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

Dr Matthew Evans

https://ml-evs.science

AIChemy seminar, Imperial College London
24/06/2025

Broadening impact via decentralized platforms

OPTIMADE

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

datalab

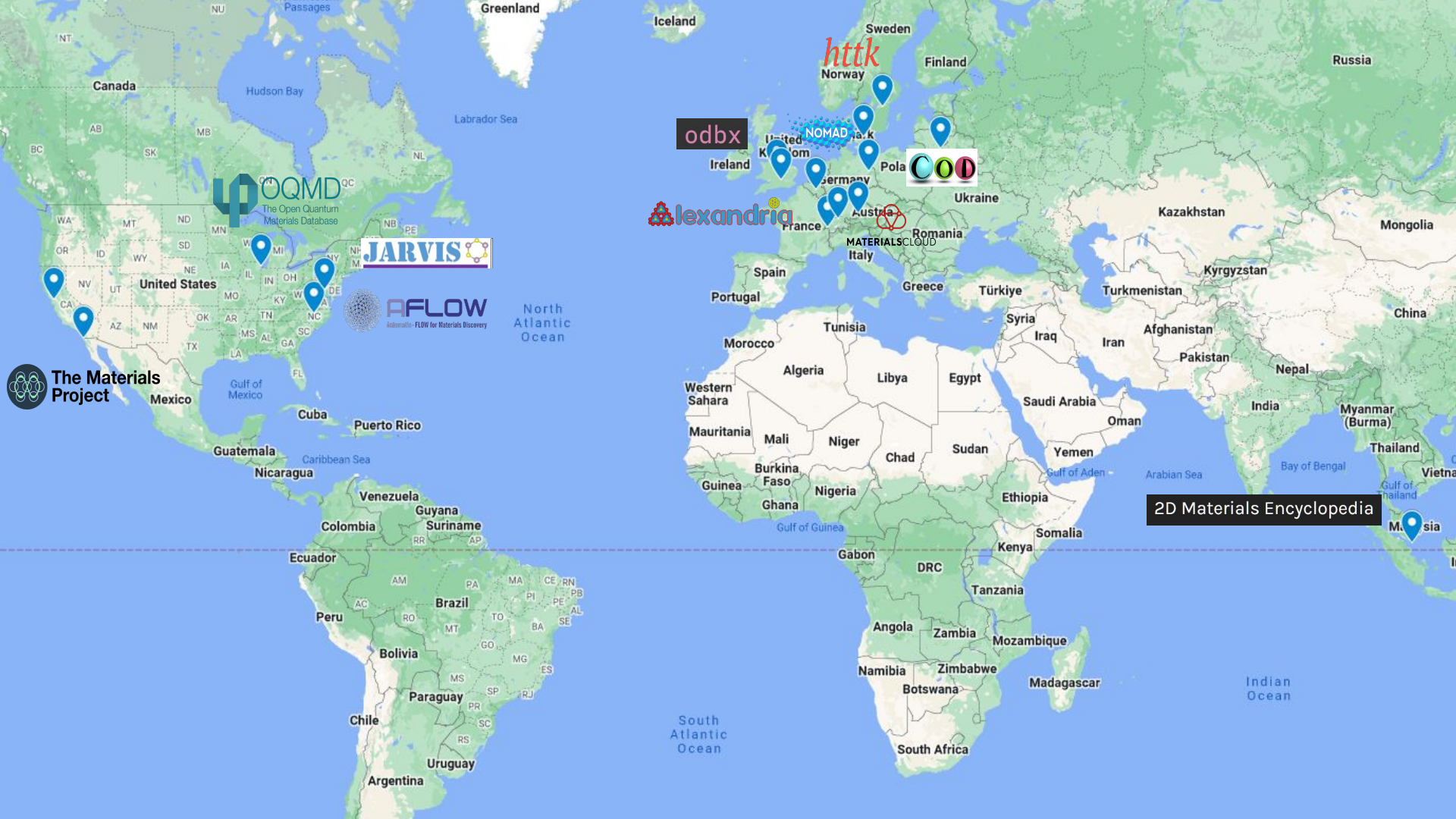
An open source data management platform for experimental materials chemistry and beyond

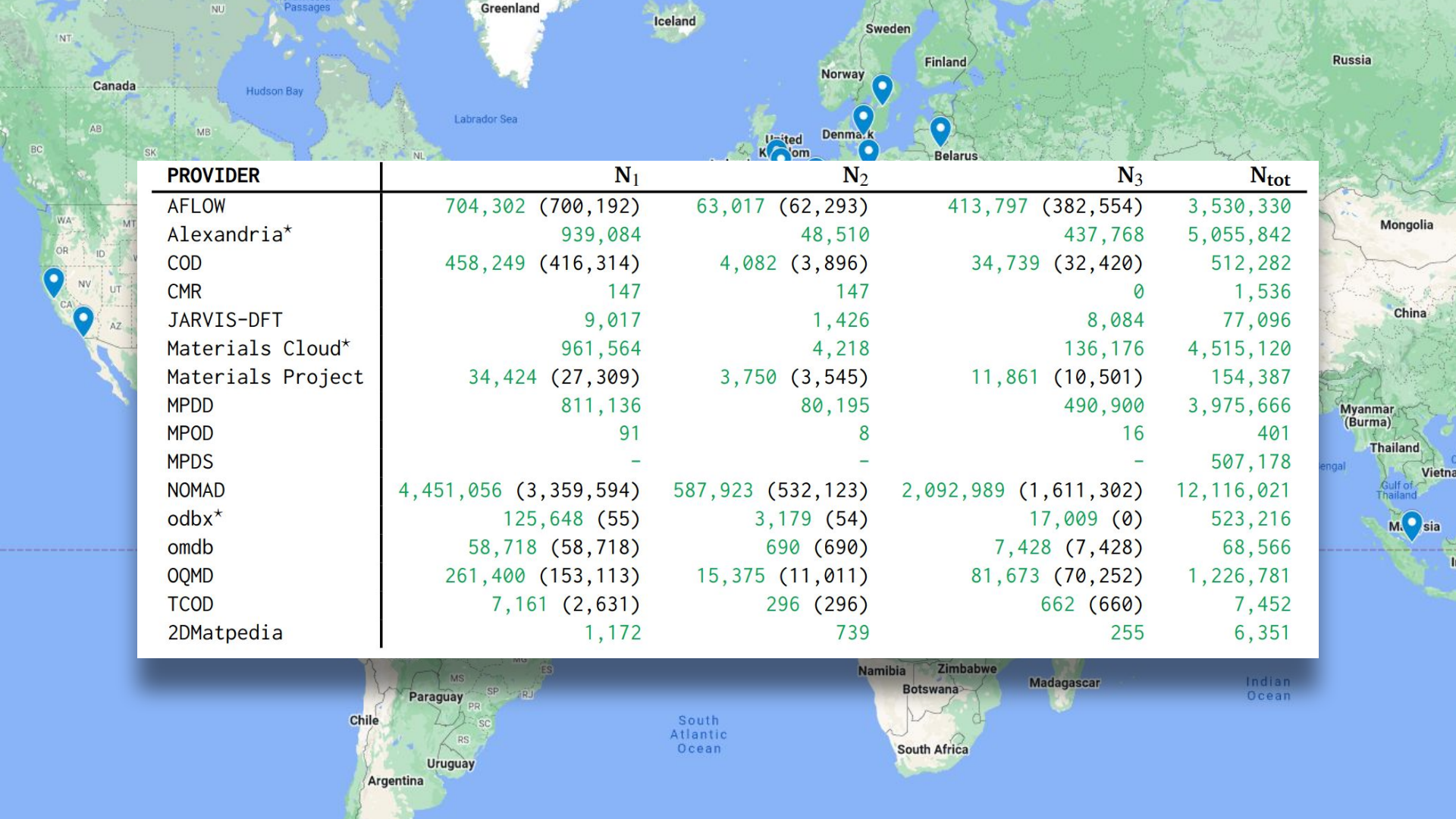
What is OPTIMADE?

