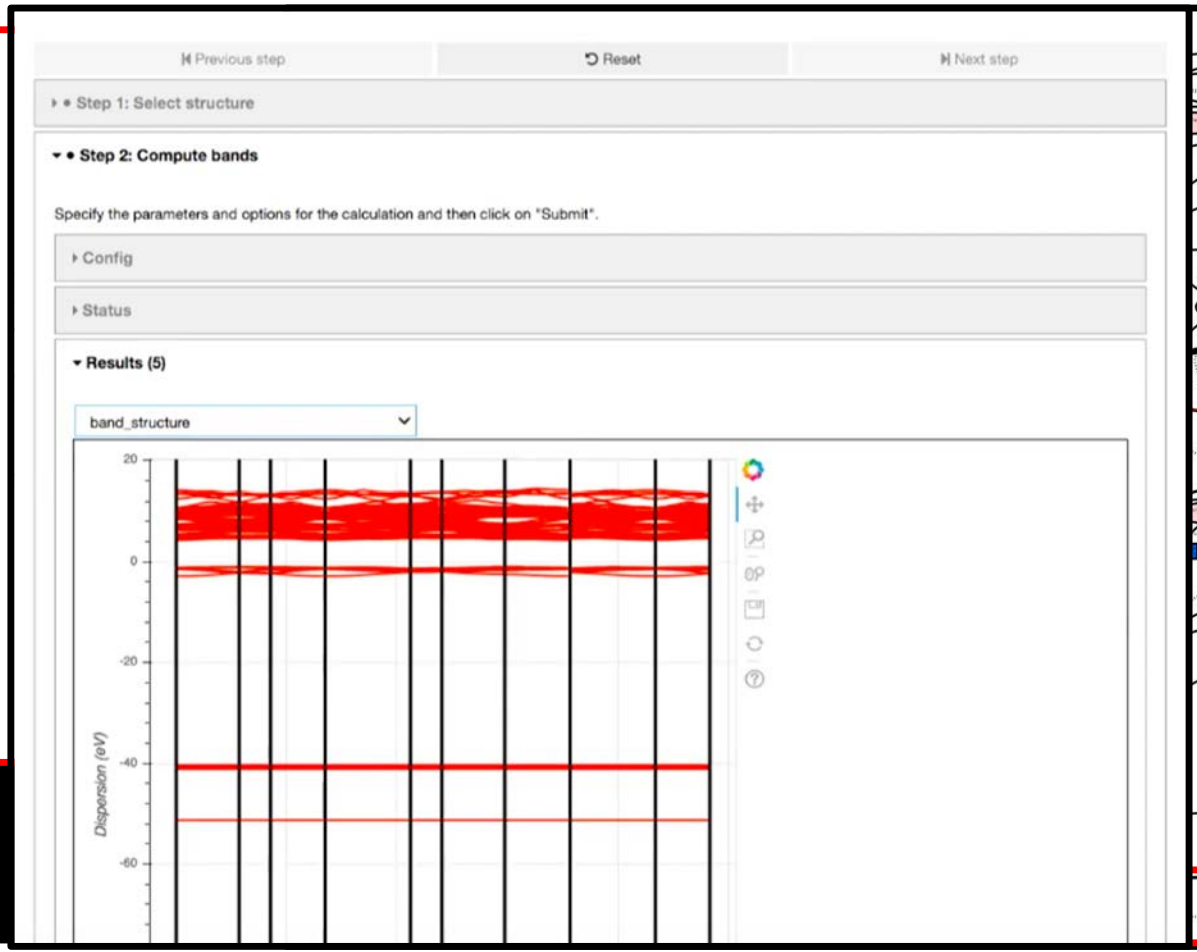
Verdi (Python) Shell:

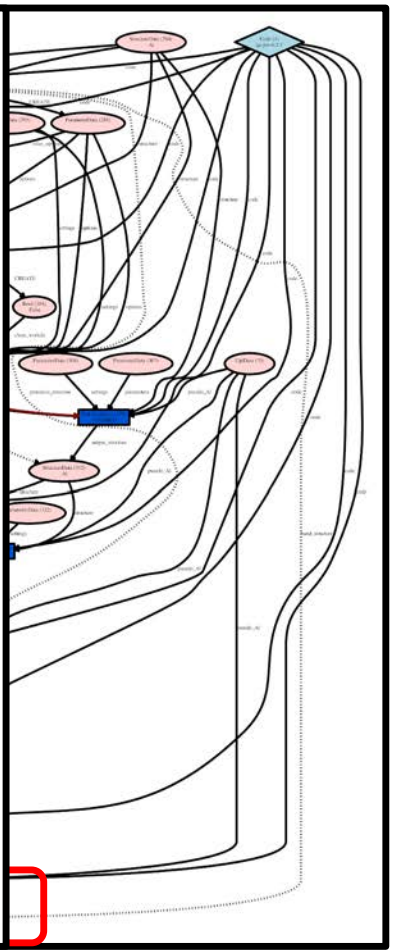
```
In [1]: QE_class = CalculationFactory('quantumespresso.pw')
In [2]: inputs = {
   ...:     'code':       load_code(label='qe-pw@cluster'),
   ...:     'kpoints':    load_node(label='AlCu3 kpoints'),
   ...:     'structure':  load_node(label='AlCu3 struct'),
   ...:     'settings':   load_node(label='AlCu3 pw settings'),
   ...:     'parameters': load_node(label='AlCu3 pw params'),
```



