

# Decentralized materials research data management, curation & dissemination for accelerated discovery

Dr Matthew Evans

<https://ml-evs.science>



FMRS24 Boston, USA, 3rd December 2024

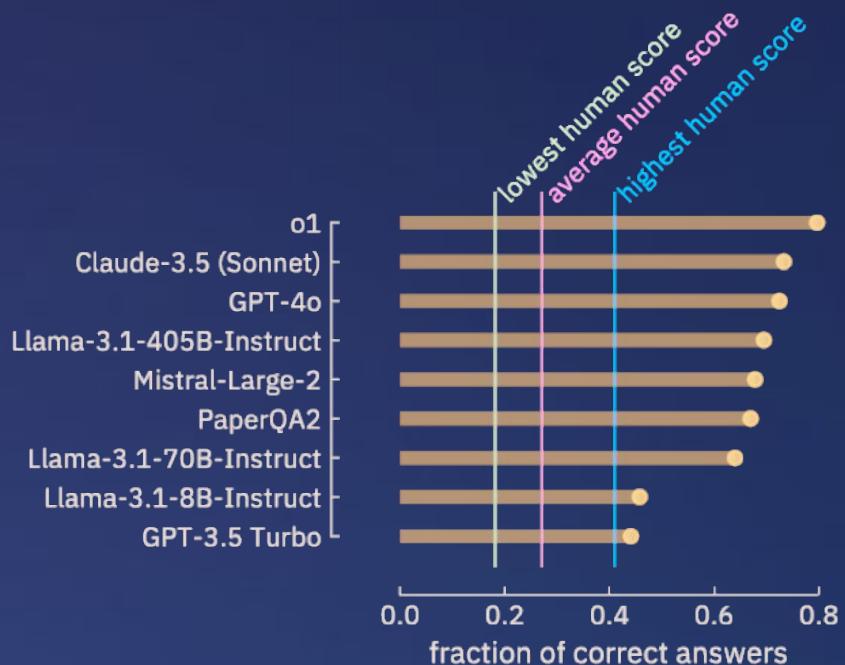
BI01: Democratizing AI in Materials Science—A Pathway to Broaden the Impact of Materials Research

# Democratizing AI

- Access to compute (GPU-poor/rich)
- Access to expertise
- Access to models
- Getting your data seen
- Bringing your own tools to the AI

# Are large language models superhuman chemists?

Adrian Mirza <sup>1,2,\*</sup>, Nawaf Alampara <sup>1,\*</sup>, Sreekanth Kunchapu <sup>1,\*</sup>,  
Martíño Ríos-García <sup>1,3,\*</sup>, Benedict Emoekabu , Aswanth Krishnan <sup>4</sup>,  
Tanya Gupta <sup>5,6</sup>, Mara Schilling-Wilhelmi <sup>1</sup>, Macjonathan Okereke <sup>1</sup>,  
Anagha Aneesh <sup>1</sup>, Mehrdad Asgari <sup>7</sup>, Juliane Eberhardt <sup>8</sup>,  
Amir Mohammad Elahi <sup>9</sup>, Hani M. Elbeheiry <sup>10</sup>, María Victoria Gil <sup>3</sup>,  
Christina Glaubitz <sup>1</sup>, Maximilian Greiner<sup>1</sup>, Caroline T. Holick <sup>1,14</sup>,  
Tim Hoffmann <sup>1, 14</sup>, Abdelrahman Ibrahim <sup>1</sup>, Lea C. Klepsch <sup>1, 14</sup>,  
Yannik Köster <sup>1</sup>, Fabian Alexander Kreth <sup>11, 12</sup>, Jakob Meyer<sup>1</sup>, Santiago Miret <sup>13</sup>,  
Jan Matthias Peschel <sup>1</sup>, Michael Ringleb <sup>1, 14</sup>, Nicole Roesner <sup>1, 14</sup>,  
Johanna Schreiber <sup>1, 14</sup>, Ulrich S. Schubert <sup>1, 2, 10, 14</sup>, Leanne M. Stafast <sup>1, 14</sup>,  
Dinga Wonanke <sup>15</sup>, Michael Pieler <sup>16, 17</sup>, Philippe Schwaller <sup>5, 6</sup>, and  
Kevin Maik Jablonka <sup>1, 2, 11, 14, 18</sup>



Broadening impact via decentralized platforms

## OPTIMADE

Federated access to  
crystal structure  
data with universal  
search for materials  
discovery & design

## *datalab*

An open source data  
management platform  
for experimental  
materials chemistry  
and beyond

**datatractor + MADICES**

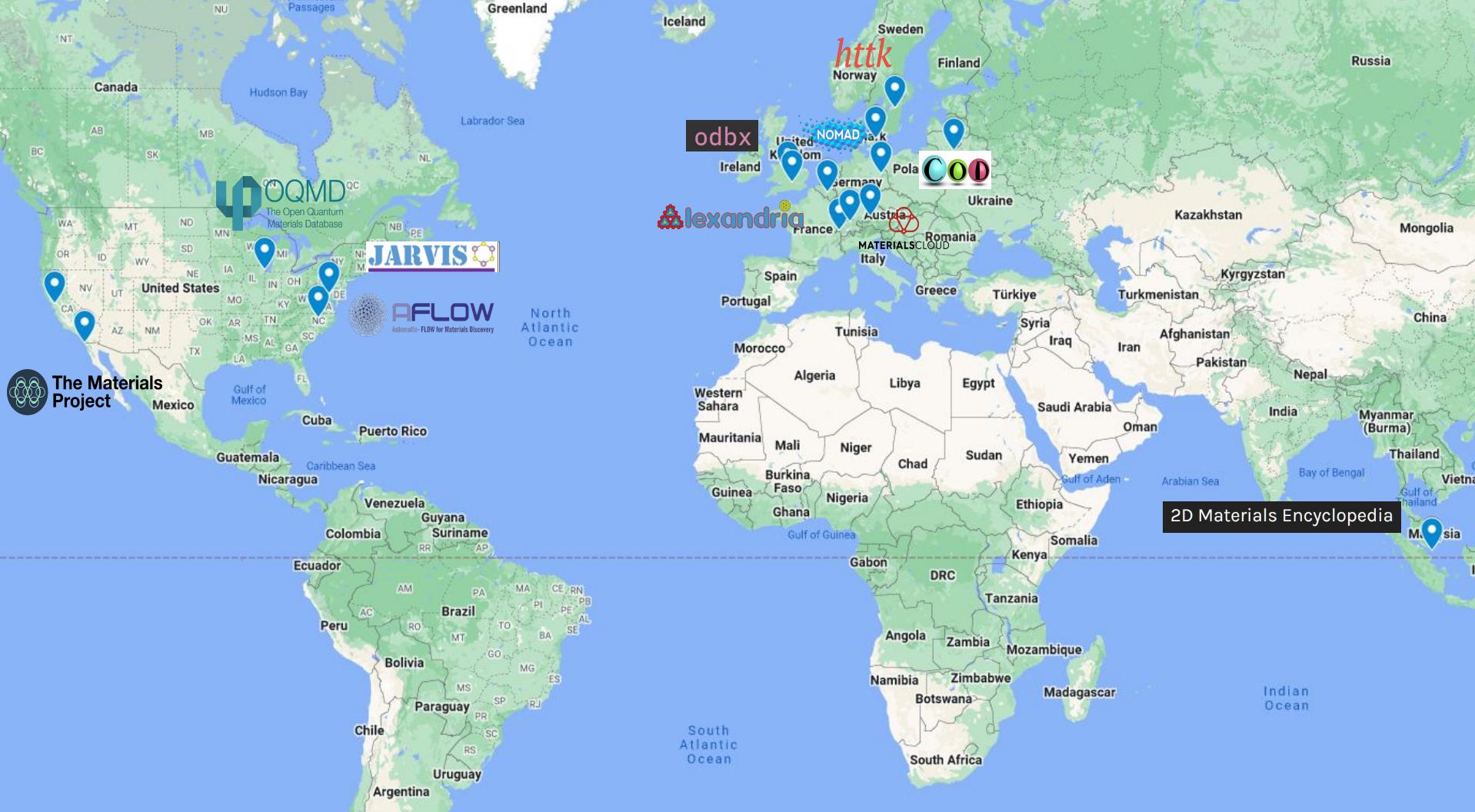
# What is OPTIMADE?

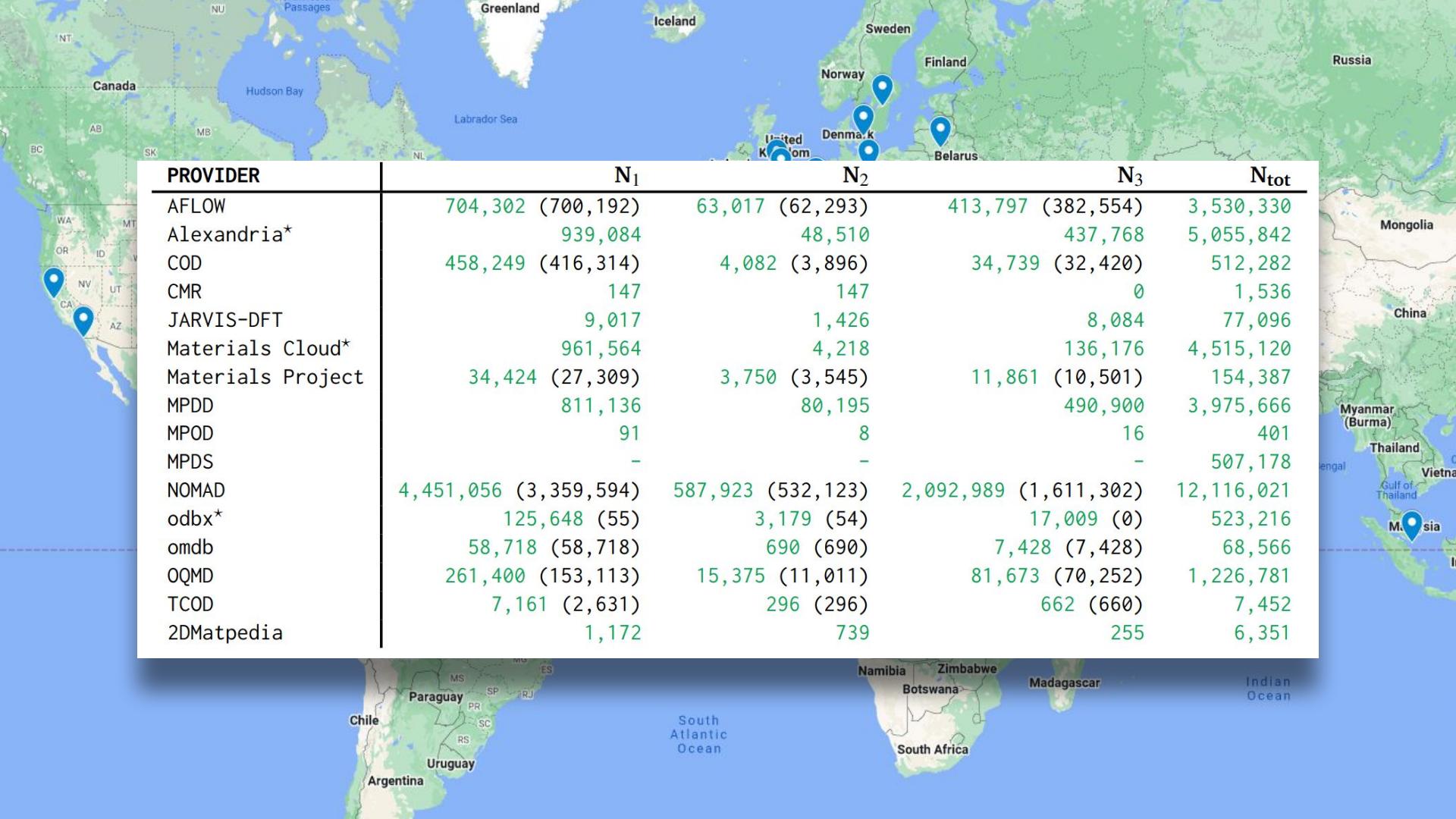


“The Open Databases Integration for Materials Design (OPTIMADE) consortium aims to make materials databases interoperational by developing a common REST API.”