“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


httk
odbx
NOMAD
COD
lexandria
MATERIALSCLOUD
OQMD
The Open Quantum Materials Database
JARVIS
AFLOW
The Materials Project
2D Materials Encyclopedia
Greenland
Iceland
Sweden
Finland
Norway
Russia
Canada
Hudson Bay
Labrador Sea
Ireland
Ukraine
Kazakhstan
Mongolia
France
Romania
Italy
Spain
Portugal
Greece
Türkiye
Turkmenistan
Kyrgyzstan
China
Syria
Iraq
Iran
Afghanistan
Pakistan
Nepal
India
Myanmar (Burma)
Thailand
Vietnam
Morocco
Tunisia
Algeria
Libya
Egypt
Western Sahara
Saudi Arabia
Oman
Yemen
Mauritania
Mali
Niger
Chad
Sudan
Burkina Faso
Guinea
Nigeria
Ethiopia
Ghana
Gulf of Guinea
Somalia
Kenya
Gabon
DRC
Tanzania
Angola
Zambia
Mozambique
Namibia
Zimbabwe
Botswana
Madagascar
South Africa
United States
Mexico
Gulf of Mexico
Cuba
Puerto Rico
Guatemala
Nicaragua
Caribbean Sea
Venezuela
Guyana
Suriname
Colombia
Ecuador
Peru
Brazil
Bolivia
Paraguay
Chile
Uruguay
Argentina
North Atlantic Ocean
South Atlantic Ocean
Indian Ocean
Arabian Sea
Gulf of Aden
Bay of Bengal
Gulf of Thailand



PROVIDER	N_1	N_2	N_3	N_{tot}
AFLOW	704,302 (700,192)	63,017 (62,293)	413,797 (382,554)	3,530,330
Alexandria*	939,084	48,510	437,768	5,055,842
COD	458,249 (416,314)	4,082 (3,896)	34,739 (32,420)	512,282
CMR	147	147	0	1,536
JARVIS-DFT	9,017	1,426	8,084	77,096
Materials Cloud*	961,564	4,218	136,176	4,515,120
Materials Project	34,424 (27,309)	3,750 (3,545)	11,861 (10,501)	154,387
MPDD	811,136	80,195	490,900	3,975,666
MPOD	91	8	16	401
MPDS	-	-	-	507,178
NOMAD	4,451,056 (3,359,594)	587,923 (532,123)	2,092,989 (1,611,302)	12,116,021
odbx*	125,648 (55)	3,179 (54)	17,009 (0)	523,216
omdb	58,718 (58,718)	690 (690)	7,428 (7,428)	68,566
OQMD	261,400 (153,113)	15,375 (11,011)	81,673 (70,252)	1,226,781
TCOD	7,161 (2,631)	296 (296)	662 (660)	7,452
2DMatpedia	1,172	739	255	6,351

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

RESOURCE REPRESENTATION

DATABASE AND DOMAIN-AGNOSTIC FILTERING GRAMMAR

PROVIDER FEDERATION

MACHINE-ACTIONABLE PROPERTY DEFINITIONS

RESOURCE LAYOUT (URL formats)

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"],
```

RESOURCE REPRESENTATION

FILTERING GRAMMAR

PROPERTY DEFINITIONS

RESOURCE LAYOUT

PROVIDER FEDERATION