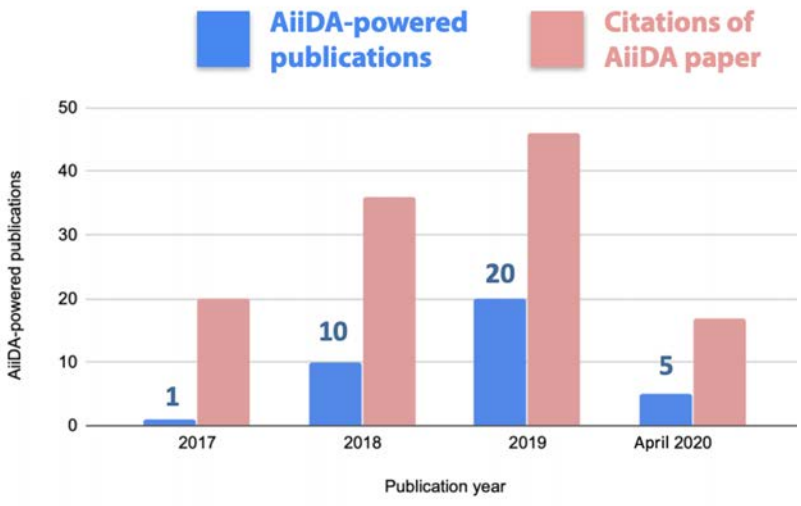
USAGE EXAMPLE

Francisco Ramirez





SCIENCE SURVEY 2020



Simulation codes / plugins used

16x CP2K

15x QUANTUMESPRESSO

3x raspa, zeo++

2x VASP, WANNIER90, TBmodels

symmetry-representation, tbextraction, …

1x Yambo, CASTEP, COD tools, Z2Pack

ddec, …

69 research projects

35 extensions used

11 different countries of affiliation

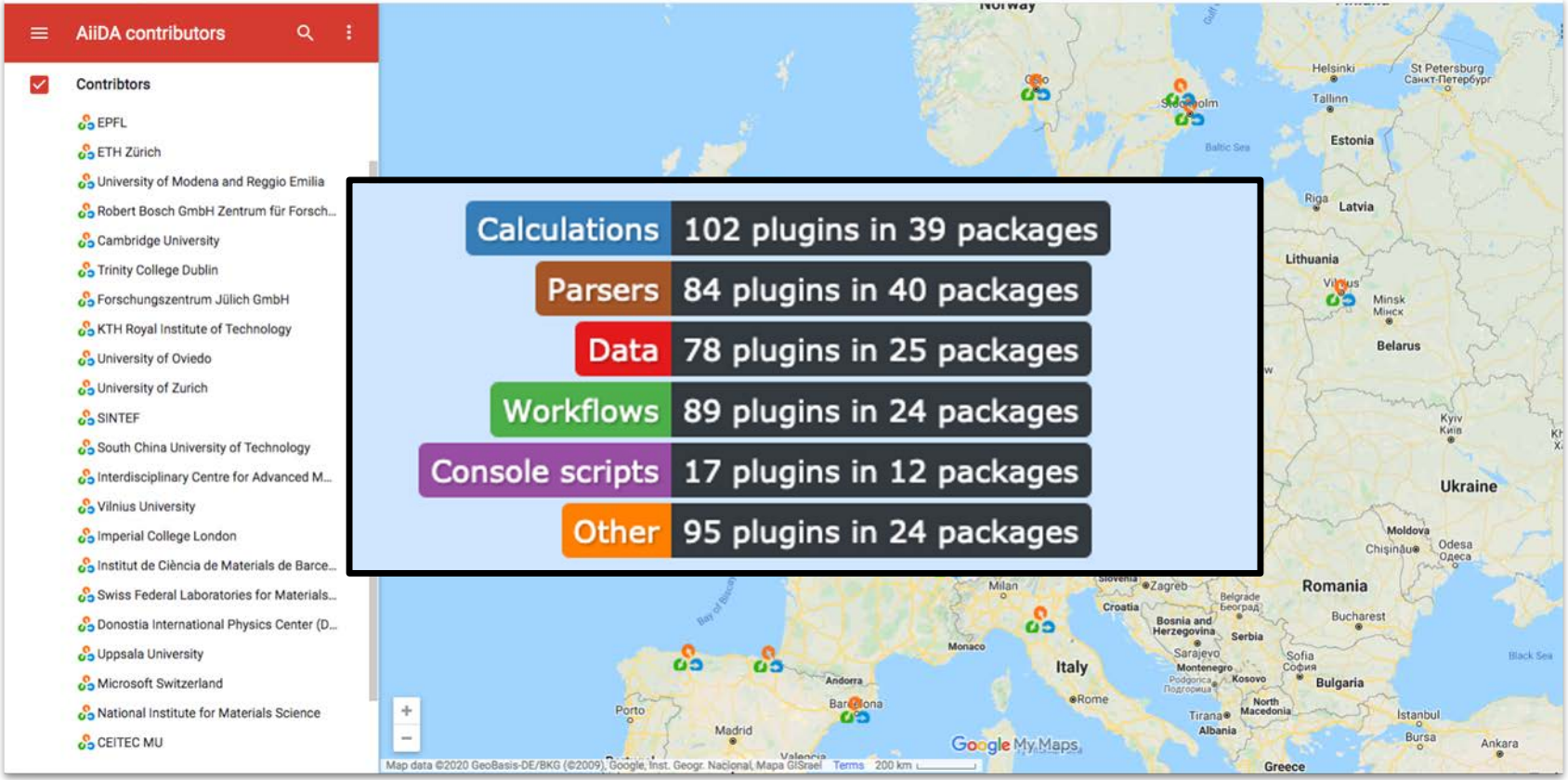
16% from research institutes/industry

More Information: www.aiida.net/science

Francisco Ramirez

AiiDA CONTRIBUTIONS

Francisco Ramirez



CONCLUSION

Francisco Ramirez

AiiDA

DRIVING REPRODUCIBLE COMPUTATIONAL SCIENCE TOWARDS EXASCALE

Efficient workflow engine

Built-in support for HPC

Automated data provenance tracking

Flexible plug-in system

“Laying the foundations for a way of doing science that seamlessly leaves behind a comprehensible and reproducible path”

ACKNOWLEDGMENTS

MARVEL National Centre for Competency in Research

European Centre of Excellence MaX

École Polytechnique Fédérale de Lausanne

Swiss National Science Foundation

swissuniversities

MarketPlace

INTERSECT

European Research Council

Platform for Advanced Scientific Computing

PRACE PARTNERSHIP FOR ADVANCED COMPUTING IN EUROPE

nffa.eu nanoscience foundries & fine analysis

The European Materials Modelling Council

Article | Open Access | Published: 08 September 2020

AiiDA 1.0, a scalable computational infrastructure for automated reproducible workflows and data provenance