- ❖ 7 annual workshops (est. 2016)
- ❖ 70+ authors/attendees
- ❖ 35,671 words
- ❖ 25 registered providers, 60M+ crystal structures

<https://www.optimade.org/providers-dashboard>





Article | Open

OPTIMADE

Casper W. An

Abhijith Gopa

Markus Sche

Curtarolo, Da

Hautier, Vinay

Kumbhar, Mo

Francesco Ri

Chris Wolverton

Scientific Dat

# Developments and applications of the OPTIMADE API for materials discovery, design, and data exchange†

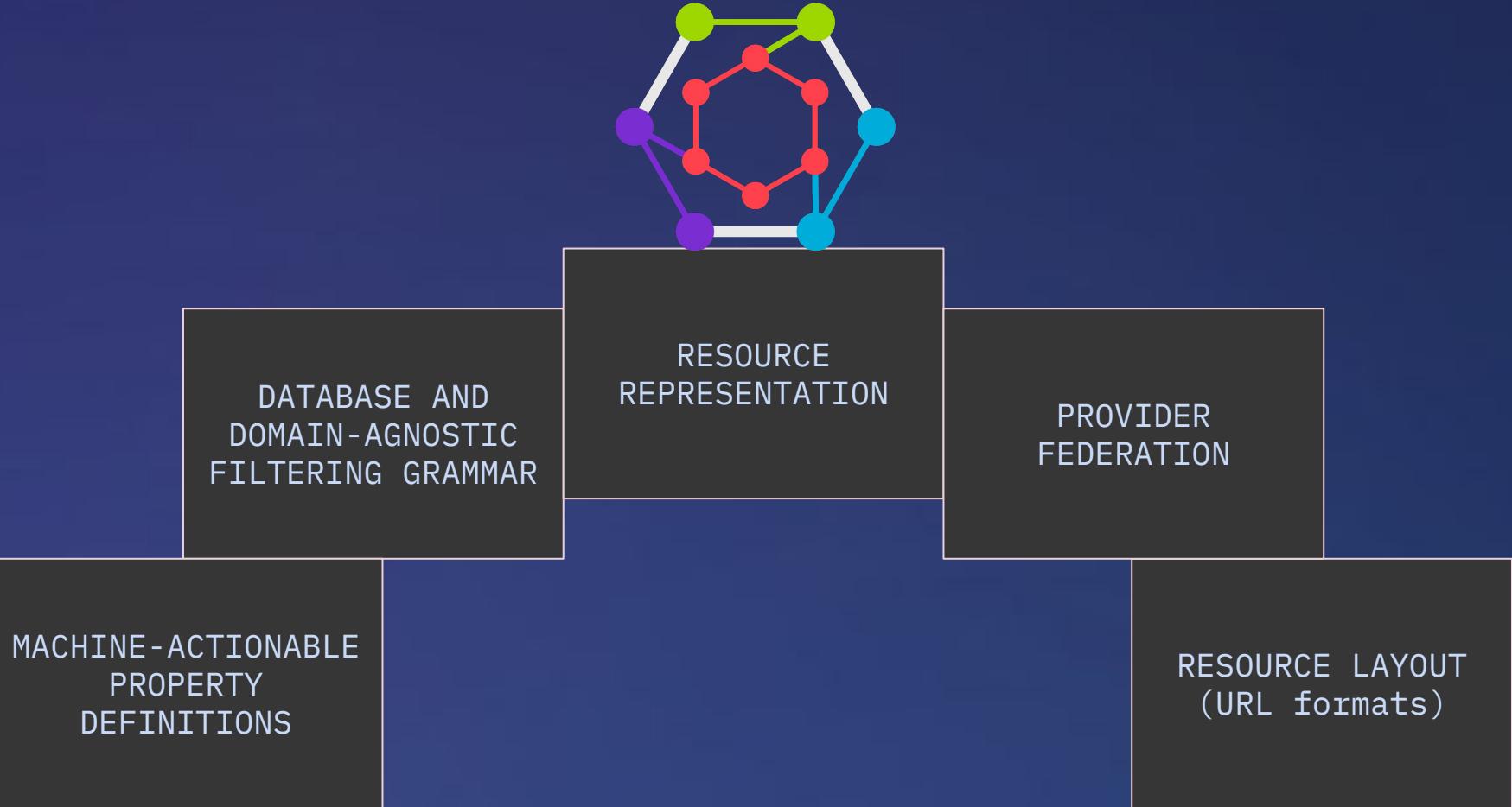
Cite this: *Digital Discovery*, 2024, 3, 1509

DOI: 10.1039/d4dd00039k

Matthew L. Evans,  <sup>a,b</sup> Johan Bergsma,  <sup>f,c</sup> Andrius Merkys,  <sup>d</sup> Casper W. Andersen,  <sup>e</sup> Oskar B. Andersson,  <sup>f</sup> Daniel Beltrán, <sup>g</sup> Evgeny Blokhin,  <sup>hi</sup> Tara M. Boland,  <sup>j</sup> Rubén Castañeda Balderas,  <sup>k</sup> Kamal Choudhary,  <sup>l</sup> Alberto Díaz Díaz,  <sup>k</sup> Rodrigo Domínguez García,  <sup>k</sup> Hagen Eckert,  <sup>mn</sup> Kristjan Eimre,  <sup>o</sup> María Elena Fuentes Montero,  <sup>p</sup> Adam M. Krajewski,  <sup>q</sup> Jens Jørgen Mortensen,  <sup>j</sup> José Manuel Nápoles Duarte,  <sup>p</sup> Jacob Pietryga,  <sup>r</sup> Ji Qi, <sup>s</sup> Felipe de Jesús Trejo Carrillo,  <sup>k</sup> Antanas Vaitkus,  <sup>d</sup> Jusong Yu,  <sup>ah</sup> Adam Zettel,  <sup>mn</sup> Pedro Baptista de Castro,  <sup>t</sup> Johan Carlsson,  <sup>u</sup> Tiago F. T. Cerqueira,  <sup>v</sup> Simon Divilov,  <sup>mn</sup> Hamidreza Hajiyani,  <sup>u</sup> Felix Hanke,  <sup>w</sup> Kevin Jose, <sup>x</sup> Corey Oses,  <sup>y</sup> Janosh Riebesell,  <sup>xz</sup> Jonathan Schmidt,  <sup>aa</sup> Donald Winston,  <sup>ab</sup> Christen Xie, <sup>s</sup> Xiaoyu Yang,  <sup>ac ad ae</sup> Sara Bonella,  <sup>c</sup> Silvana Botti,  <sup>af</sup> Stefano Curtarolo,  <sup>mn</sup> Claudia Draxl,  <sup>ag</sup> Luis Edmundo Fuentes Cobas,  <sup>k</sup> Adam Hospital,  <sup>g</sup> Zi-Kui Liu,  <sup>q</sup> Miguel A. L. Marques,  <sup>af</sup> Nicola Marzari,  <sup>oh</sup> Andrew J. Morris,  <sup>ai</sup> Shyue Ping Ong,  <sup>s</sup> Modesto Orozco,  <sup>g</sup> Kristin A. Persson,  <sup>zaj</sup> Kristian S. Thygesen,  <sup>j</sup> Chris Wolverton,  <sup>r</sup> Markus Scheidgen,  <sup>ag</sup> Cormac Toher,  <sup>nak</sup> Gareth J. Conduit,  <sup>xan</sup> Giovanni Pizzi,  <sup>oah</sup> Saulius Gražulis,  <sup>dal</sup> Gian-Marco Rignanese  <sup>\*ab am</sup> and Rickard Armiento  <sup>\*f</sup>

Fekete,  
se ,  
tefano  
ni, Geoffroy  
nehal  
cell,  
Faroquier,

Indian  
Ocean



## RESOURCE REPRESENTATION

## FILTERING GRAMMAR

## PROPERTY DEFINITIONS

## RESOURCE LAYOUT

## PROVIDER FEDERATION

```
{  
  "data": [  
    {  
      "id": "mp-7000",  
      "type": "structures",  
      "attributes": {  
        "immutable_id": "645d2ba4bcd30f748b475981",  
        "last_modified": "2023-03-11T14:56:30Z",  
        "elements": ["O", "Si"],  
        "nelements": 2,  
        "elements_ratios": [0.3333333333333333, 0.6666666666666666],  
        "chemical_formula_descriptive": "O6Si3",  
        "chemical_formula_reduced": "O2Si",  
        "chemical_formula_hill": "O6Si3",  
        "chemical_formula_anonymous": "A2B",  
        "dimension_types": [1, 1, 1],  
        "nperiodic_dimensions": 3,  
        "lattice_vectors": [  
          [4.914966, -1e-8, 0],  
          [-2.45748252, 4.2564861, 0],  
          [0, 0, 5.43130114]  
        ],  
        "nsites": 9,  
        "species_at_sites": ["Si", "Si", "Si", "O", "O", "O", "O", "O", "O"]  
      }  
    }  
  ]  
}
```





```
/v1/structures?filter=elements HAS ANY "C",
"Si", "Ge", "Sn", "Pb"

/v1/structures?filter=elements HAS ANY "C",
"Si", "Ge", "Sn", "Pb" AND nelements=2

/v1/structures?filter=elements HAS ANY "C",
"Si", "Ge", "Sn" AND NOT elements HAS "Pb" AND
elements LENGTH 3
```

## RESOURCE REPRESENTATION

## FILTERING GRAMMAR

## PROPERTY DEFINITIONS

## RESOURCE LAYOUT

## PROVIDER FEDERATION

```
{  
    "$id": "https://schemas.optimade.org/defs/v1.2/properties/optimade_nsites",  
    "$schema": "https://schemas.optimade.org/meta/v1.2/optimade_property_definition.json",  
    "title": "number of sites",  
    "x-optimade-type": "integer",  
    "x-optimade-definition": {  
        "label": "nsites_optimade_structures",  
        "kind": "property",  
        "version": "1.2.0",  
        "format": "1.2",  
        "name": "nsites"  
    },  
    "type": [  
        "integer",  
        "null"  
    ],  
    "description": "An integer specifying the length of the  
    `cartesian_site_positions`  
    property.\n\n**Requirements/Conventions:**  
    - MUST be equal to  
    the lengths of the list properties elements and elements_ratios,  
    if they are provided.  
    **Query examples:**  
    - Match only structures with exactly 4 sites: `nsites=4`  
    - Match structures that have between 2 and 7 sites: `nsites>=2 AND nsites<=7`",  
    "examples": [  
        42  
    ],  
    "x-optimade-unit": "dimensionless"  
}
```





# RESOURCE REPRESENTATION

# FILTERING GRAMMAR

# PROPERTY DEFINITIONS

# RESOURCE LAYOUT

# PROVIDER FEDERATION

**base URL** **endpoint**

`https://optimade.materialsproject.org/v1/structures`

?filter=`chemical_formula_reduced="O2Si"`

**OPTIMADE filter**

## RESOURCE REPRESENTATION

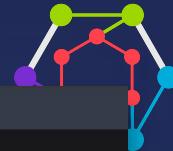
## FILTERING GRAMMAR

## PROPERTY DEFINITIONS

## RESOURCE LAYOUT

## PROVIDER FEDERATION

```
▼ 10:
  type:          "links"
  id:           "mcloudarchive"
  ▼ attributes:
    name:        "Materials Cloud Archive"
    ▼ description:
      base_url:  "https://www.materialscloud.org/optimade/archive"
      homepage:   "https://archive.materialscloud.org"
      link_type:  "external"
  ▼ 11:
  type:          "links"
  id:           "mp"
  ▼ attributes:
    name:        "The Materials Project"
    ▼ description:
      "An open database of computed materials properties to accelerate materials
       discovery and design"
      base_url:  "https://providers.optimade.org/index-metadbs/mp"
      homepage:   "https://www.materialsproject.org"
      link_type:  "external"
```