```
/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


```
{
    "$id": "https://schemas.optimade.org/defs/v1.2/properties/optimade
    ↪ /structures/nsites",
    "$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**:\n\n- MUST be equal to
    ↪ the lengths of the list properties elements and elements_ratios,
    ↪ if they are provided.\n\n**Query examples**:\n\n- Match only
    ↪ structures with exactly 4 sites: `nsites=4`\n- 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

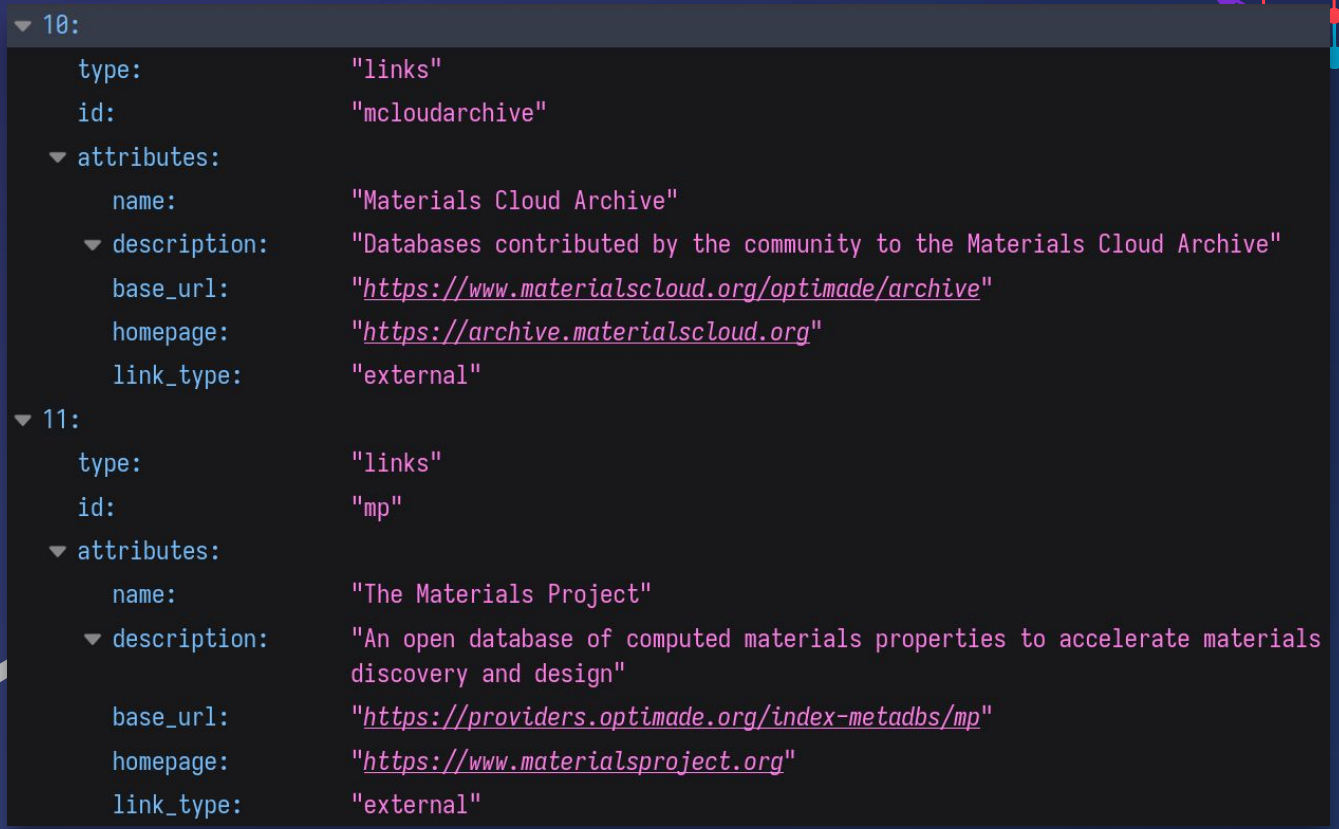
RESOURCE REPRESENTATION

FILTERING GRAMMAR

PROPERTY DEFINITIONS

RESOURCE LAYOUT

PROVIDER FEDERATION



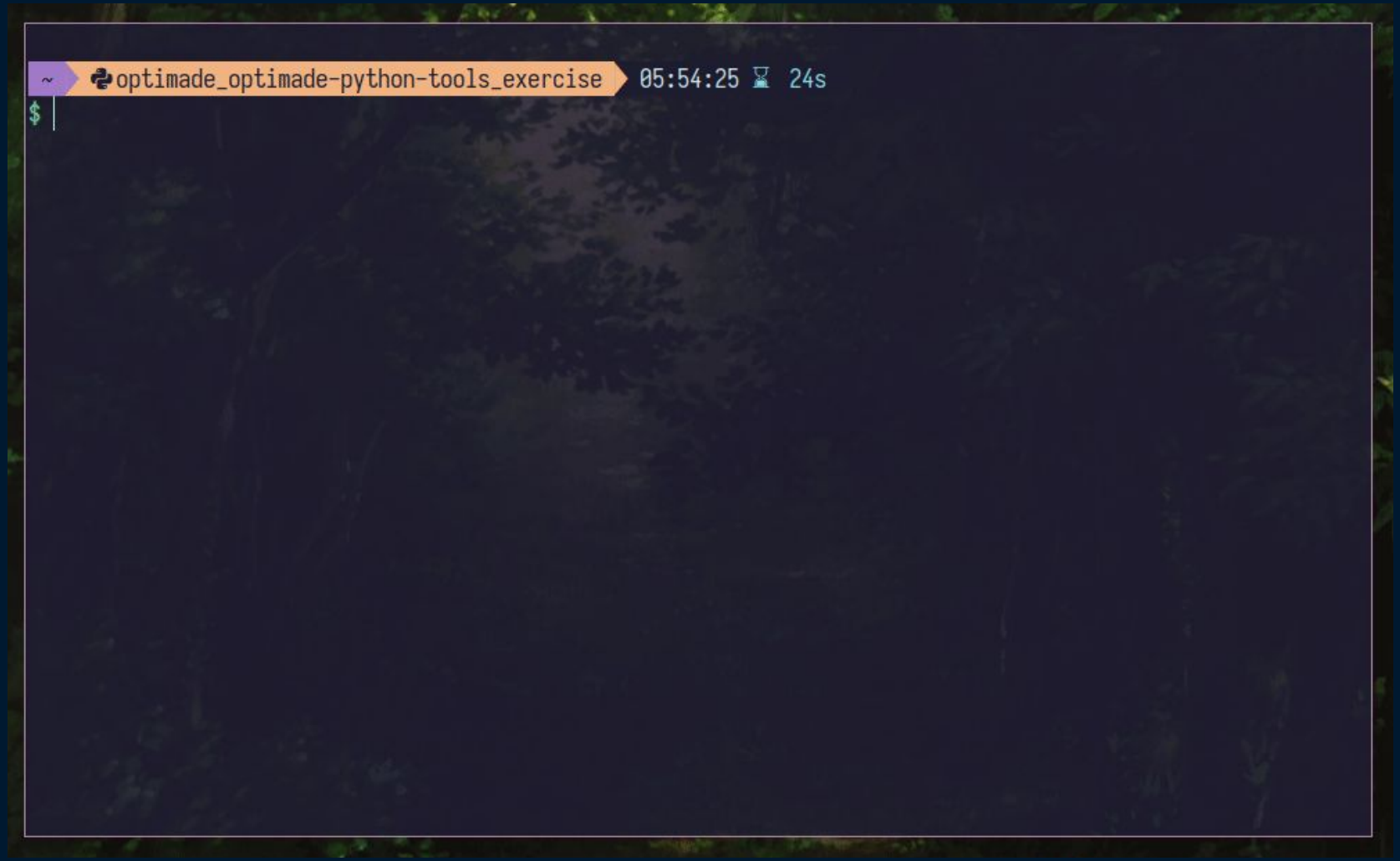
Materials-Consortia/optimade-python-tools

optimade-python-tools

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

- 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

Evans, M. L., Andersen, C. W. A. *et al.*,
Journal of Open Source Software, 6(65), 3458, 2021
10.21105/joss.03458


~
optimade_optimade-python-tools_exercise
05:54:25
24s
$

How can we screen the burgeoning number of hypothetical (inorganic) compounds 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

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*

MODNet

matminer

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

ATOMATE2

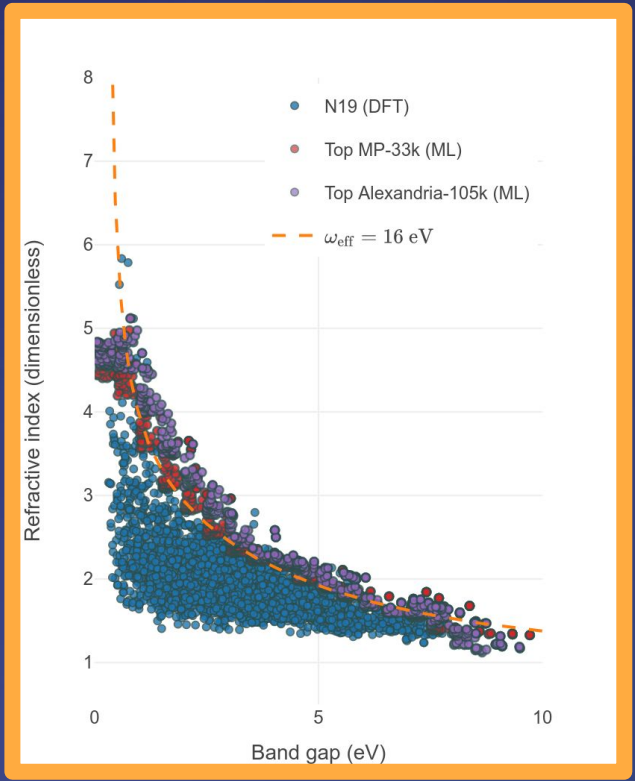
Robust automated DFT+beyond workflows

jobflow-remote

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)

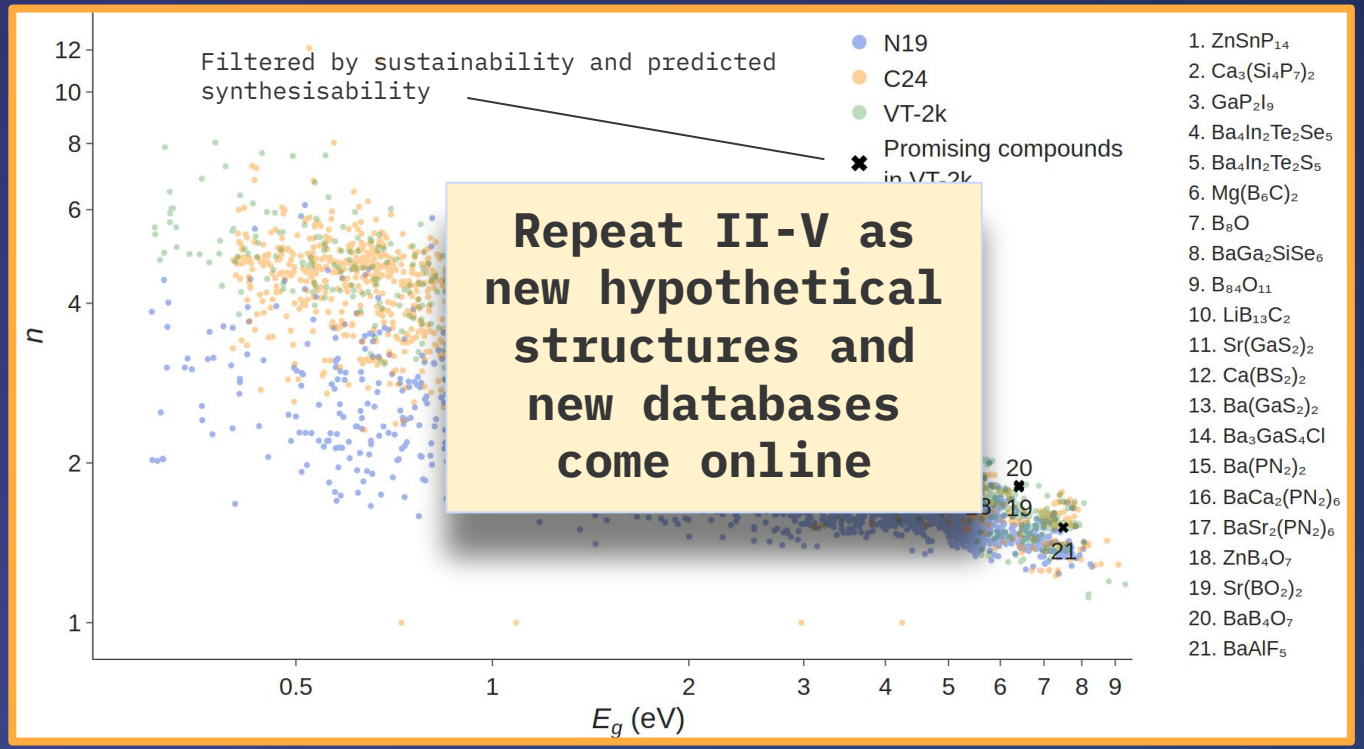
III. Filter OPTIMADE entries for stable materials with non-zero band gap (~150k structures), featurize and predict

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

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



"Optical materials discovery and design via federated databases and machine learning", Trinquet, **Evans** *et al*, Faraday Discussions (2024) 10.1039/D4FD00092G

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

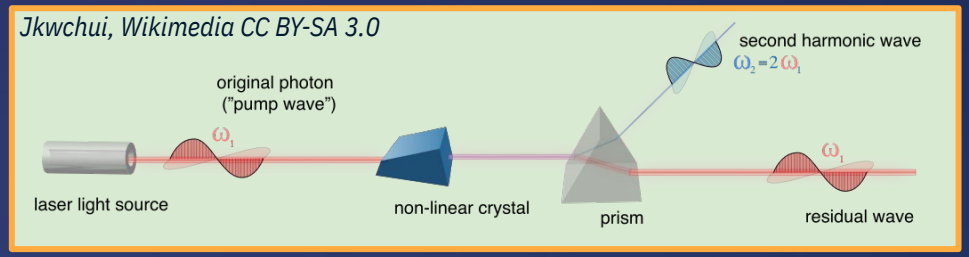
- AL especially necessary when properties are rare + expensive to compute
- Therefore, we initialised a new campaign for non-linear optical response

"Optical materials discovery and design via federated databases and machine learning", Trinquet, **Evans** *et al*, Faraday Discussions (2024) 10.1039/D4FD00092G

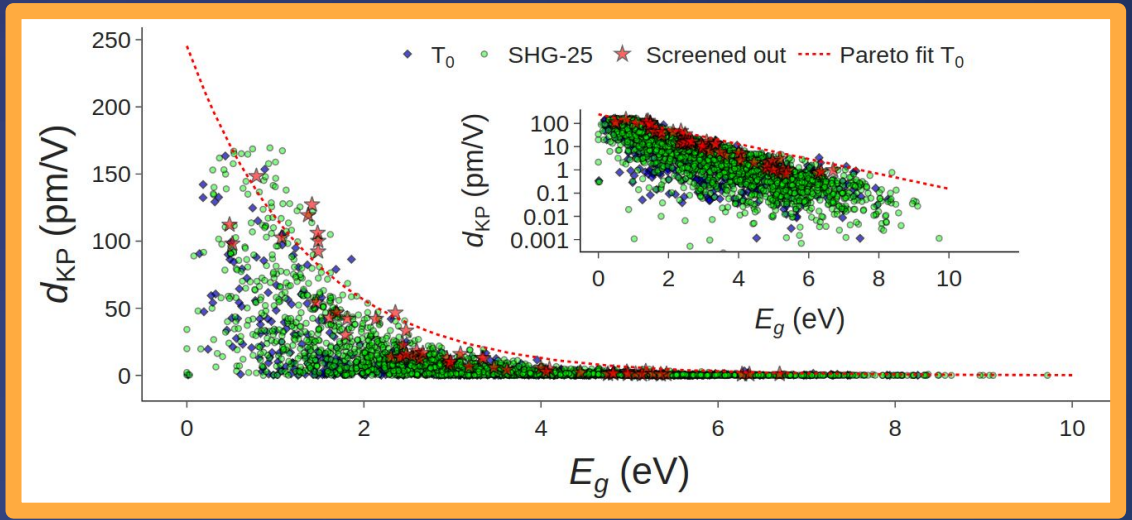
Accelerating the discovery of high-performance nonlinear optical materials using active learning and high-throughput screening