Sebastiaan P. Huber, Spyros Zoupanos, Martin Uhrin, Leopold Talirz, Leonid Kahle, Rico Häuselmann, Dominik Gresch, Tiziano Müller, Aliaksandr V. Yakutovich, Casper W. Andersen, Francisco F. Ramirez, Carl S. Adorf, Fernando Gargiulo, Snehal Kumbhar, Elsa Passaro, Conrad Johnston, Andrius Merkys, Andrea Cepellotti, Nicolas Mounet, Nicola Marzari, Boris Kozinsky & Giovanni Pizzi

Scientific Data **7**, Article number: 300 (2020) | Cite this article

1037 Accesses | **6** Altmetric | Metrics

We also acknowledge code contributions, bug fixes, documentation, and suggestions from:

Oscar Arbelaez, Michael Atambo, Valentin Bersier, Marco Borelli, Jocelyn Boullier, Jens Bröder, Ivano E. Castelli, Andrea Cepellotti, Keija Cui, Vladimir Dikan, Marco Dorigo, Y.-W. Fang, Fernando Gargiulo, Marco Gibertini, Davide Grassano, Conrad Johnston, Rico Häuselmann, Daniel Hollas, Eric Hontz, Jianxing Huang, Christoph Koch, Espen Flage Larsen, Boris Kozinsky, Ian Lee, Daniel Marchand, Antimo Marrazzo, Andrius Merkys, Nicolas Mounet, Simon Pintarelli, Gianluca Prandini, Philipp Rüßmann, Riccardo Sabatini, Ole Schütt, Philippe Schwaller, Chris Sewell, Andreas Stamminger, Atsushi Togo, Daniele Tomerini, Nicola Varini, Jason Yu, Austin Zadoks, Bonan Zhu and Mario Zic.

ONLINE RESOURCES AND CONTACT

Francisco Ramirez

AiiDA

WEBSITE

http://www.aiida.net

SOURCE CODE

github.com/aiidateam/aiida-core

DOCUMENTATION

https://aiida.readthedocs.io

CONTACT INFORMATION

francisco.ramirez@epfl.ch