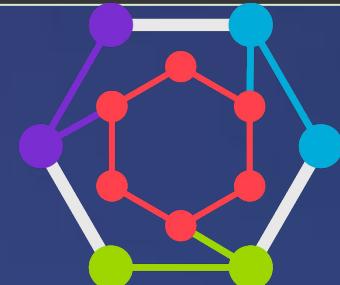
DATA-DRIVEN  
WORKFLOWS

INTEROPERABILITY OF  
REPRESENTATIONS

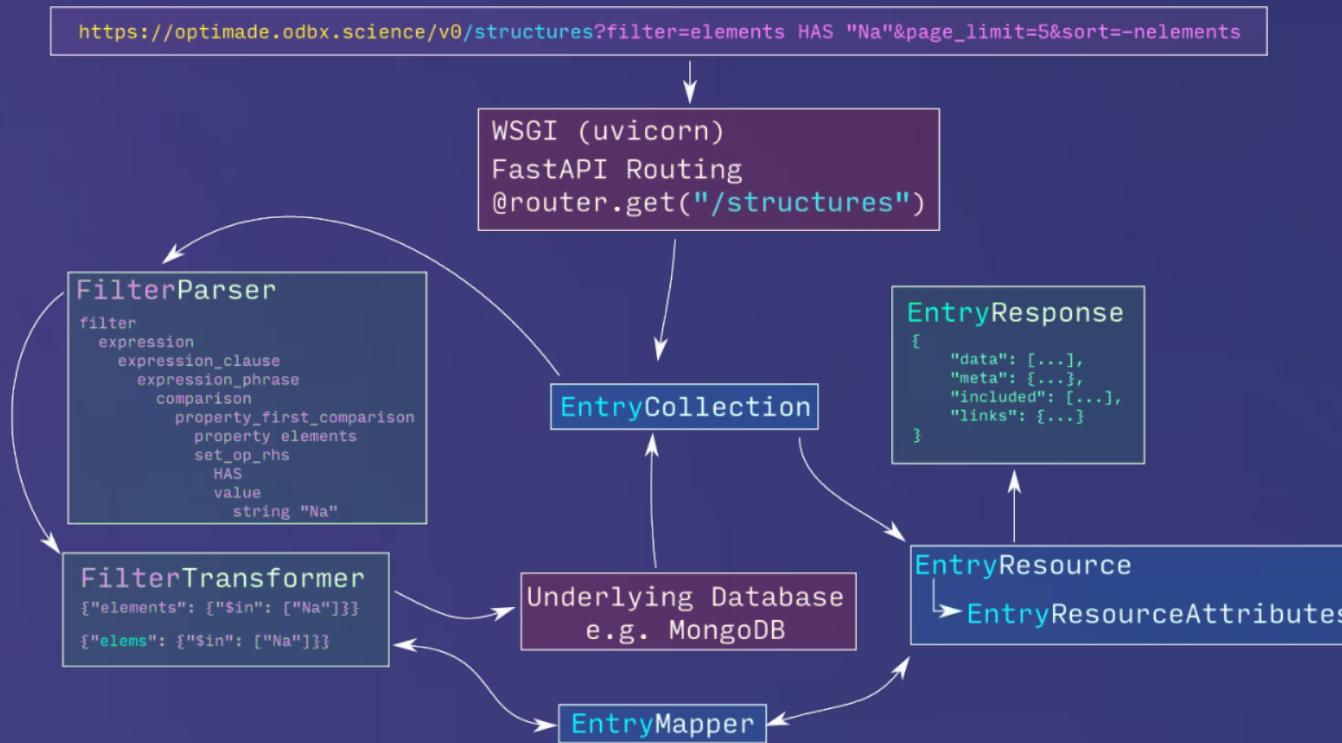
AUTOMATED  
VALIDATION

DATA TO COMPLIANT  
DATABASE PIPELINES

MULTI-PROVIDER  
CLIENTS & QUERIES



# What makes an OPTIMADE API?



# *optimade-python-tools*

An open source Python package for consuming  
and implementing OPTIMADE APIs.



Evans, M. L., Andersen, C. W. A. et al.,  
Journal of Open Source Software, 6(65), 3458, 2021  
[10.21105/joss.03458](https://doi.org/10.21105/joss.03458)



[Materials-Consortia/optimade-python-tools](https://github.com/Materials-Consortia/optimade-python-tools)

- ❖ Spin-up OPTIMADE API with “no code”
  - Built with pydantic and FastAPI
  - Annotated data models with data validation
  - Auto-generated OpenAPI 3.0 and JSONSchema
  - EBNF grammar implementation with filter transformers for MongoDB & Elasticsearch
  - Mappers between existing formats (ASE, pymatgen, CIF) and OPTIMADE, supporting aliases etc.
- ❖ Client for asynchronously querying multiple databases
- ❖ Used by Materials Project, NOMAD, `odbx`, 2DMatPedia and Materials Cloud
- ❖ Provides tools for validating remote implementations

~ ➜ optimade\_optimade-python-tools\_exercise ➜ 05:54:25 ✘ 24s

\$

How can we screen the  
burgeoning number of  
hypothetical materials for  
real applications?

# ML-accelerated high-throughput workflows for optical materials discovery & design: *re2fractive*

*with Victor Trinquet, Cameron Hargreaves, Pierre-Paul De  
Breuck & Gian-Marco Rignanese (MODL-IMCN UCLouvain)*

*MODL-IMCN, UCLouvain*



[modl-uclouvain/re2fractive](https://github.com/modl-uclouvain/re2fractive)

Finding materials with the largest refractive index for a given band gap: *re2fractive*

Leveraging open source tools -- broaden impact by investing in maintenance



60M+  
Hypothetical  
inorganic crystal  
structures\*



Property  
prediction from  
small datasets  
with kitchen sink  
featurisation and  
ensemble  
uncertainties

ATOMATE2

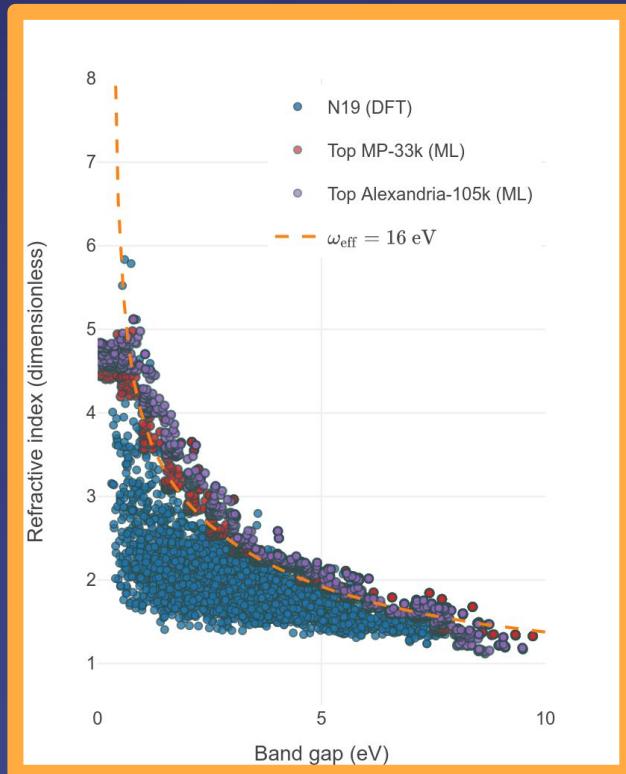
Robust  
automated  
DFT+beyond  
workflows



Automated  
submission and  
job handling

**Publish repeatable workflows**

# Finding materials with the largest refractive index for a given band gap: *re2fractive*



## Active learning loop

- I. Starting from ~4000 materials (Naccarato et al, Phys. Rev. Mat. 2019), train MODnet model for refractive index
- II. Screen databases with appropriate annotations in OPTIMADE (4.2m structures)
- III. Filter OPTIMADE entries for stable materials with non-zero band gap (~150k structures), featurize and predict
- IV. Select a mix of a) most promising and b) highest uncertainty predictions, compute them with DFT
- V. Retrain model with the new data and repeat until Pareto front is sufficiently explored

*"Optical materials discovery and design via federated databases and machine learning"*, Trinquet, Evans et al, Faraday Discussions (2024) [10.1039/D4FD00092G](https://doi.org/10.1039/D4FD00092G)

# Finding materials with the largest refractive index for a given band gap: *re2fractive*

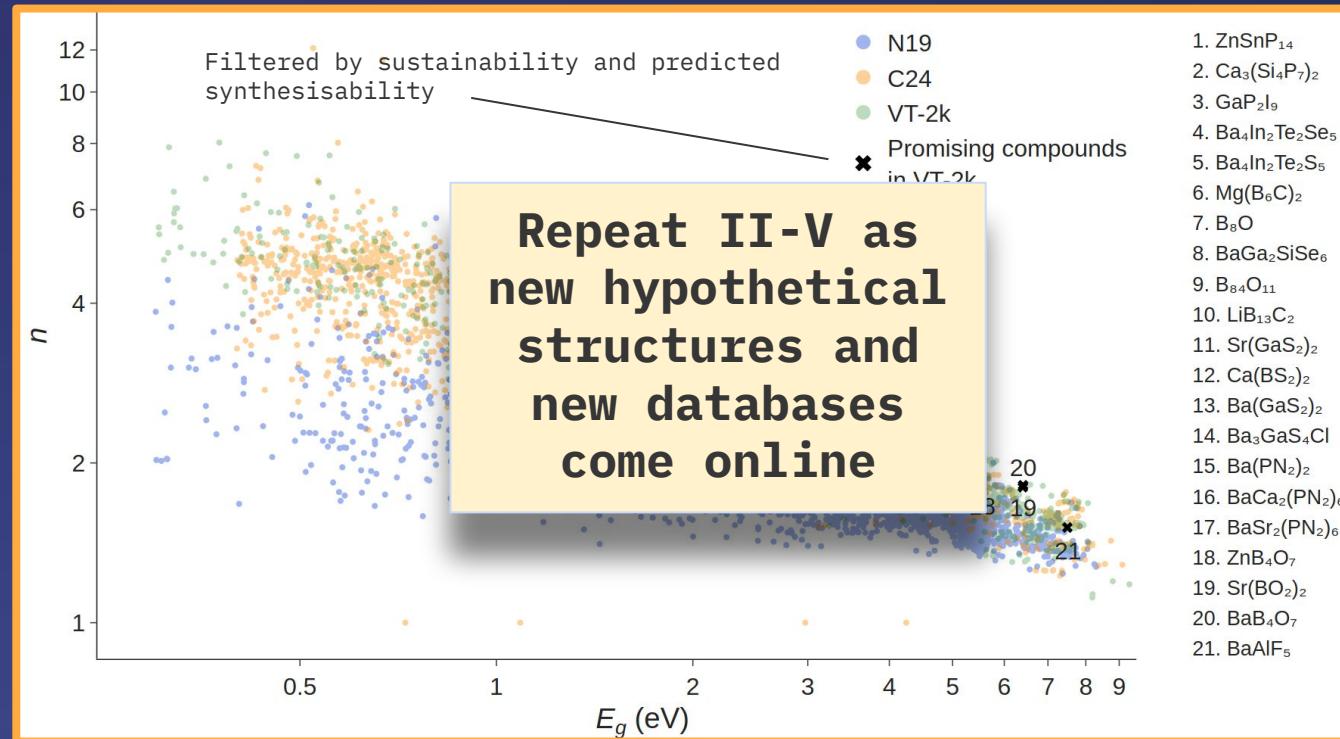
Repeat II-V as new hypothetical structures and new databases come online

Design with this in mind!

- I. Starting from ~4000 materials (Naccarato et al, Phys. Rev. Mat. 2019), train MODnet model for refractive index (#1 matbench leaderboard for dielectric task)
- II. Screen databases with appropriate annotations in OPTIMADE (4.2m structures)
- III. Filter OPTIMADE entries for stable materials with non-zero band gap (~150k structures), featurize and predict
- IV. Select a mix of a) most promising and b) highest uncertainty predictions, compute them with DFT
- V. Retrain model with the new data and repeat until Pareto front is sufficiently explored

*"Optical materials discovery and design via federated databases and machine learning"*, Trinquet, Evans et al, Faraday Discussions (2024) [10.1039/D4FD00092G](https://doi.org/10.1039/D4FD00092G)

# Finding materials with the largest refractive index for a given band gap: *re2fractive*



# Why stop at refractive index?

- Decentralised AL workflows scanning OPTIMADE databases daily for potential new compounds
- Created datasets themselves are then hosted via OPTIMADE, tracking provenance of initial structures
- Self-reinforcing loops for multi-fidelity property prediction can help prioritise new compounds for synthesis
- AL especially necessary when properties are rare + expensive to compute
- We have initialised a new campaign for non-linear optical response

Data Descriptor | [Open access](#) | Published: 11 July 2024

## Second-harmonic generation tensors from high-throughput density-functional perturbation theory

[Victor Trinquet](#) , [Francesco Naccarato](#), [Guillaume Brunin](#), [Guido Petretto](#), [Ludger Wirtz](#), [Geoffroy Hautier](#) & [Gian-Marco Rignanese](#) 

[Scientific Data](#) 11, Article number: 757 (2024) | [Cite this article](#)

# *optimade-maker*

with Kristjan Eimre, & Giovanni Pizzi

## From static data to an API

- Enhancing discoverability, explorability and expressiveness of structural data on the MCloud
- User annotates their MCloud entry with a simple config file
- Additional services watch for such entries and run a processing pipeline to convert to our OPTIMADE JSONL format
- Container launches to expose the structural data and properties as public OPTIMADE APIs

```
database_description: >-
  This database contains some example CIFs.

data_paths:
  - structures.zip
  - data.tar.gz

entries:
  - entry_type: structures
    entry_paths:
      - structures/cifs/*.cif
    property_paths:
      - data/data.csv
      - data/data2.csv
    property_definitions:
      - name: energy
        title: Total energy per atom
        description: The total energy per atom as computed by DFT
        unit: eV/atom
        type: float
      - name: property_b
        title: Property B
        description: Alias for some more complicated property_b
        maps_to: https://schemas.optimade.org/v1.1/dft/dft_hull_distance
      - name: description
        title: Entry description
        description: Provides a human-readable description for this particular entry_type
        type: str
    # - entry_type: references
    #entry_paths:
    #  - refs.bib
```



[materialscloud.org/optimade-maker](https://materialscloud.org/optimade-maker)