Victor Trinquet, Matthew Evans & Gian-Marco Rignanese

10.48550/arXiv.2504.01526



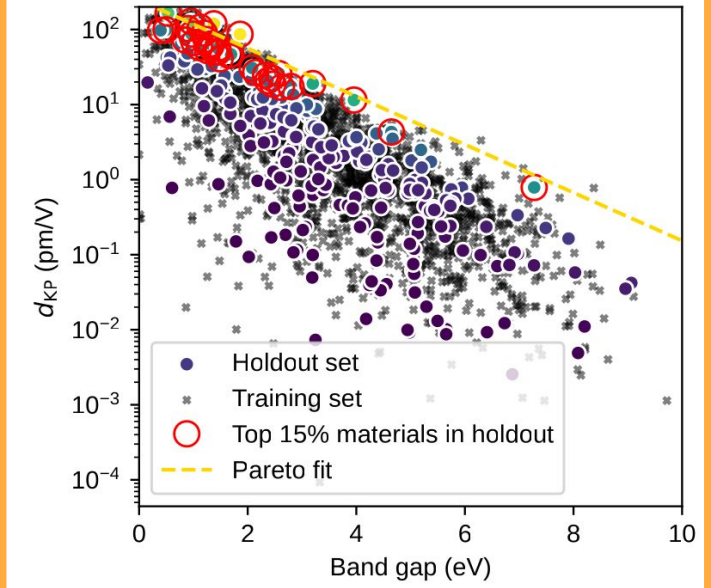
- Starting dataset of ~600 compounds with GGA SHG tensors using Abinit workflows from Trinquet et al, Sci Data
- Use MODNet + AL to explore the most promising , most uncertain and most compatible materials accessible via OPTIMADE
- Expand dataset to around ~2200 SHG tensors, compute HSE tensors for a subset and then fit/learn the scissor correction on the SHG coefficient

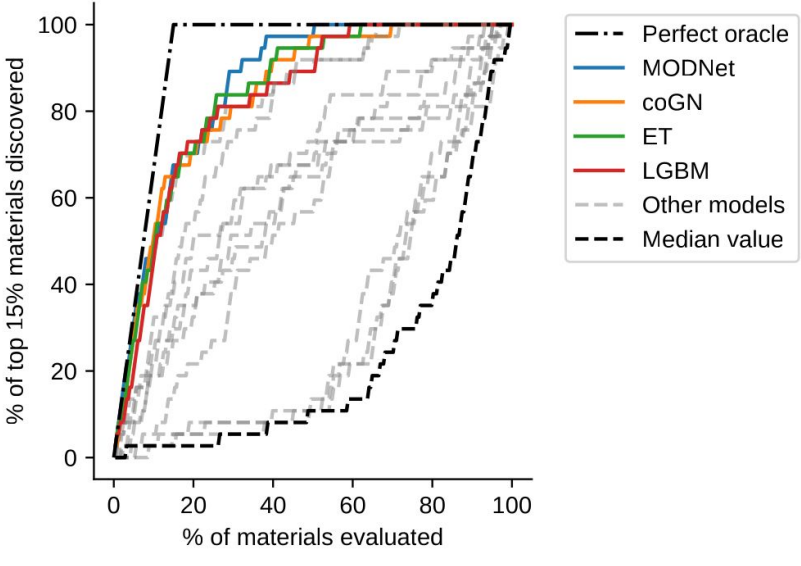


Accelerating the discovery of high-performance nonlinear optical materials using active learning and high-throughput screening

Victor Trinquet, Matthew Evans & Gian-Marco Rignanese

10.48550/arXiv.2504.01526

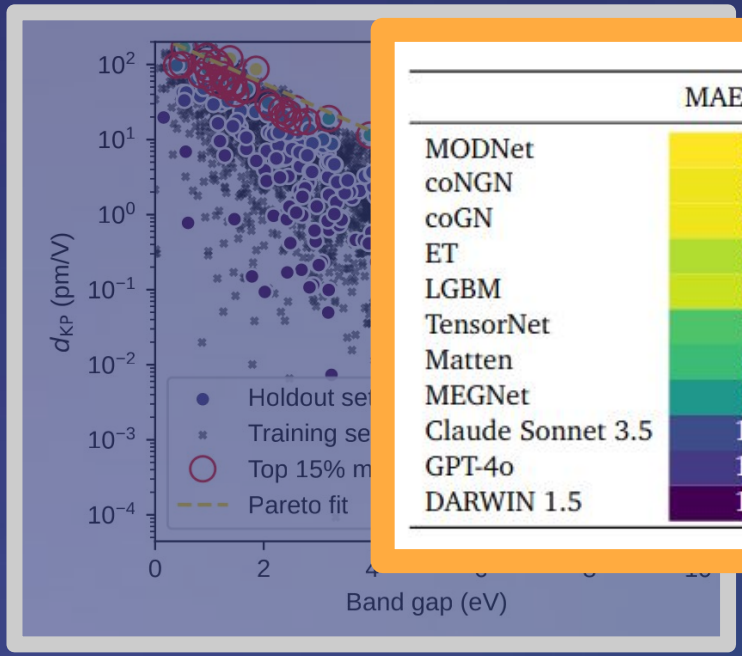




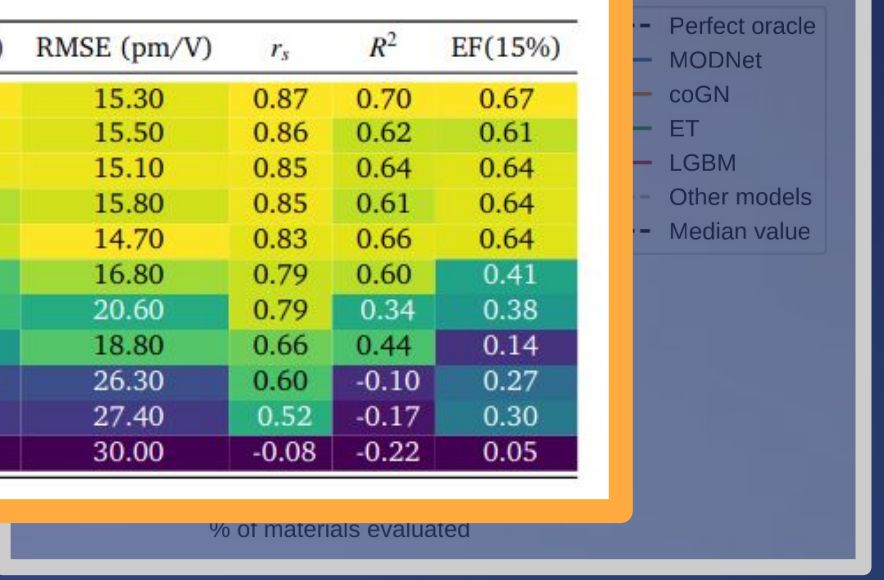
Accelerating the discovery of high-performance nonlinear optical materials using active learning and high-throughput screening

Victor Trinquet, Matthew Evans & Gian-Marco Rignanese

10.48550/arXiv.2504.01526



	MAE (pm/V)	RMSE (pm/V)	r_s	R^2	EF(15%)
MODNet	5.80	15.30	0.87	0.70	0.67
coNGN	6.00	15.50	0.86	0.62	0.61
coGN	6.00	15.10	0.85	0.64	0.64
ET	6.70	15.80	0.85	0.61	0.64
LGBM	6.40	14.70	0.83	0.66	0.64
TensorNet	7.90	16.80	0.79	0.60	0.41
Matten	8.20	20.60	0.79	0.34	0.38
MEGNet	9.30	18.80	0.66	0.44	0.14
Claude Sonnet 3.5	11.60	26.30	0.60	-0.10	0.27
GPT-4o	12.00	27.40	0.52	-0.17	0.30
DARWIN 1.5	13.30	30.00	-0.08	-0.22	0.05



Accelerating the discovery of high-performance nonlinear optical materials using active learning and high-throughput screening

Victor Trinquet, Matthew Evans & Gian-Marco Rignanese

10.48550/arXiv.2504.01526

Identifier	Formula	E_g^{HSE} (eV)	d_{HSE} (pm/V)	Δn_{HSE}	Spacegroup	Wang et al.[14]	Chu et al.[13]
mp-966800	InP	1.261	27.499	0.108	$P6_3mc$	-	-
mp-1215429	$ZnSnP_2$	1.325	49.348	0.092	$P\bar{4}m2$	-	-
agm003450028	$Mg(InTe_2)_2$	1.461	63.528	0.038	$P\bar{4}2m$	-	-
mp-35777	$Mg(InTe_2)_2$	1.464	49.854	0.055	Cm	-	-
mp-571195	ZnTe	1.938	46.356	0.035	$P3_1$	-	-
mp-1222182	$Mg(InTe_2)_2$	1.960	51.373	0.041	$I\bar{4}$	-	-
agm002160623	$Mg(GaTe_2)_2$	2.140	61.522	0.076	$I\bar{4}$	-	-
agm002088965	InTeI	2.158	55.801	0.385	$P2_1$	-	-
agm002156796	$Li_2SiSnSe_4$	2.306	30.451	0.257	Cm	-	-
agm005605697	GaTeI	2.314	44.027	0.411	$Pmn2_1$	-	-
agm002790067	$AlInP_2$	2.399	27.310	0.039	$I\bar{4}2d$	-	-
agm005056337	$MgInGaSe_4$	2.497	21.144	0.052	$I\bar{4}$	-	-