# *optimade-maker*

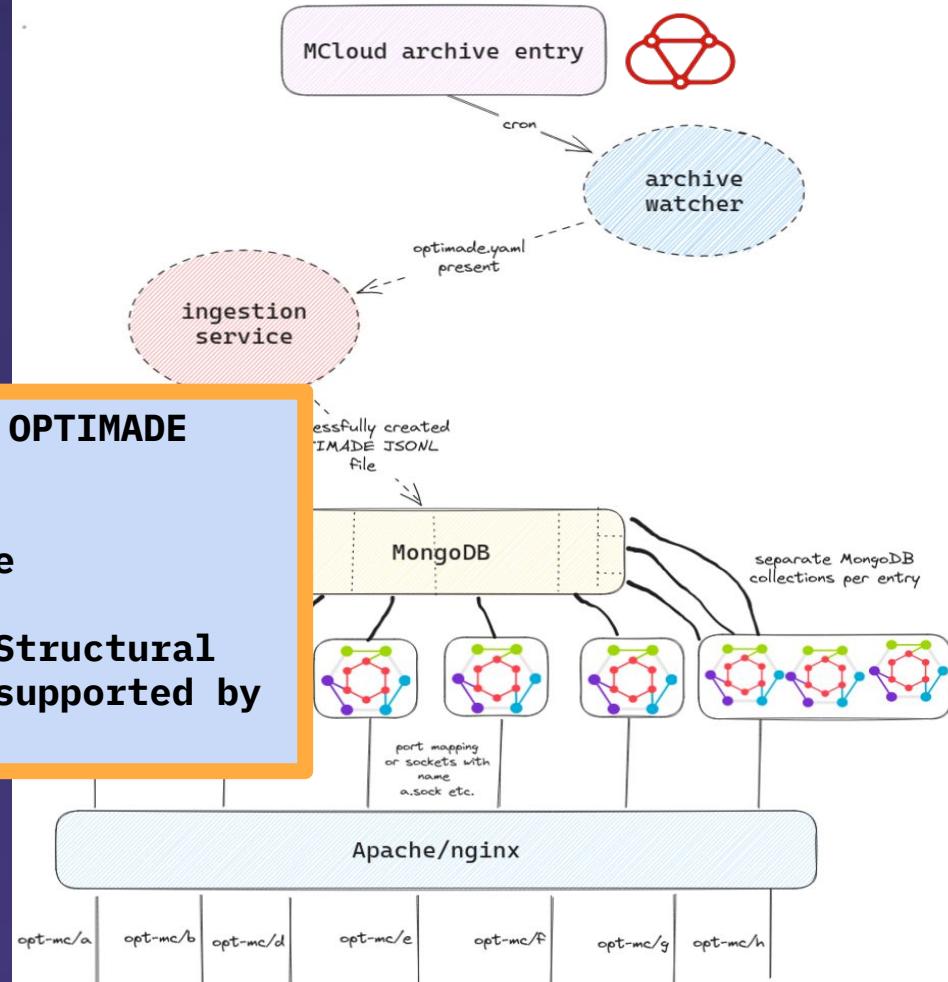
with Kristjan Eimre, & Giovanni Pizzi

## From static data to an API

- Enhancing discoverability, explorability and expressiveness of structural data in the MCloud
- User annotates their MCloud data with a simple config file
- Additional services watch for changes and run a processing pipeline to our OPTIMADE JSONL format
- Container launches to expose the structural data and properties as public OPTIMADE APIs

**Used to create OPTIMADE APIs for:**

- **Alexandria**
- **re2fractive**
- **GNome**
- **Cambridge Structural Database (supported by PSDI)**



[materialscloud.org/optimade-maker](https://materialscloud.org/optimade-maker)

**Well-defined  
abstraction: crystal  
structure**

**Flat(ish) data  
structure**

**Extensible property  
definitions**

**Well-defined search**

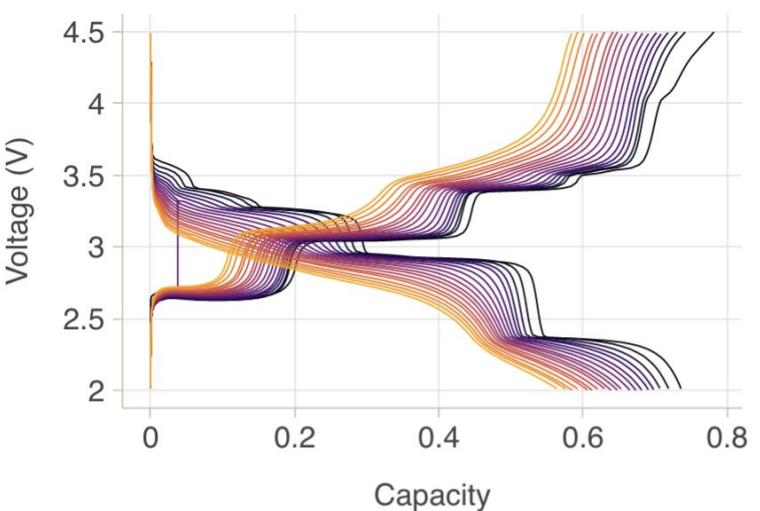
**Tools for federation**

What about  
experimental  
data?

# Data management tech tree



# Managing experimental lab data



What information would need to be recorded to make this information useful after the student leaves?

What information would need to be recorded to make this useful to a machine learning algorithm?

What type of cell is this? What are the electrodes? What batch are they from? Active mass? Where was it cycled? What instrument? Cycling parameters? Surface area? When was it made? Who made it? When? What batch of electrolyte? Was it made in a glovebox? Which one? Have other measurements been made on this cell? Has this result been repeated with other cells? Is there any characterization on the electrodes? On the active material? On the electrolyte? What was the temperature of the room? Were there any spikes? What does the  $dV/dQ$  look like? Who "owns" this data?

# Experimental data is only useful in its context

Na<sub>2</sub>O<sub>2</sub> from ECO glovebox

NiO (from collaborator)

Jmas-1-4a

Jmas-1-4b

Jmas-1-4c

Jmas-1-4d

Super P

PVDF

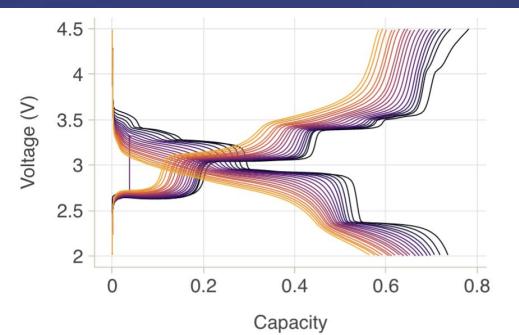
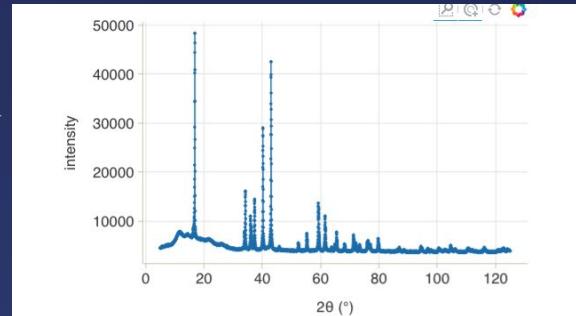
NMP

Jmas-1-4a\_e1

NaPF6

PC

electrolyte



Jmas1-4a-e1-c1

Jmas1-4a-e1-c2

Jmas1-4a-e1-c3

Na metal

# Initial motivations: *datalab*



[datalab-org/datalab](https://github.com/datalab-org/datalab)

- **Automating tedious data handling processes**
- **Reproducible, robust science**
- **Enhanced dissemination**
- **Putting the FUN in data management**
- **Decentralized data unification**



Prof Joshua  
Bocarsly  
(Cambridge ✈  
UHouston)



Josh Bocarsly  Logout

About | [Samples](#) | Collections | Inventory |  Graph View

Add an item 

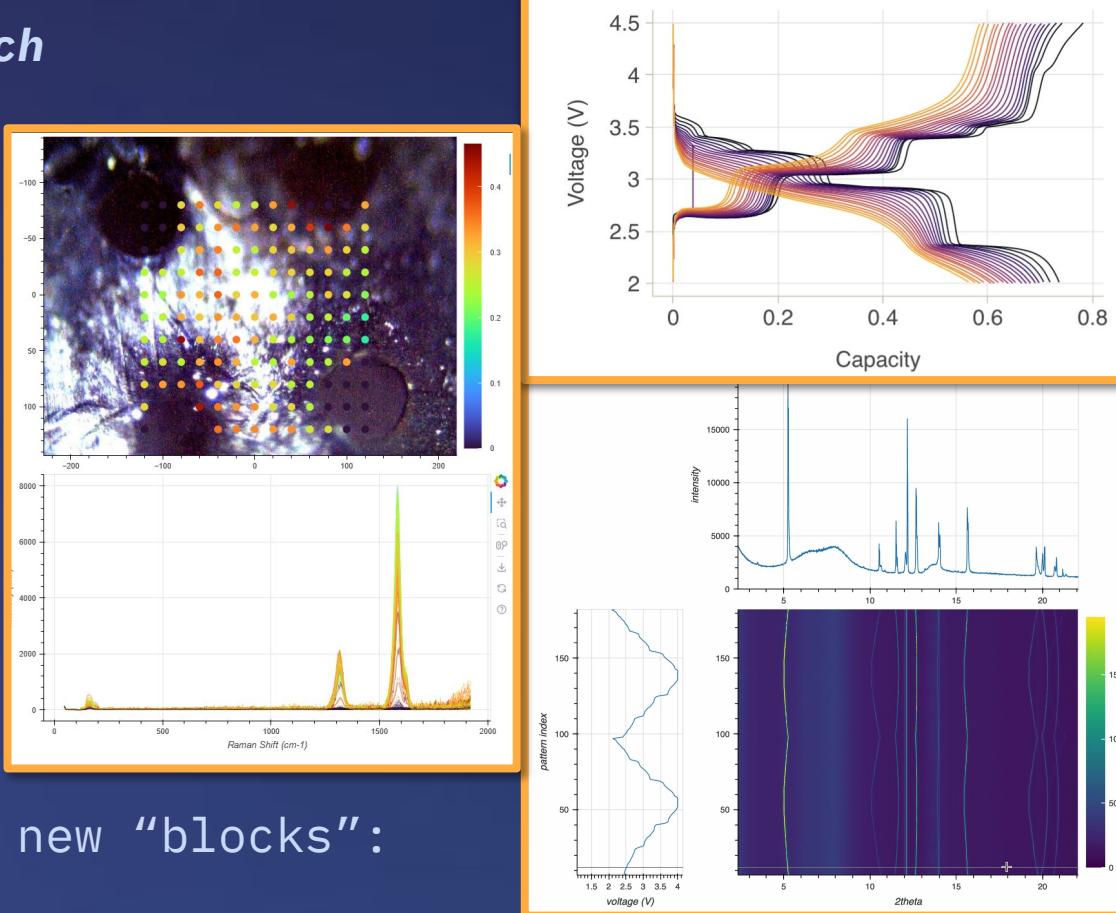
Add batch of samples

ID	Type	Sample name	Formula	Date	Collections	Creators	# of blocks
cell_13b	sample			2023-09-12			1 
test-02	cell	test cell		2023-09-08			2 
test-01	sample	test		2023-09-08			1 
jdb1003	sample	a sample of sodium solution		2023-09-07			1 
Ni3Al_1	sample	solid state sample of Ni-Al		2023-09-05			1 
jdb_test1	sample	NaCoO2 made from solid state synthesis		2023-09-05			1 
mp2028_T15	cell	Titration SEI 15 cell		2023-09-05			1 
mp2028_T14	cell	Titration SEI T14 cell		2023-09-05			1 
mp2028_T13	cell	Titration SEI T13 cell		2023-09-05			1 
Umi-33_PTFE_SP_38-14-48_SSF-OC1	cell	Umi-33_PTFE_SP_38-14-48_SSF-OC1		2023-09-03			0 
Umi-34_PTFE_SP_38-14-48_SSF-cell1	cell	Umi-34_PTFE_SP_38-14-48_SSF-cell1		2023-09-03			0 
Umi-33_PTFE_SP_38-14-48_SSF-cell1	cell	Umi-33_PTFE_SP_38-14-48_SSF-cell1		2023-09-03			0 

# **Modular “Data block” approach**

Currently, we support:

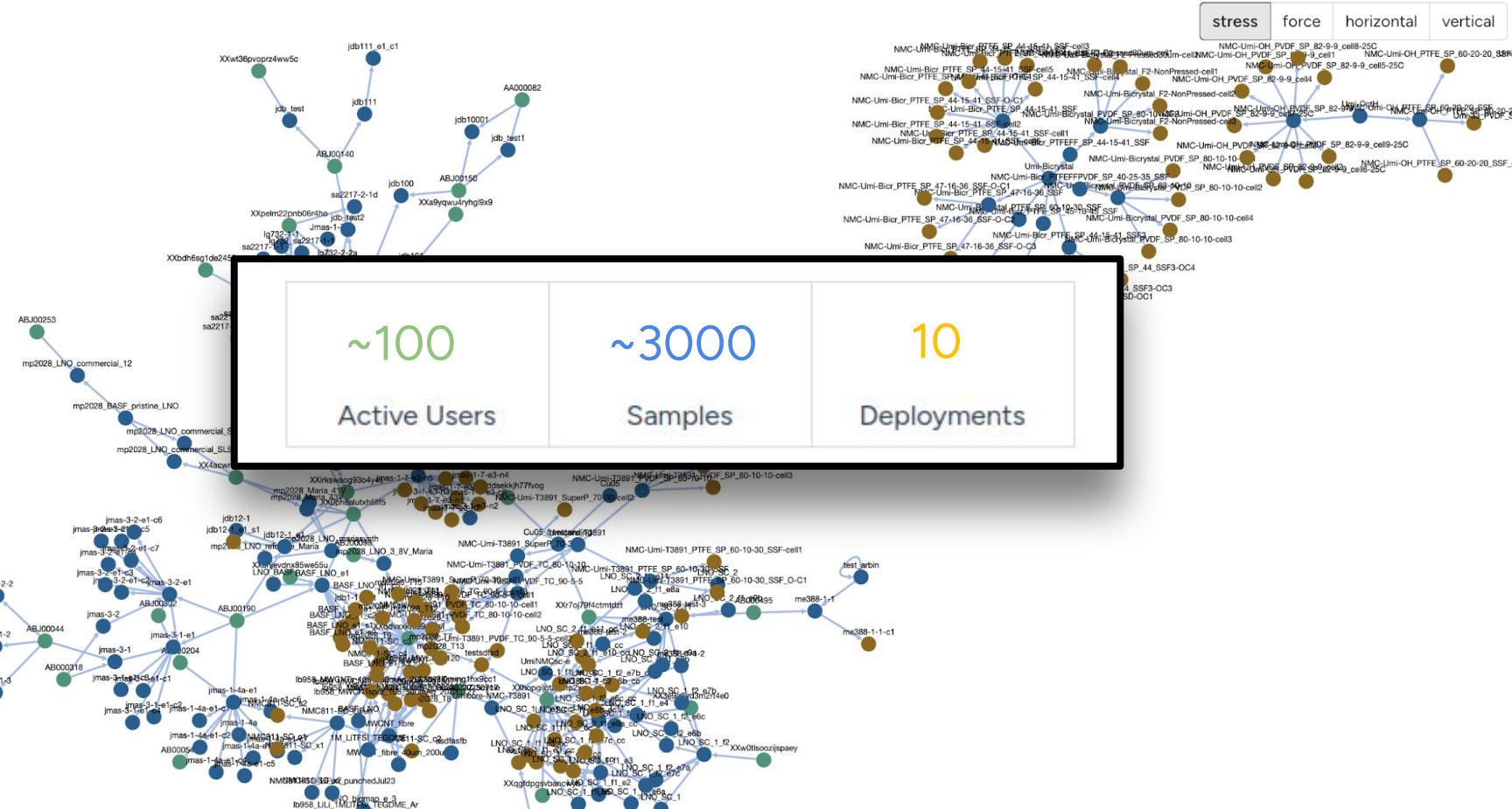
- Electrochemical cycling
- EIS
- Materials synthesis
- Cell preparation
- Images (common + raw)
- Videos
- Powder X-ray diffraction
- Raman (1D + 2D maps)
- NMR
- Mass spectrometry/TGA
- Your method?



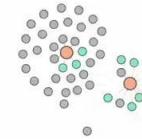
Purposefully simple to add new “blocks”:

- File loader
- Interactive plot
- Baseline corrections, background subtraction, schemas, models, validation, search, comparative analysis...

**Grey Group Datalab pilot: connected, contextualised data from ~5 users in ~3 months**

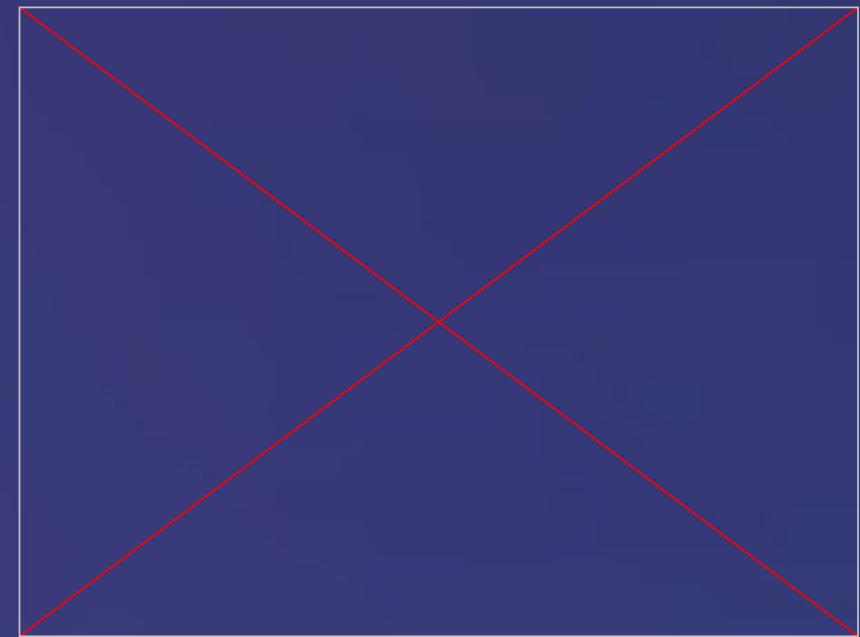


# Growth of the Bocarsly group data



## ***datalab* is open-source software (MIT)**

- Collaborative development is performed on GitHub using modern best practices for open source software development.
  - ◆ ~30,000 lines of code
  - ◆ 439 merged PRs
- Automated CI with testing of server and GUI
- Simplified deployment via Docker
- Automated cloud deployments with Ansible & Terraform
- Decentralizing maintenance costs via the community



*[github.com/datalab-org/datalab](https://github.com/datalab-org/datalab)*

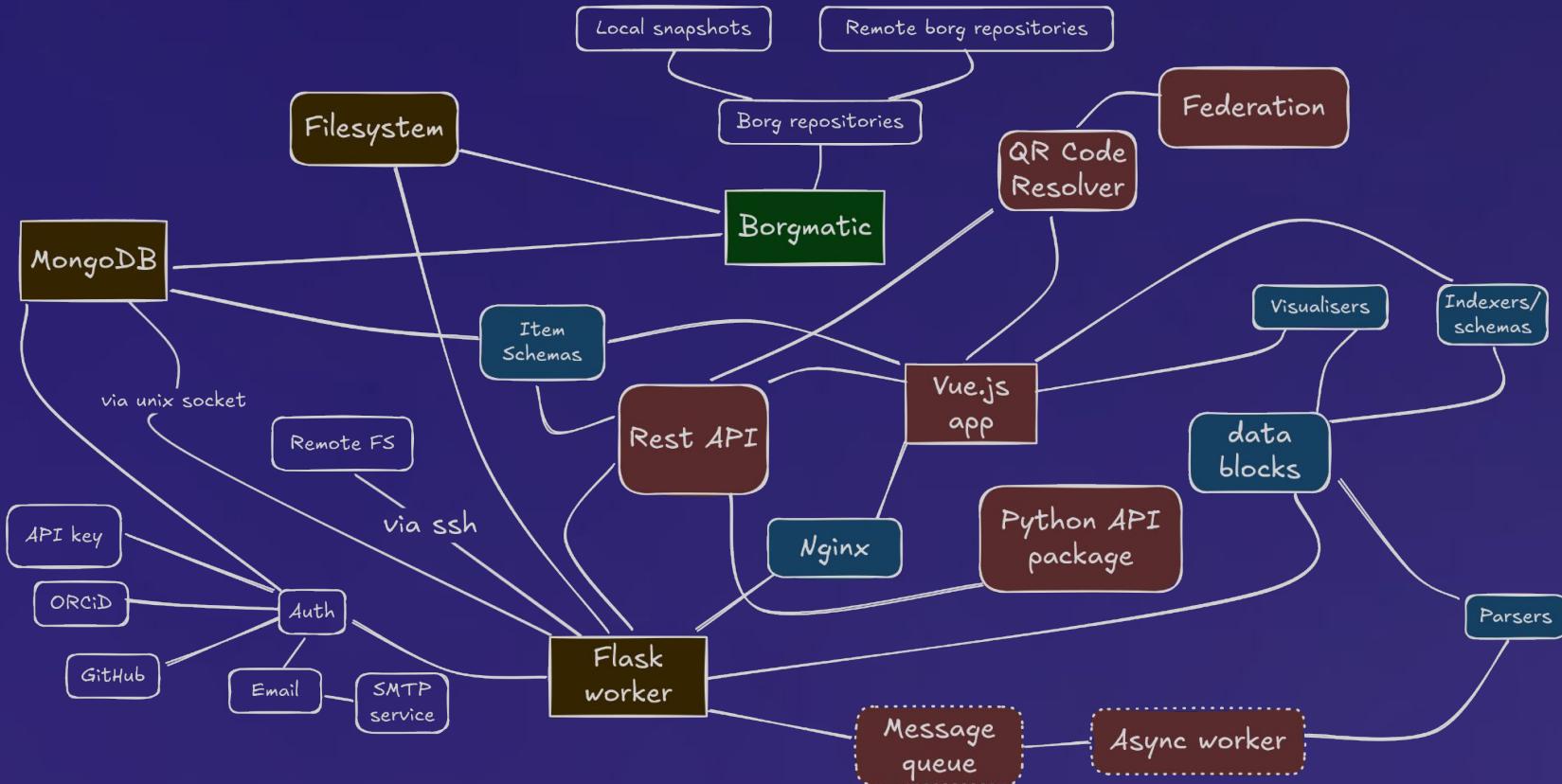
# *datalab* is open-source software (MIT)

- Collaborative development is performed on GitHub using modern best practices for open source software development.
  - ◆ ~30,000 lines of code
  - ◆ 439 merged PRs
- Automated CI with testing of server and GUI
- Simplified containerized deployment via Docker (instructions online)
- Automated cloud deployments with Ansible & Terraform

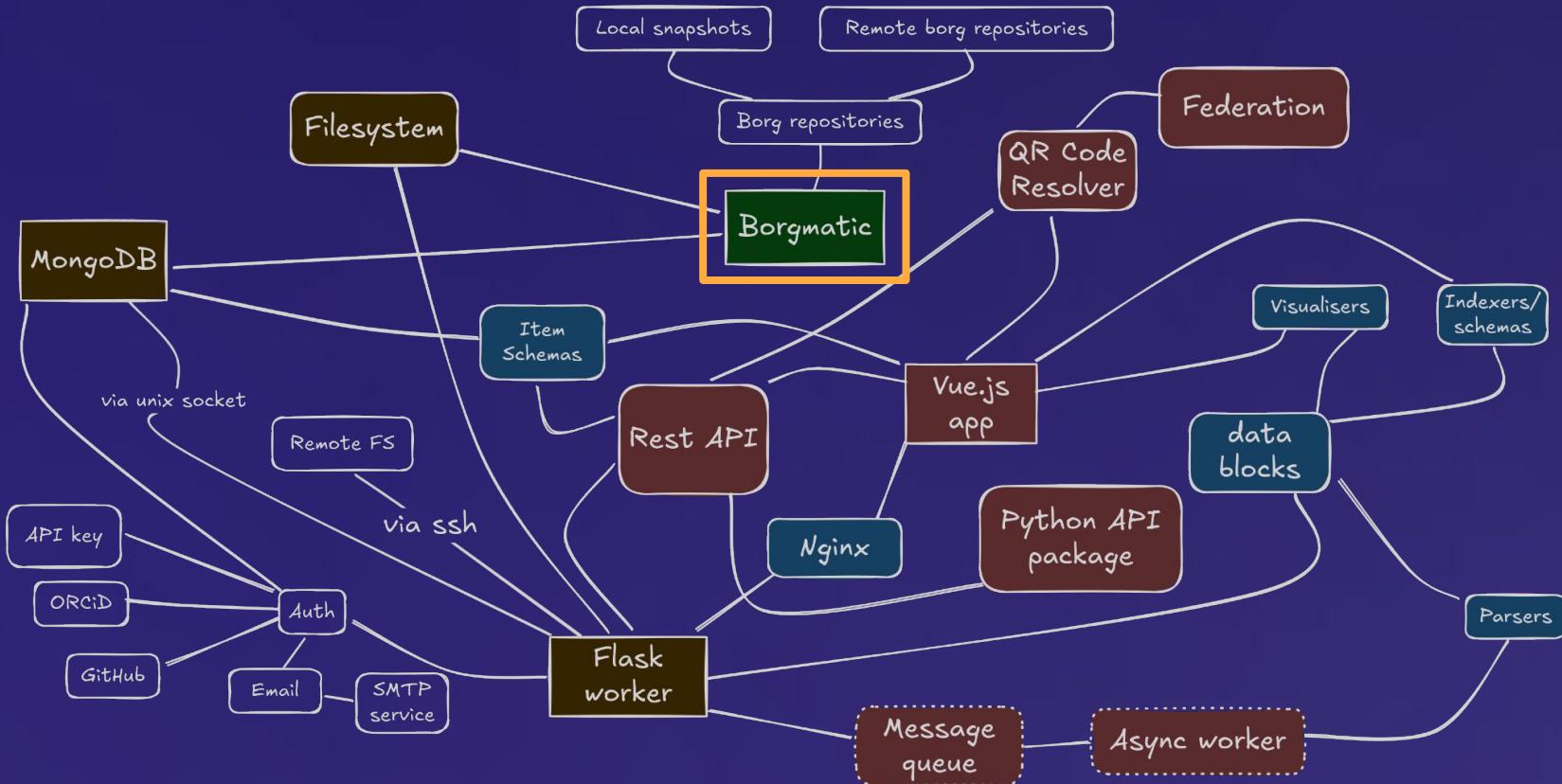
The screenshot shows the GitHub Issues page for the 'datalab' repository. The URL at the bottom is [github.com/datalab-org/datalab](https://github.com/datalab-org/datalab). The page displays 154 open issues and 167 closed issues. The search bar at the top contains the query 'is:issue is:open'. The issues listed include:

- Handling 2D Raman maps (#781) - 1 of 6 tasks
- Add landing page to documentation (documentation)
- Login/register button should not render until the /get-current-user/ call has completed (polish)
- Suggestion: Replace default unauthenticated page with a login splash screen (suggestions)
- Copying from existing entry when creating new equipment does nothing (bug)
- Missing features for Equipment type (schema) (Suggestions)
- Status field for items and equipment (suggestions)
- Suggestion: Gas chromatography block (datablock) (suggestions)

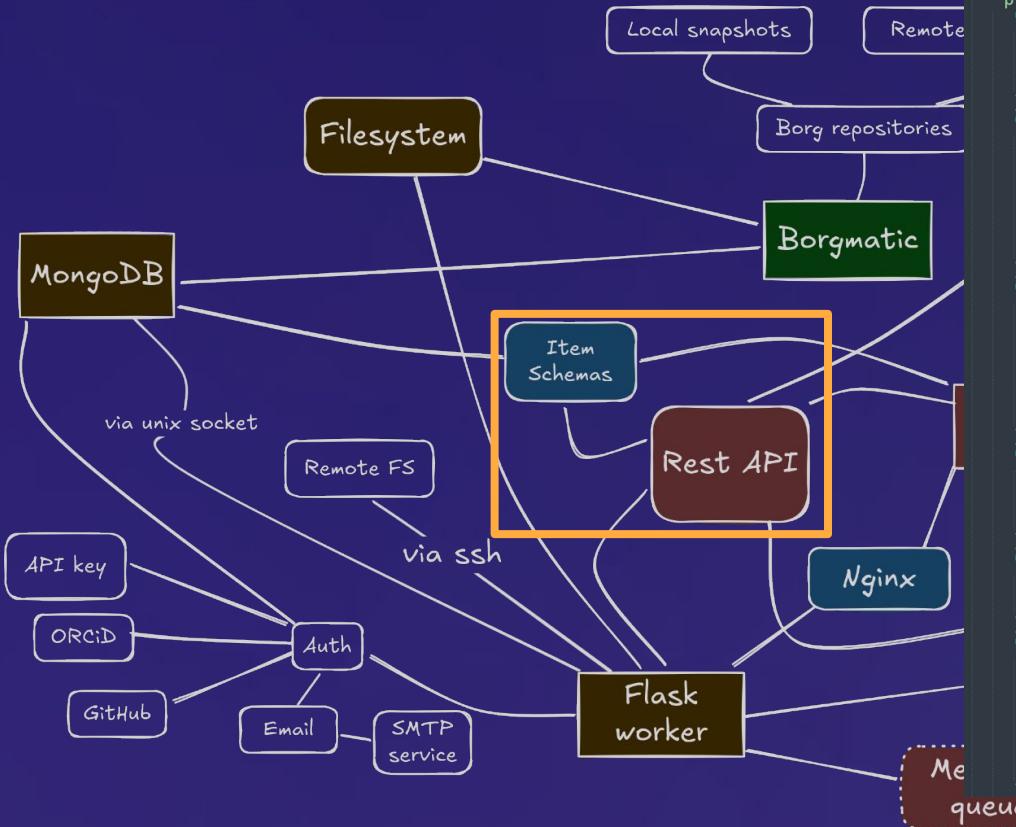
# *datalab* Architecture



# *datalab* Architecture

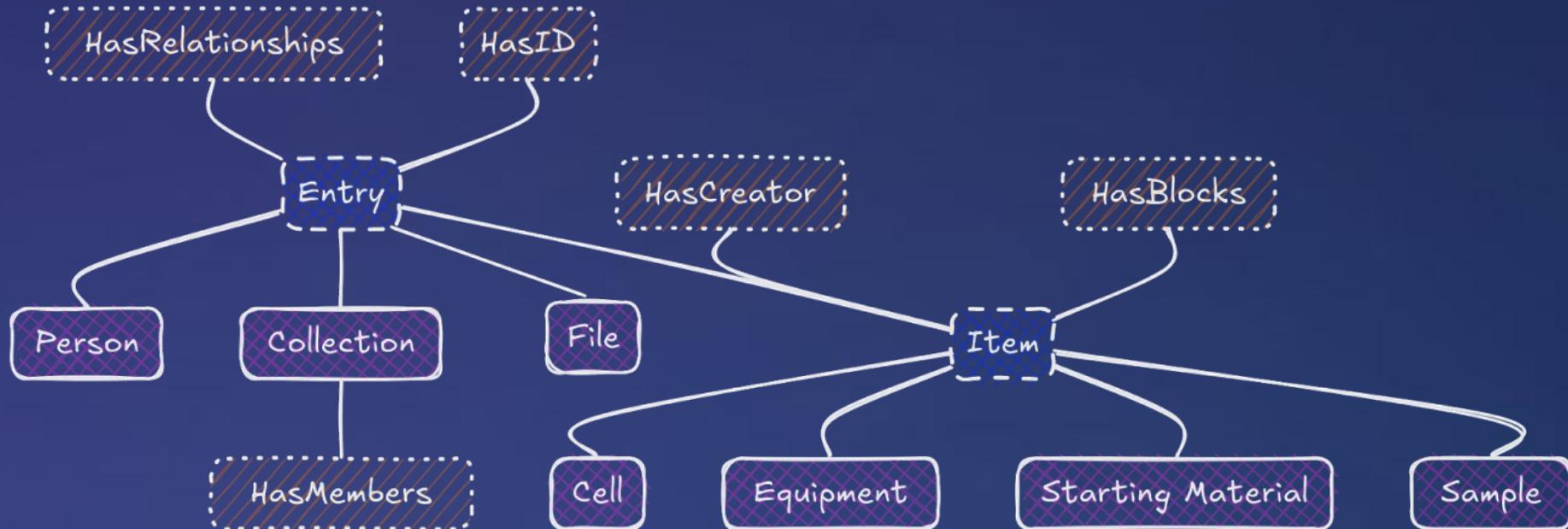


# *datalab* Architecture

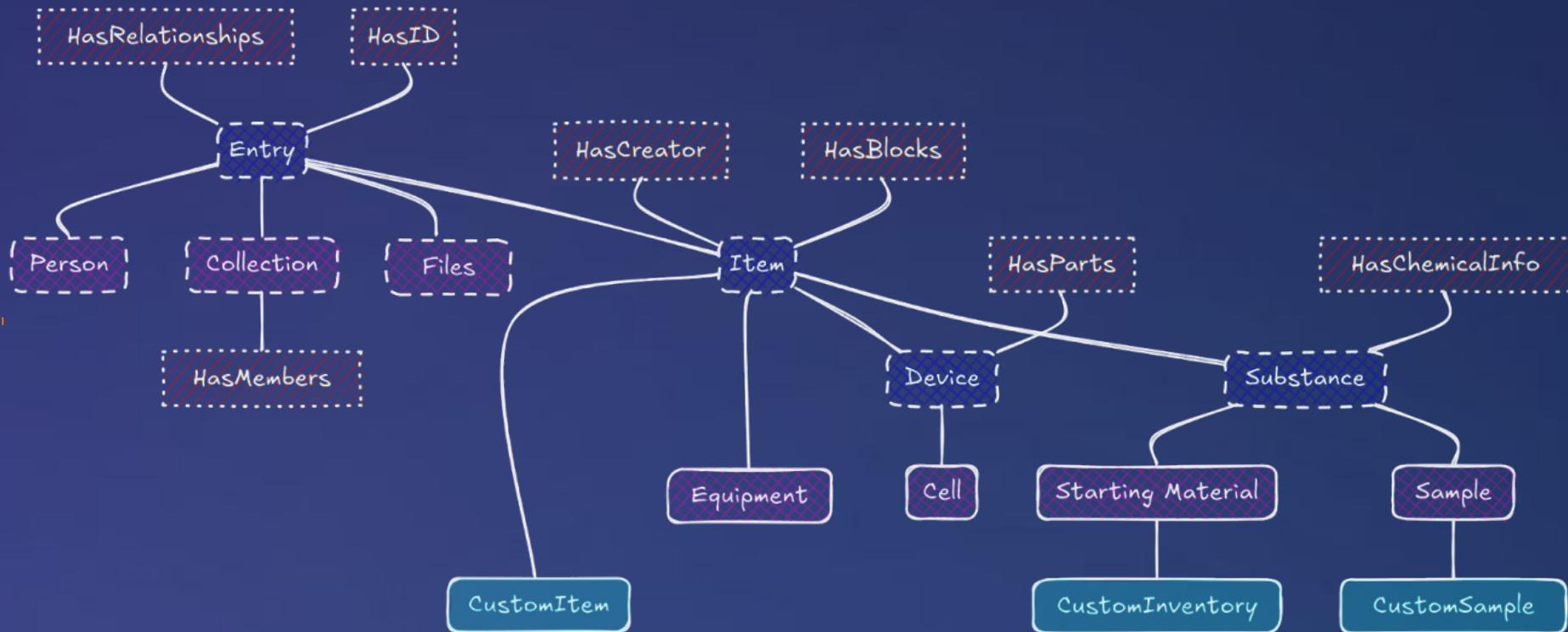


```
{  
  "title": "Sample",  
  "description": "A model for representing an experimental sample.",  
  "type": "object",  
  "properties": {  
    "blocks_obj": {  
      "title": "Blocks Obj",  
      "default": {},  
      "type": "object"  
    },  
    "display_order": {  
      "title": "Display Order",  
      "default": [],  
      "type": "array",  
      "items": {  
        "type": "string"  
      }  
    },  
    "collections": {  
      "title": "Collections",  
      "default": [],  
      "type": "array",  
      "items": {  
        "$ref": "#/definitions/Collection"  
      }  
    },  
    "revision": {  
      "title": "Revision",  
      "default": 1,  
      "type": "integer"  
    },  
    "revisions": {  
      "title": "Revisions",  
      "type": "object"  
    },  
    "creator_ids": {  
      "title": "Creator Ids",  
      "default": [],  
      "type": "array",  
      "items": {  
        "type": "string"  
      }  
    }  
}
```

# *datalab* data model: traits-based

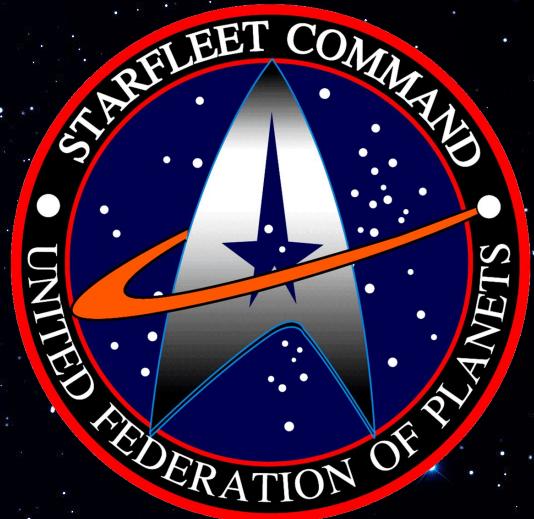
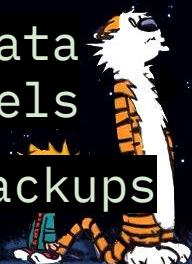


# *datalab* data model: custom schemas



# *datalab* Federation

- Share and discover data between collaborators
- Share analysis workflows & schemas
- Share infrastructure: QR codes and persistent labelling
  - Decentralize & preserve safety information
- Private machine learning
  - opt-in with automatic data credit to train new models
- Robust offsite encrypted backups



- Grey group (Cambridge)
- Bocarsly group (UHouston)
- Cliffe group (Nottingham)
- Titirici group (Imperial)
- ConCat lab (TUBerlin)
- 20+ person startup (USA)
- Menkin group (Cambridge)
- Materials Research Lab (UCSB)

# Whinchat

- Chat interface to *datalab* (LLM hackathon 2023)
  - Stuff structured data from a collection of samples into context window of an LLM, then interrogate



Jablonka et al, Digital Discovery, 2023, 2

**Model:** Claude-3-5-socket-20241032

---

**Current conversation token count:** 22299/2000000  
**est. cost for next message:** \$0.079

**temperature:**

---

**system prompt:**  
 You are **whetcharhat** (Downescale w), a virtual data management assistant that helps materials chemists manage their experimental data and plan experiments. You are deployed in the group of Professor Clare Grey in the Department of Chemistry at the University of Cambridge. You are embedded within the program distalab, where you have access to JSON describing an item, or a collection of items, with connections to other items. These items may include experimental samples, starting materials, and devices (e.g. battery cells made out of experimental samples and starting materials). Answer questions in markdown. Specify the language for all markdown code blocks. You can make diagrams by writing a `mermaid` code block or an `svg` code block. When writing `mermaid` code, you must use quotations around each of the labels (e.g. `graph TD; A["label1"] --> B["label2"]`) as concise as possible. When saying your name, type a bird emoji right after **whetcharhat**.