agm002283412	Li_2SnGeS_4	2.647	16.802	0.239	Cm	-	-
agm002160619	$Mg(GaSe_2)_2$	2.748	21.159	0.064	$I\bar{4}$	-	-
agm002160138	$Mg(AlTe_2)_2$	2.872	23.966	0.041	$I\bar{4}$	-	-
agm2000111340	GaTeCl	3.128	8.074	0.536	$Pmn2_1$	-	-
mp-4586	$LiAlTe_2$	3.156	15.120	0.058	$I\bar{4}2d$	-	✓
mp-27529	PI_3	3.199	8.224	0.265	$P6_3$	-	-
agm002793928	$NaBSe_2$	3.361	9.264	0.058	$I\bar{4}2d$	-	-
mp-690	P_4S_5	3.413	6.937	0.164	$P2_1$	✓	-
agm2000135800	GaTeCl	3.499	7.012	0.507	$Pca2_1$	-	-
agm006047631	Ga_4SnS_7	3.527	8.098	0.099	$Pmn2_1$	-	-
mp-20790	$InPS_4$	3.538	18.361	0.057	$I\bar{4}$	-	✓
mp-30294	Sr_2SnS_4	3.630	7.997	0.076	$Ama2$	✓	✓
agm005605595	GaSeCl	3.653	9.915	0.384	$Pmn2_1$	-	-
mp-1227993	$BaGa_2SiS_6$	3.909	10.020	0.063	$P1$	-	✓
agm002157245	$LiAlSe_2$	3.962	5.415	0.041	$I\bar{4}2d$	-	-
mp-2646995	Li_3PS_4	3.991	4.472	0.039	$I\bar{4}2m$	-	-
agm002158826	$LiGaS_2$	4.078	7.071	0.060	$I\bar{4}2d$	-	-
agm005605654	GaSCl	4.264	5.416	0.260	$Pmn2_1$	-	-
mp-559065	NaI_3O_8	4.474	4.325	0.149	$P\bar{4}$	-	-
agm002157243	$LiAlS_2$	4.620	2.641	0.031	$I\bar{4}2d$	-	-
mp-561104	$Ga(IO_3)_3$	4.699	8.119	0.098	$P6_3$	-	-
mp-555903	$Al(IO_3)_3$	4.747	7.942	0.082	$P6_3$	-	-
mp-559545	SeO_2	4.835	3.674	0.299	$Pmc2_1$	-	-
mp-27367	$SeOF_2$	5.537	2.830	0.051	$Pca2_1$	✓	-
agm002163269	$CaSiN_2$	5.561	3.424	0.034	$I\bar{4}2d$	-	-
mp-22909	$ZnCl_2$	5.574	1.373	0.031	$I\bar{4}2d$	-	-
agm005604809	AlSCl	5.743	1.164	0.179	$Pca2_1$	-	-
agm002137165	$CaAl_2B_2O_7$	6.405	0.713	0.066	$R32$	-	-
mp-5730	$Ba(BO_2)_2$	6.418	1.365	0.115	$R3c$	-	-
mp-5853	$LiSi_2N_3$	6.518	0.821	0.037	$Cmc2_1$	✓	-
mp-557391	$Na_2Ca_2(CO_3)_3$	6.533	0.631	0.034	$Amm2$	-	-
mp-753671	PNO	6.705	1.844	0.121	$I2_12_12_1$	✓	-
mp-36066	PNO	6.736	1.612	0.151	Cc	✓	-
agm005607967	SiNF	6.737	0.471	0.188	$Pmn2_1$	-	-
mp-6524	$CaMg_3(CO_3)_4$	6.898	0.800	0.154	$R32$	-	-
mp-1195844	$Ba_3B_6O_{11}F_2$	6.940	0.761	0.046	$P2_1$	-	-
mp-1202821	$Sr_3B_6O_{11}F_2$	7.143	0.624	0.041	$P2_1$	-	-
mp-1020019	Li_2PNO_2	7.271	0.449	0.103	$Cmc2_1$	-	-
mp-1200209	$Li_2B_6O_9F_2$	7.993	0.485	0.061	Cc	-	-
mp-3660	LiB_3O_5	8.166	0.561	0.042	$Pna2_1$	-	-
mp-1019509	$B_2S_2O_9$	9.158	0.656	0.034	$C2$	-	-

<10 meV/atom away from hull, HSE gap > 1 eV, elemental availability, birefringence, dKP criteria threshold

optimade-maker

with Kristjan Eimre, Gian-Marco Rignanese & 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

`materialscloud-org/optimade-maker`