Here is the JSON data for the current item, KPI-1. If the item is in a collection, the data for the rest of the items in the collection are also given. [collections:[]][removable\_id:1]kpi-1

characteristic\_mole\_weight: 107.02; positive\_electrode: Li|Pm<sub>x</sub>|Pyne  
 samples\_name: SODIUM iron(II) oxide electrodes; negative\_electrode: KO<sub>1-x</sub>Li<sub>x</sub>  
 chemform: Na<sub>0.74</sub>Co<sub>0.26</sub>, quantity: 10.4 mg (g); negative\_electrode  
 [Item Type: starting\_materials; name: Sodium; item\_id: V000242  
 chemform: Na; quantity: 10.0 mg (g); electrolyte: LiClO<sub>4</sub>  
 samples\_name: Li|Pm<sub>x</sub>|Pyne in PC, Na-battery electrolyte, thermal:  
 KO<sub>1-x</sub>Li<sub>x</sub>; chemform: NaPF<sub>6</sub>, quantity: 10.0 mg (g);  
 active\_ion\_charge: 1.0; blocktype: 1|blocktype: 2|media: title: Media; Net  
 K<sub>0.74</sub>Li<sub>0.26</sub>(Co<sub>0.26</sub>Fe<sub>0.74</sub>)<sub>x</sub>; blocktype: media; title: Media; Net  
 KO<sub>1-x</sub>Li<sub>x</sub>(Co<sub>0.26</sub>Fe<sub>0.74</sub>)<sub>x</sub>; blocktype: cycle\_derivative; model: null  
 testroom\_comment: <br><br>very many cycling, initial capacity 80 mAh/g,  
 but fading quickly. <br> - Dspine: 5, Suprise: 5, title: Electrochemical  
 properties of the new sodium-iron(II) oxide battery

# Whinchat

- Chat interface to *datalab* (LLM hackathon 2023)
- Stuff structured data from a collection of samples into context window of an LLM, then interrogate



Hello! I'm whinchat 🦜, your virtual data management assistant. It looks like you've provided a comprehensive JSON dataset for a series of sodium cobalt oxide ( $\text{NaCoO}_2$ ) coin cells and related samples, all made to test their electrochemical performance.

[raw](#) | [formatted](#)

In these experiments, is there a relationship between the battery performance and the synthesis or processing conditions?

[raw](#) | [formatted](#)

whinchat 🦜 here. Let me analyze the relationships between synthesis and performance:

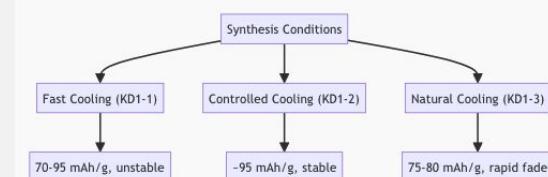
The samples (KD1-1, KD1-2, KD1-3) were all synthesized similarly at 850°C under  $\text{O}_2$ , but with some key differences:

- KD1-1: 20 hours dwell, fast cooling (4 hours)
- KD1-2: 24 hours dwell, controlled cooling (5°C/min)
- KD1-3: 24 hours dwell, natural cooling (15 hours)

Looking at cell performance:

- KD1-1 cells: ~70-95 mAh/g, significant degradation
- KD1-2 cells: Best performer with cell KD1-2\_e1\_c3 showing ~95 mAh/g with "reasonably stable cycling"
- KD1-3 cells: 75-80 mAh/g, all showing rapid capacity fade

The trend suggests that the controlled cooling rate (KD1-2) produced the most stable material, while both fast cooling (KD1-1) and very slow cooling (KD1-3) resulted in materials with poorer cycling stability. This could be related to the formation of optimal crystal structure or sodium content during the cooling process.



[raw](#) | [formatted](#)

# YeLlowhaMMer

[datalab.org/yellowhammer](https://datalab.org/yellowhammer) (with Josh Bocarsly & Yue Wu)

- Agentic interface to *datalab* -- still chat driven, but now uses our Python API directly (LLM hackathon 2024)
- Can automate complicated (not complex) data analysis tasks, e.g.
  - Search/RAG
  - Writing tools that we don't have time for
  - Multimodal/historic data input from spreadsheets, physical lab notebooks or even video
- Saves us writing UI workflows for complicated features!

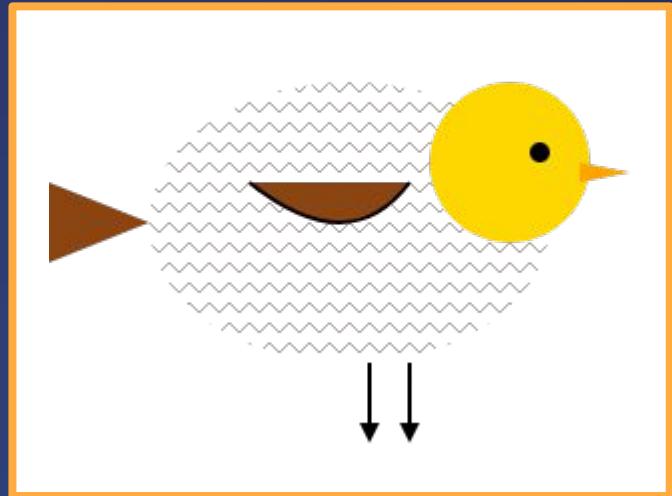


Zimmerman *et al* [arXiv:2411.15221](https://arxiv.org/abs/2411.15221) (2024)

# YeLlowhaMMer

[datalab.org/yellowhammer](https://datalab.org/yellowhammer) (with Josh Bocarsly & Yue Wu)

- Agentic interface to *datalab* -- still chat driven, but now uses our Python API directly (LLM hackathon 2024)
- Can automate complicated (not complex) data analysis tasks, e.g.
  - Search/RAG
  - Writing tools that we don't have time for
  - Multimodal/historic data input from spreadsheets, physical lab notebooks or even video
- Saves us writing UI workflows for complicated features!

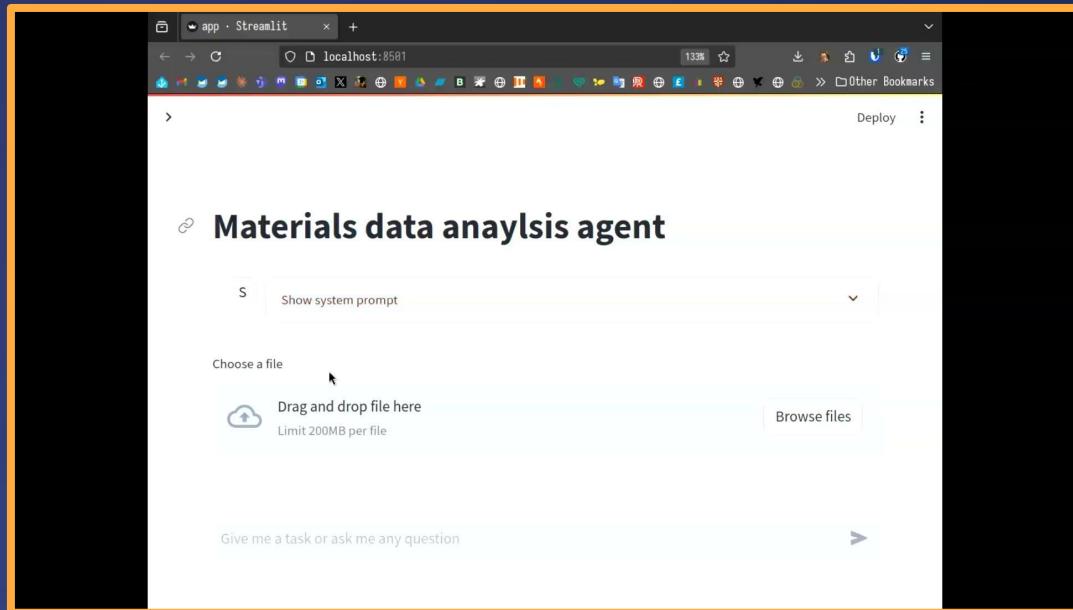


Zimmerman *et al* [arXiv:2411.15221](https://arxiv.org/abs/2411.15221) (2024)

# YeLlowhaMMer

[datalab.org/yellowhammer](https://datalab.org/yellowhammer) (with Josh Bocarsly & Yue Wu)

- Agentic interface to *datalab* -- still chat driven, but now uses our Python API directly (LLM hackathon 2024)
- Can automate complicated (not complex) data analysis tasks, e.g.
  - Search/RAG
  - Writing tools that we don't have time for
  - Multimodal/historic data input from spreadsheets, physical lab notebooks or even video
- Jupyter "magic" `%%llm` interface reduces complexity



# YeLLowhaMMer

[datalab.org/yellowhammer](https://datalab.org/yellowhammer) (with Josh Bocarsly & Yue Wu)

- Jupyter “magic” `%%llm` interface reduces complexity
- Can make sure to track the provenance of AI-generated data for human inspection
- UI components for guiding the AI through common workflows
- Designing our tools for AI:
  - /llms.txt and agent-readable documentation that is always up to date
  - API schemas and introspective endpoints useful for tool-calling
  - Expose in decentralized way?



# *datalab* Roadmap + sustainability

- Plugin ecosystem
- Deployment-level schema customisation
- Sidecar Jupyter “Hub”
- Secure some sustainable funding (<https://datalab.industries> & <https://matgenix.com>)
- Proper/automatic exports to archive servers and integrations with other ELNs
- Dashboards of live experiments
- ~~Ingesting historic data~~
- Integration with lab “knowledge bases”
  - Wiki, Scheduling, Slack
- ~~Advanced chat block:~~
  - ~~Custom plots and API calls~~
- **Deep search via OPTIMADE**
- Integrating ML models fine-tuned on “private” data for e.g., DoE
  - Mroz *et al.*, Web-B0 Faraday Discussions (2024)

Decentralisation means we  
don't have to all use the  
same **tools** or **schemas**

but agreeing on **protocols**  
can be very powerful

# Datatractor (MaRDA extractors working group)

w/*Peter Kraus* (TUBerlin) & *David Elbert* (JHU)

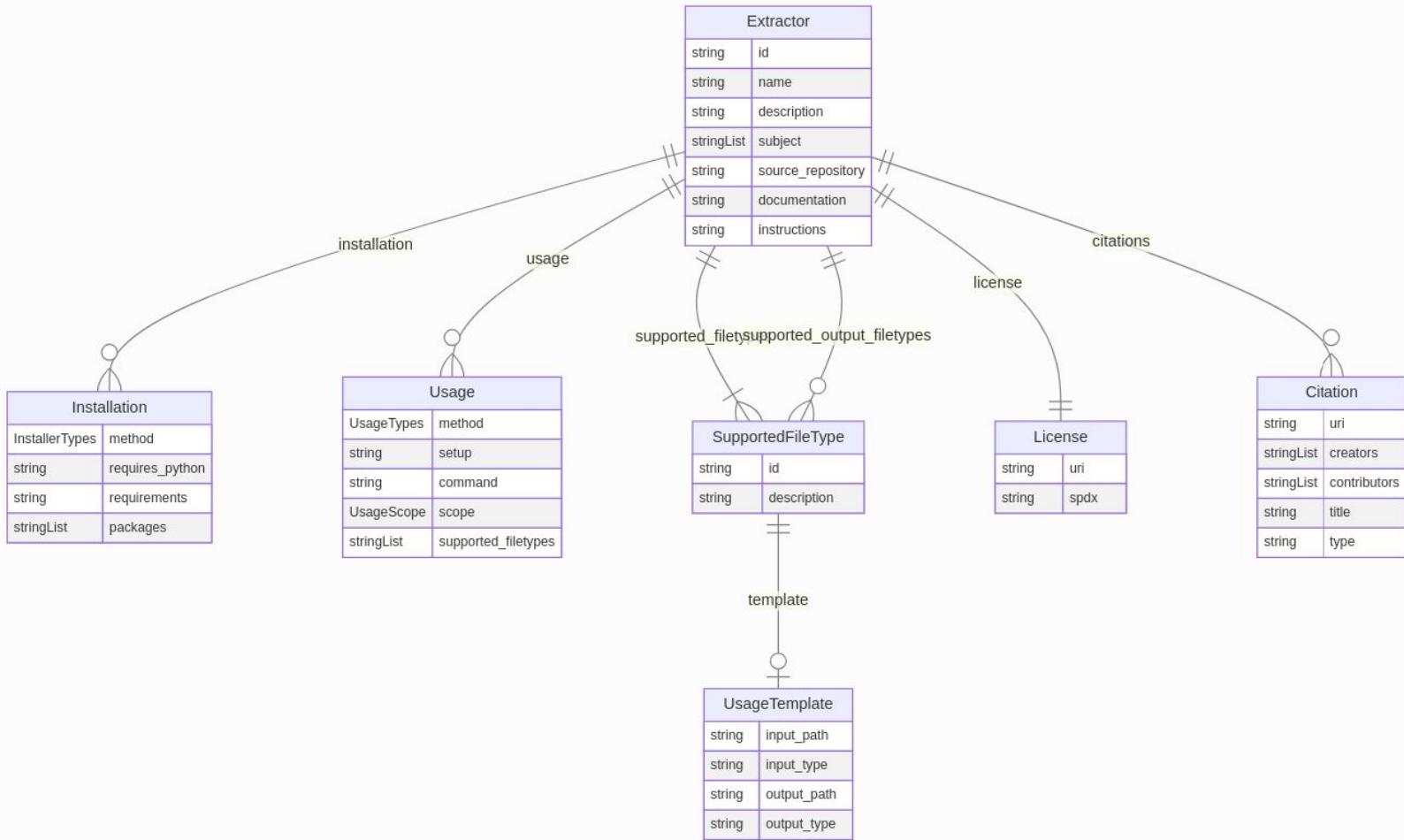
Supporting all possible instrument file format types in a single platform/framework is not possible

How can we communalize or decentralize this work?