```yaml
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
```

optimade-maker

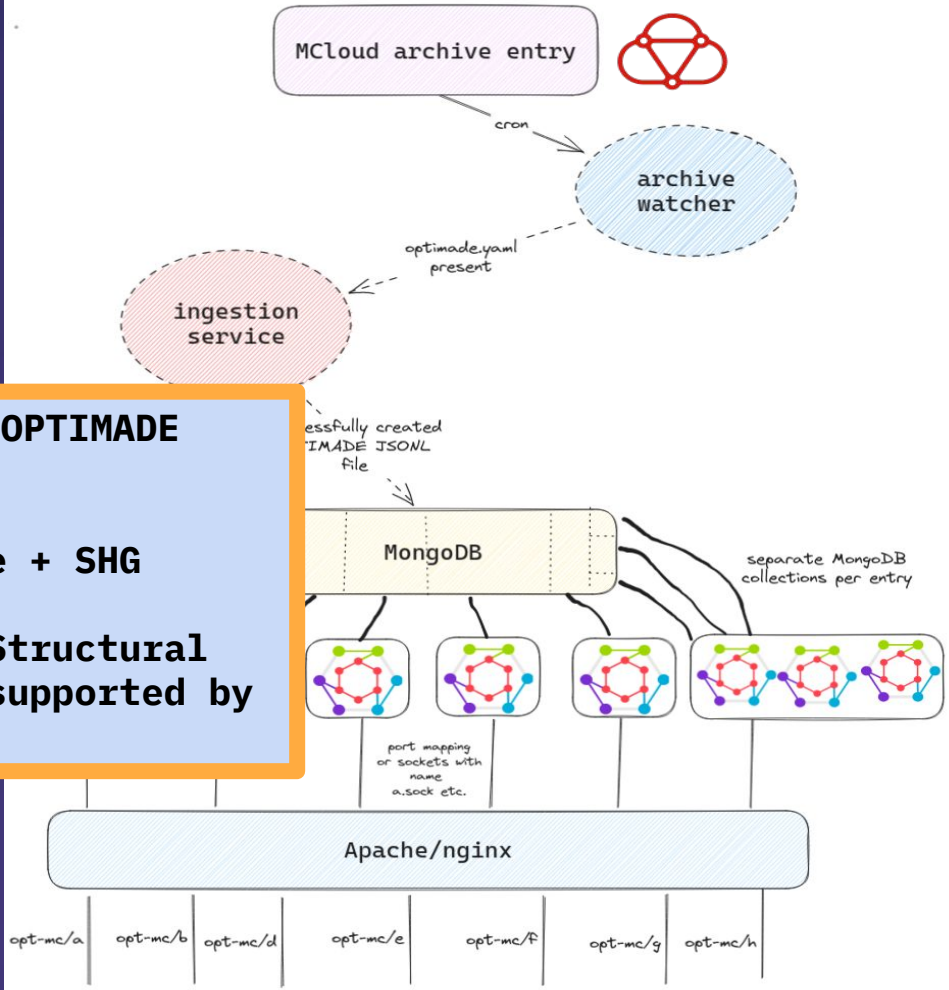
with Kristjan Eimre, Gian-Marco Rignanese & Giovanni Pizzi

From static data to an API

- Enhancing discoverability, explorability and expressiveness of stru… the MCloud
- User annotates their MClou… simple config file
- Additional services watch f… and run a processing pipel… our OPTIMADE JSONL forma…
- Container launches to expose the structural data and properties as public OPTIMADE APIs

Used to create OPTIMADE APIs for:

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



materialscloud-org/optimade-maker

Well-defined abstraction: crystal structure

Flat(ish) data structure

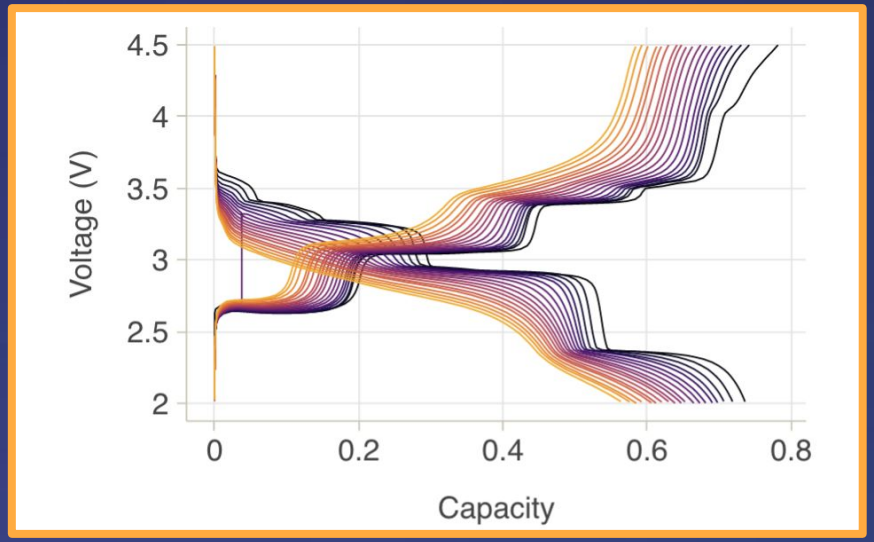
Extensible property definitions

Well-defined search

Tools for federation

What about experimental data?

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_2O_2 from ECO glovebox

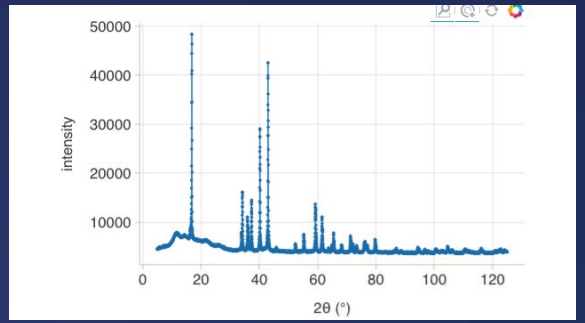
NiO (from collaborator)

Jmas-1-4a

Jmas-1-4b

Jmas-1-4c

Jmas-1-4d

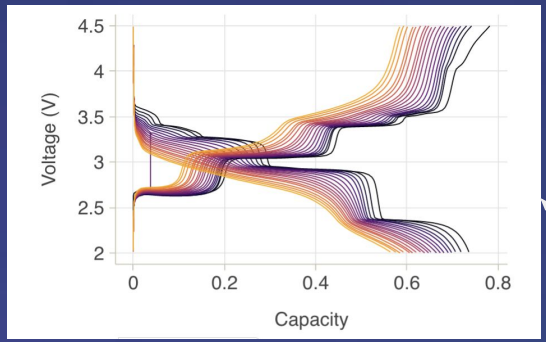


Super P

PVDF

NMP

Jmas-1-4a_e1



NaPF6

PC

Jmas1-4a-e1-c1

Jmas1-4a-e1-c2

Jmas1-4a-e1-c3

electrolyte

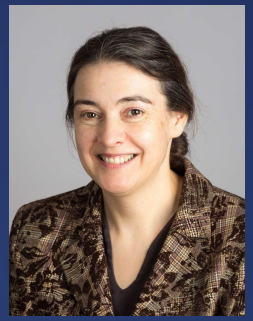
Na metal

Initial motivations: *datalab*

datalab-org/datalab

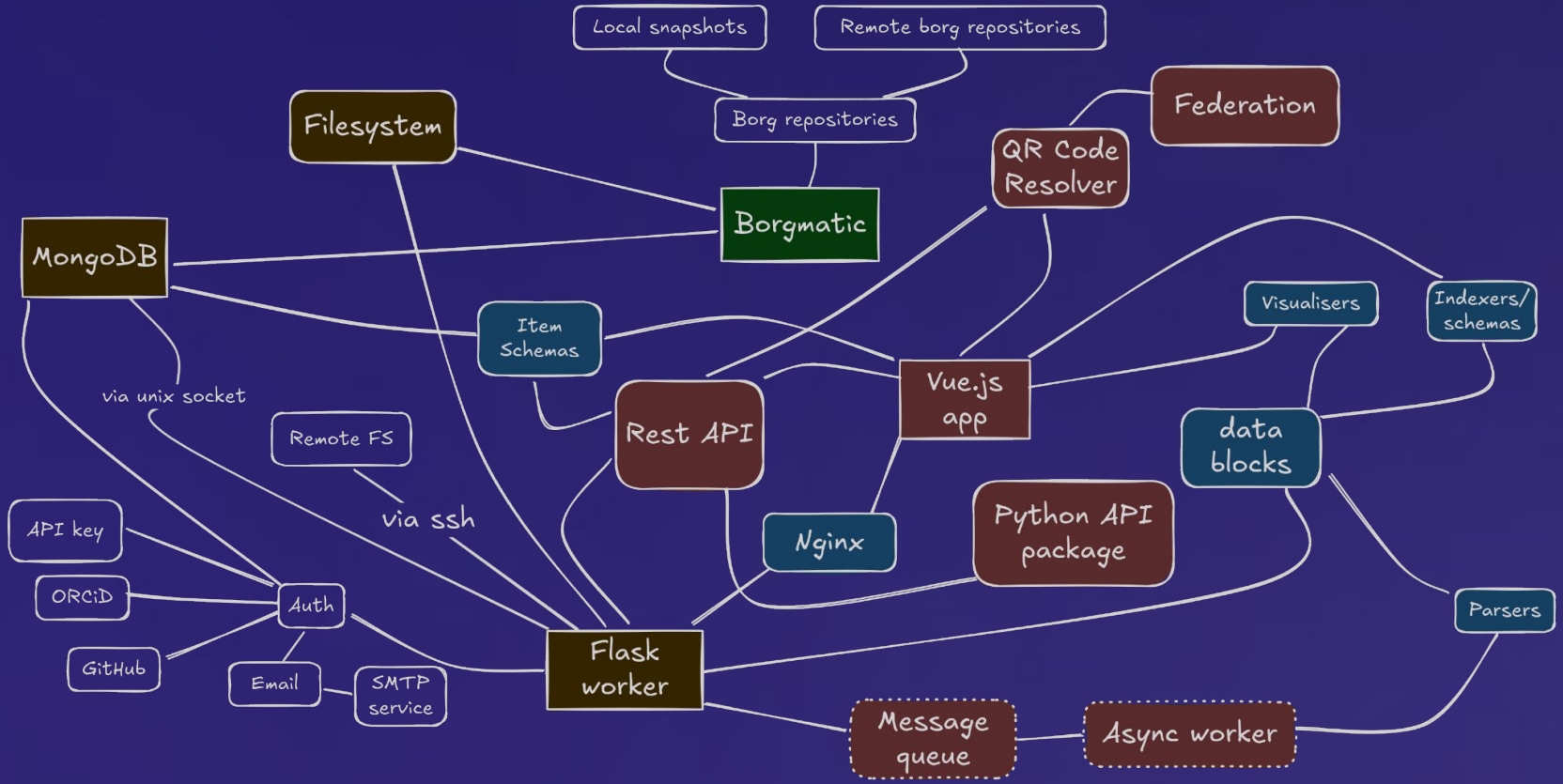
- **Automating tedious data handling processes**
- **Reproducible, robust science**
- **Enhanced dissemination (internal and public)**
- **Putting the FUN in data management**
- **Decentralized data unification**

Prof Joshua Bocarsly (Cambridge ✈️ UHouston)

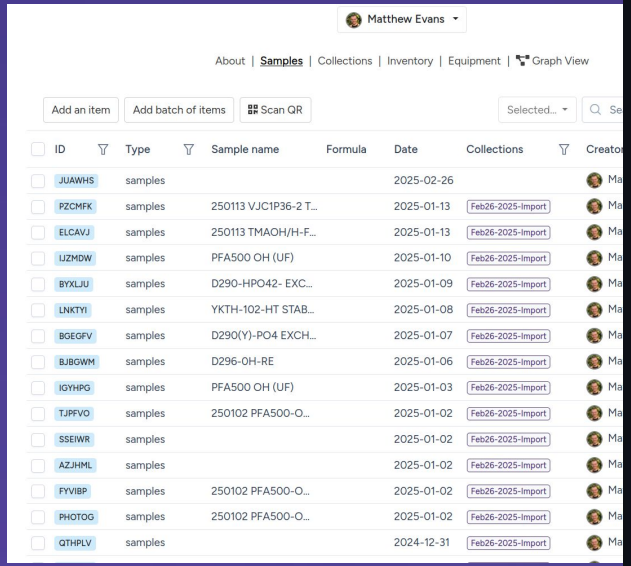


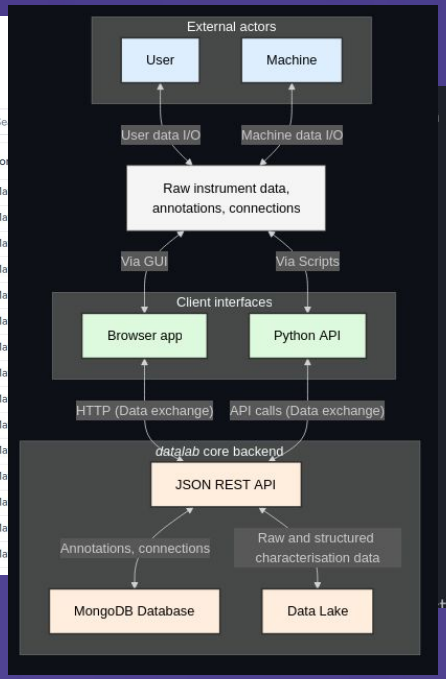
Dame Prof Clare Grey FRS (Cambridge)

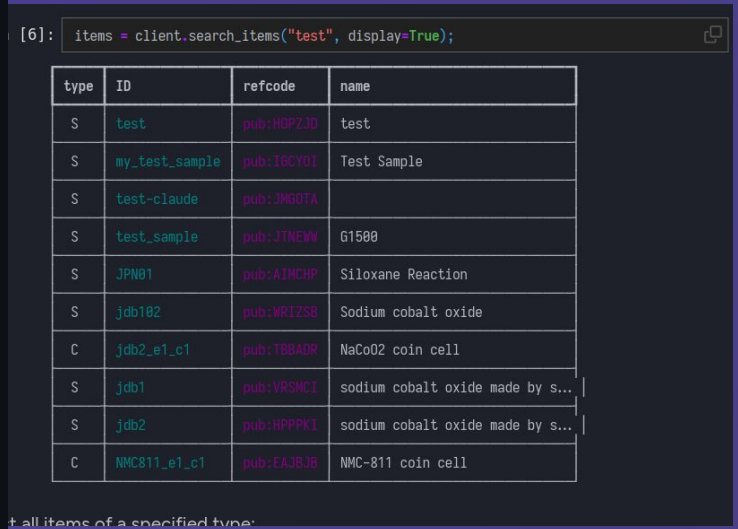
datalab ~~Architecture~~ Spaghetti



What is *datalab*?



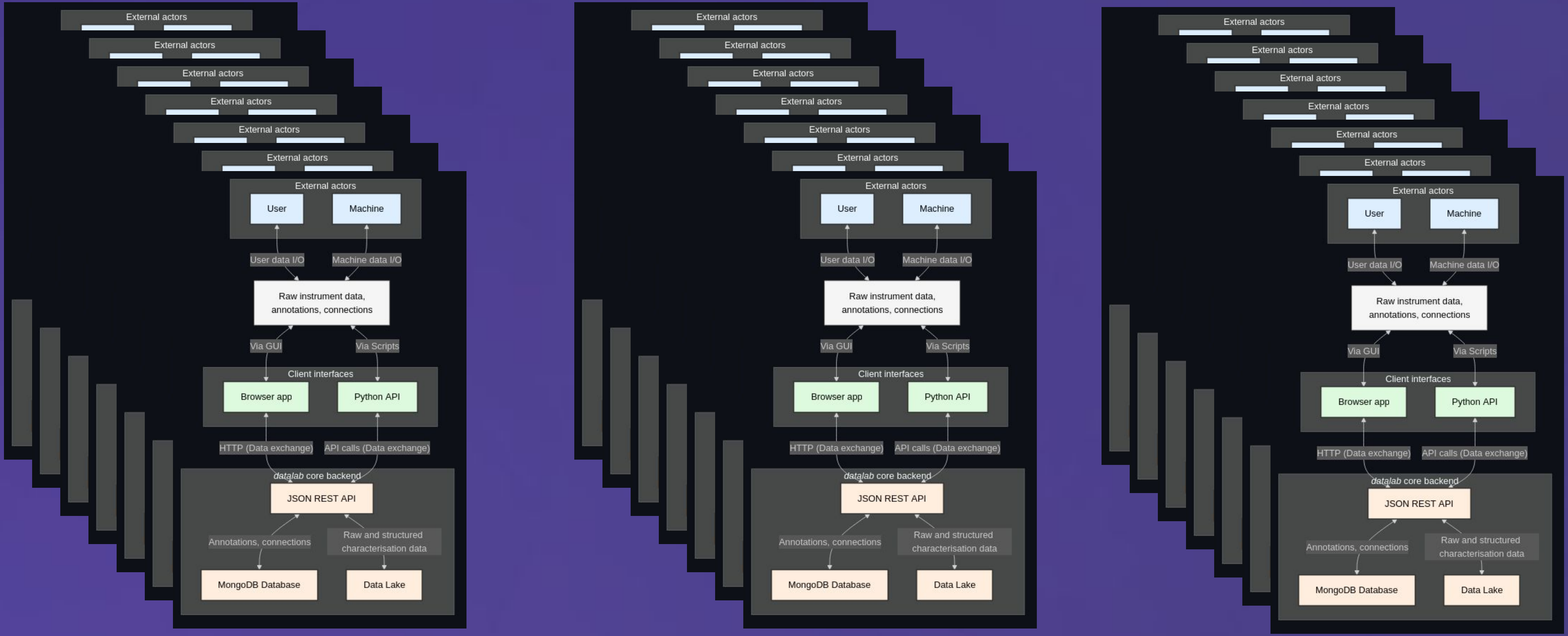




datalab-org/datalab

datalab-org/datalab-api

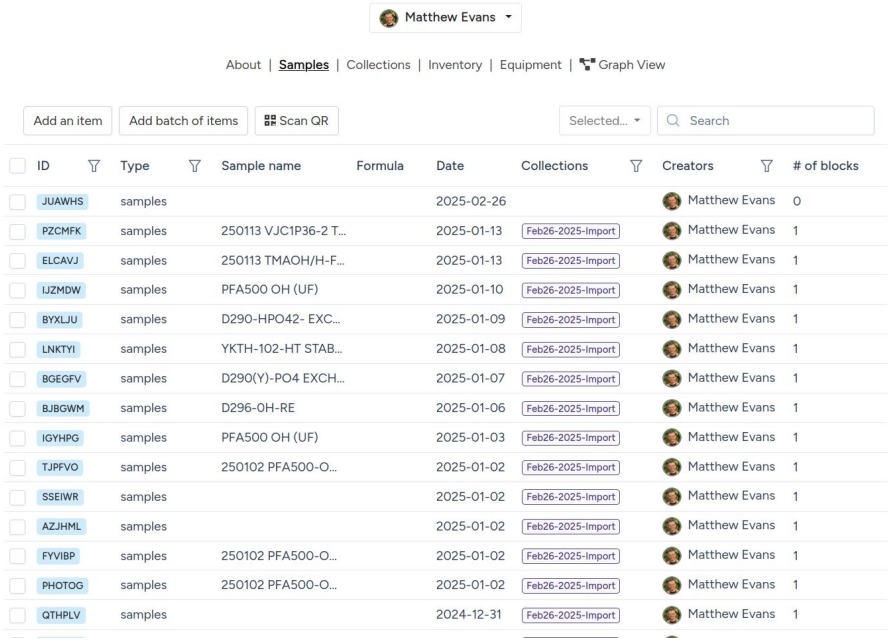
What is *datalab*?



datalab-org/datalab-federation

datalab: data management for batteries & materials characterisation data

- Open source ELN/LIMS, with development funded by academic work and consultancy
- Deployed specifically for a given company, lab or research group
- Runs in the cloud or on-prem on your existing hardware
- Accessible from anywhere (the lab, at home etc)



ID	Type	Sample name	Formula	Date	Collections	Creators	# of blocks
JUAWHS	samples			2025-02-26		Matthew Evans	0
PZCMFK	samples	250113 VJC1P36-2 T...		2025-01-13	Feb26-2025-Import	Matthew Evans	1
ELCAVJ	samples	250113 TMAOH/H-F...		2025-01-13	Feb26-2025-Import	Matthew Evans	1
IJZMDW	samples	PFA500 OH (UF)		2025-01-10	Feb26-2025-Import	Matthew Evans	1
BYXLJU	samples	D290-HPO42- EXC...		2025-01-09	Feb26-2025-Import	Matthew Evans	1
LNKTYI	samples	YKTH-102-HT STAB...		2025-01-08	Feb26-2025-Import	Matthew Evans	1
BGEGFV	samples	D290(Y)-PO4 EXCH...		2025-01-07	Feb26-2025-Import	Matthew Evans	1
BJBGWM	samples	D296-0H-RE		2025-01-06	Feb26-2025-Import	Matthew Evans	1
IGYHPG	samples	PFA500 OH (UF)		2025-01-03	Feb26-2025-Import	Matthew Evans	1