- Common schema for describing file types and extractor code
- Machine-actionable definitions for extractor installation & usage
- Online registry of such file types and extractors
- API package that can automatically install and extract





# Datatractor (MaRDA extractors working group)

w/Peter Kraus (TUBerlin) & David Elbert (JHU)

The screenshot shows the datatractor website interface. At the top, there's a navigation bar with links for 'About', 'JSON API', 'GitHub', 'File types', and 'Extractors'. Below this, there are two main sections:

- agilent-ch (Agilent ChemStation export file)**: Described as a binary data file created by Agilent ChemStation OpenLab software, containing retention limits and signal data. It includes a 'More details' link.
- agilent-dx (Agilent OpenLab file)**: Described as a raw data export file, actually a zipped archive containing other data formats.

A screenshot of the yadg JSON editor interface. The JSON structure for the agilent-ch file is displayed:

```
JSON Raw Data Headers
Save Copy Collapse All Expand All Filter JSON

data:
  id: "yadg"
  name: "yet another datagram"
  description: "yadg is a set of tools to use raw instrument data."
  subject: []
  citations: []
  license: {}
  supported_filetypes: []
  supported_output_filetypes: null
  source_repository: "https://github.com/dgbowl/yadg"
  documentation: "https://dgbowl.github.io/yadg"

usage:
  0:
    method: "cli"
    setup: null
    command: "yadg extract {{ input_type }} {{ input_path }} {{ output_path }}"
    scope: "meta+data"
    supported_filetypes: null
  1:
  2:

installation:
  0:
    method: "pip"
    requires_python: ">=3.10"
    requirements: null
```

yet another datagram)  
only  
github.com/dgbowl/yadg  
/dgbowl.github.io/yadg  
set of tools to extract and parse raw instrument  
ns:  
e package into a Python 3.9+ environment with `pip  
dg`. After activating the environment, the `extract`  
will be available at the command-line.  
file types:  
gic-mpr  
Note: Several fields in file headers are not  
translated into meaningful parameters.  
gic-mpt  
ent-ch  
ent-dx  
Note: Only .ch files contained in the .dx archive  
are parsed.  
spe  
tical-xrdm1  
:  
le: P. Kraus, N. Vetsch, C. Battaglia, *yadg: yet  
er datagram*

# Datatractor (MaRDA extractors working group)

w/Peter Kraus (TUBerlin) & David Elbert (JHU)

```
from beam import extract

# extract(<input_type>, <input_path>
data = extract("./example.mpr", "biologic-mpr")
```

Follows installation and usage instructions then installs in container, piping resulting data back

```
data>>> data
<xarray.Dataset> Size: 196kB
Dimensions:                                (uts: 721)
Coordinates:
  * uts                                     (uts) float64 6kB 1.557e+09 ... 1.557e+09
Data variables: (12/31)
  Ns                                         (uts) int64 6kB 0 0 0 0 0 0 ... 0 0 0 0 0 0
  time                                       (uts) float64 6kB 1.089e+05 ... 1.521e+05
  time_std_err                               (uts) float64 6kB 5e-05 5e-05 ... 5e-05 5e-05
  control_V                                  (uts) float64 6kB 0.3 0.3 0.3 ... 0.3 0.3 0.3
  control_V_std_err                         (uts) float64 6kB 2e-05 2e-05 ... 2e-05 2e-05
  Ewe                                         (uts) float64 6kB 0.1464 0.2987 ... 0.2986
  ...
  mode                                        (uts) int64 6kB 2 2 2 2 2 2 ... 2 2 2 2 2 2
  ox or red                                  (uts) int64 6kB 1 1 1 1 1 1 ... 1 1 1 1 1 1
  error                                       (uts) int64 6kB 0 0 0 0 0 0 ... 0 0 0 0 0 0
  control changes                           (uts) int64 6kB 1 1 1 1 1 1 ... 1 1 1 1 1 1
  Ns changes                                 (uts) int64 6kB 0 0 0 0 0 0 ... 0 0 0 0 0 0
  counter inc.                             (uts) int64 6kB 0 0 0 0 0 0 ... 0 0 0 0 0 0
Attributes:
  provenance:      yadg extract
  date:            2024-02-21 16:36:20
  datagram_version: 5.0.2
  yadg_extract_filename: /tmp/tmpo276u26f
  yadg_extract_filetype: filetype='marda:biologic-mpr' timezone='Europe/Lo...
```

# Datatractor (MaRDA extractors working group)

w/Peter Kraus (TUBerlin) & David Elbert (JHU)

## Remaining challenges

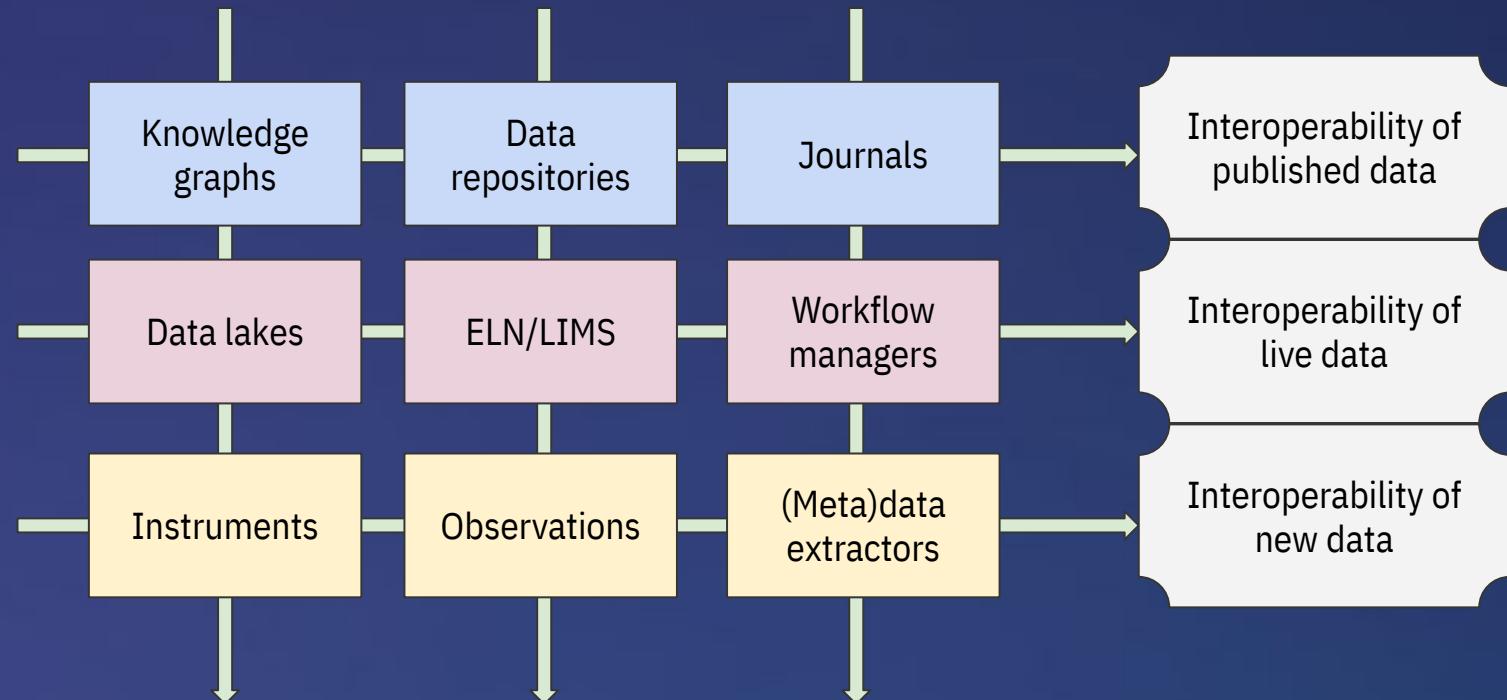
- More robust file type detection
- Automatically running registered extractors against example files
- Providing a place in the schema for announcing output schemas
- Dockerized execution environment (with security features, e.g., no access to network or other files)
- Streamlining contributions to registry and deciding policies for versioning (without implementing our own package manager!)

## Potential use cases

- Create dynamic *datalab* plugins from datatractor to render data in an ELN
  - And publish *datalab* plugins as datatractor extractors!
- Extending domain-specific archive APIs
  - Currently allows an `optimade.yaml` config file to create an OPTIMADE API on top of archived data, but could use `datatractor.yaml` in the future
  - Enhances sustainability of platforms like *datalab*

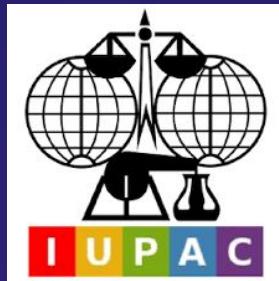
Electronic Lab Notebook Comparison Matrix 2021 (<https://doi.org/10.5281/zenodo.4723753>)

# Interoperability hierarchy



# Machine-Actionable Data Interoperability for Chemical Sciences (MADICES)

- CECAM workshop series: 2022 and 2024 -- hopefully 2025
- Practicable standards for interoperability across:



NATIONAL CENTRE OF COMPETENCE IN RESEARCH



# Concluding thoughts

- Federation enables qualitatively distinct applications
- Treat data management as an investment, rather than a cost
- LLMs != data management
- Can your data be discovered accidentally?
- Can you easily repeat and share systematic screening when new data arises?

# Call to arms

## OPTIMADE:

- Interesting material dataset?
  - Materials data providers
  - Potential users
  - Developers
  - Interested in adopting a generic federated API format?
- 
- + Monthly meetings to discuss extensions, yearly CECAM workshops
  - + Check out Materials-Consortia on GitHub

## *datalab, datatractor & MADICES*

- Dealing with tricky interconnected data?
- Interested groups/companies
- Individual researchers doing this work already, plugin or schema development?
- Looking for a postdoc? Talk to Josh!
  - + Monthly *datalab* meetings
  - + Communal slack
  - + Demo server [demo.datalab-org.io](https://demo.datalab-org.io)
  - + [datalab-org](https://github.com/datalab-org) and [datatractor](https://github.com/datatractor) on GitHub
  - + Talk to Josh and I this week!
  - + MADICES mailing list and pre-meetings

# Acknowledgements



Prof Josh Bocarsly  
(Houston)



Prof Clare Grey  
(Cambridge)



Prof Gian-Marco Rignanese  
(UCLouvain)



Dr David Waroquiers  
(Matgenix)

- + OPTIMADE consortium
- + Materials Research Data Alliance (MaRDA)
- + Physical Sciences Data Infrastructure (PSDI)
- + *re2fractive*: Victor Trinquet (UCLouvain)
- + datatractor: Dr Peter Kraus (TUBerlin)
- + Ben Charmes & Dr Yue Wu
- + The Grey Group (Cambridge)

## Useful links

### OPTIMADE

<https://optimade.org>

*Digital Discovery*, 2024, **3**, 1509-1533  
[10.1039/D4DD00039K](https://doi.org/10.1039/D4DD00039K)

### datalab ecosystem

<https://datalab-org.io>

<https://datalab.industries>

<https://demo.datalab-org.io>

### Datatractor / MaRDA Extractors WG

<https://yard.datatractor.org>

### MADICES

<https://madices.github.io>