TJPFVO	samples	250102 PFA500-O...		2025-01-02	Feb26-2025-Import	Matthew Evans	1
SSEIWR	samples			2025-01-02	Feb26-2025-Import	Matthew Evans	1
AZJHML	samples			2025-01-02	Feb26-2025-Import	Matthew Evans	1
FYVIBP	samples	250102 PFA500-O...		2025-01-02	Feb26-2025-Import	Matthew Evans	1
PHOTOG	samples	250102 PFA500-O...		2025-01-02	Feb26-2025-Import	Matthew Evans	1
QTHPLV	samples			2024-12-31	Feb26-2025-Import	Matthew Evans	1

Grey Group

Josh Bocarsly Logout

About | Samples | Collections | Inventory | Graph View

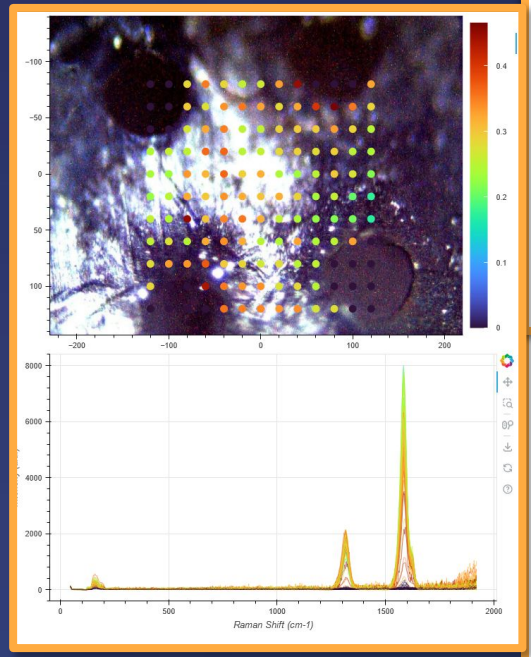
Add an item | Add batch of samples

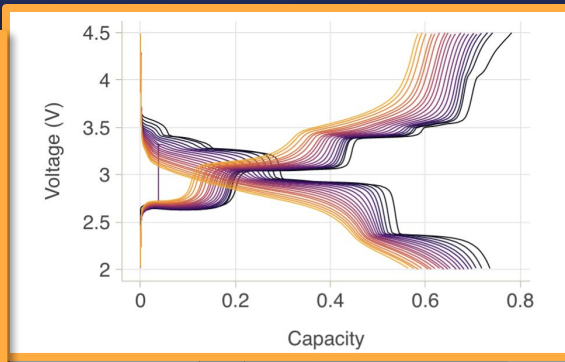
ID	Type	Sample name	Formula	Date	Collections	Creators	# of blocks
cell_13b	sample			2023-09-12			1
test-02	cell	test cell		2023-09-08			2
test-01	sample	test		2023-09-08			1
jdb1003	sample	a sample of sodium solution		2023-09-07			1
Ni3Al_1	sample	solid state sample of Ni-Al		2023-09-05			1
jdb_test1	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

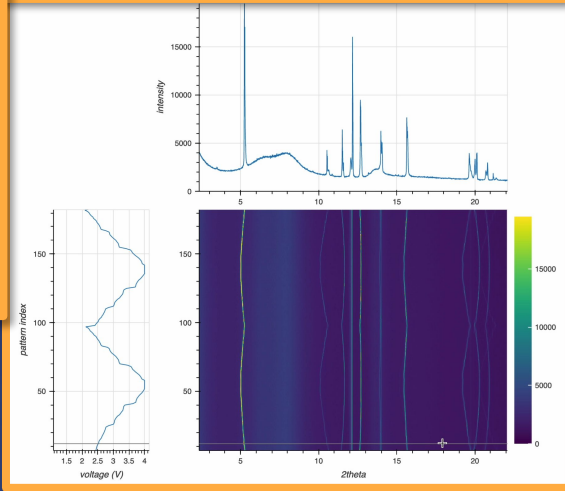
Modular “Data block” approach

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: extensible via plugins





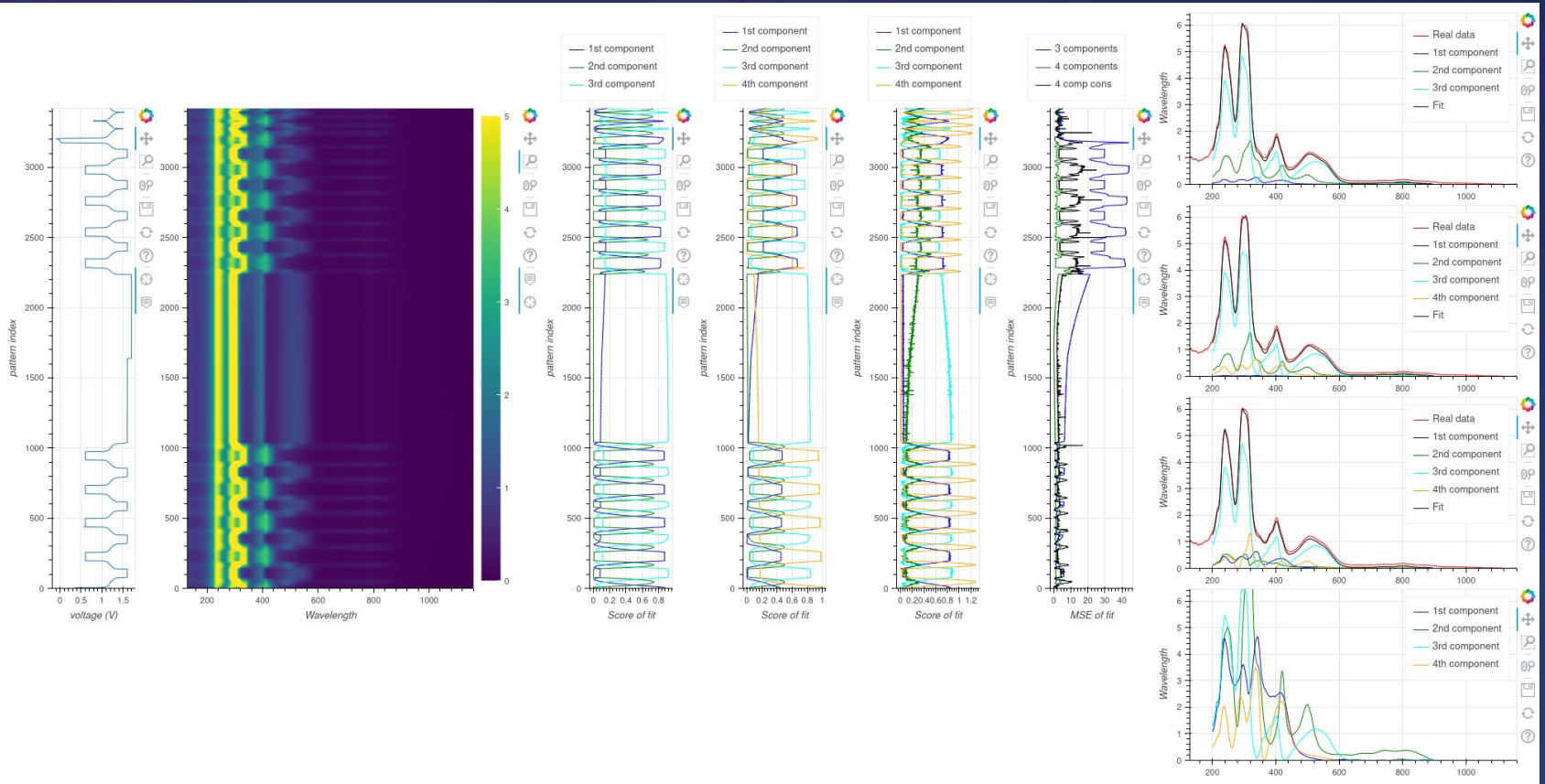


Purposefully simple to add new “blocks”:

- File loader
- Interactive plot (Bokeh)
- Baseline corrections, background subtraction, schemas, models, validation, search, QC, comparative analysis…

in situ UV-Vis data block

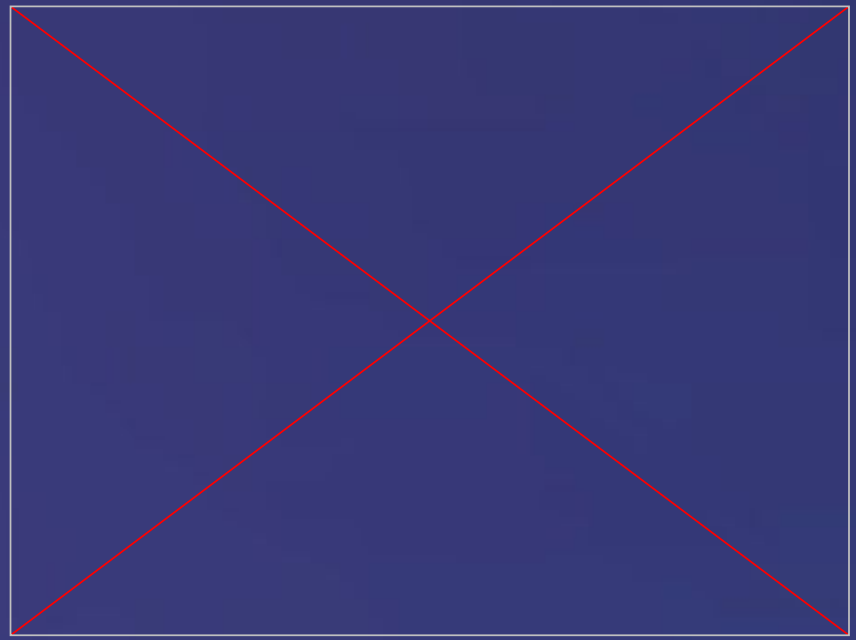
Ben Smith, Dominic Hey, Clare Grey (submitted)



datalab is open-source software (MIT)

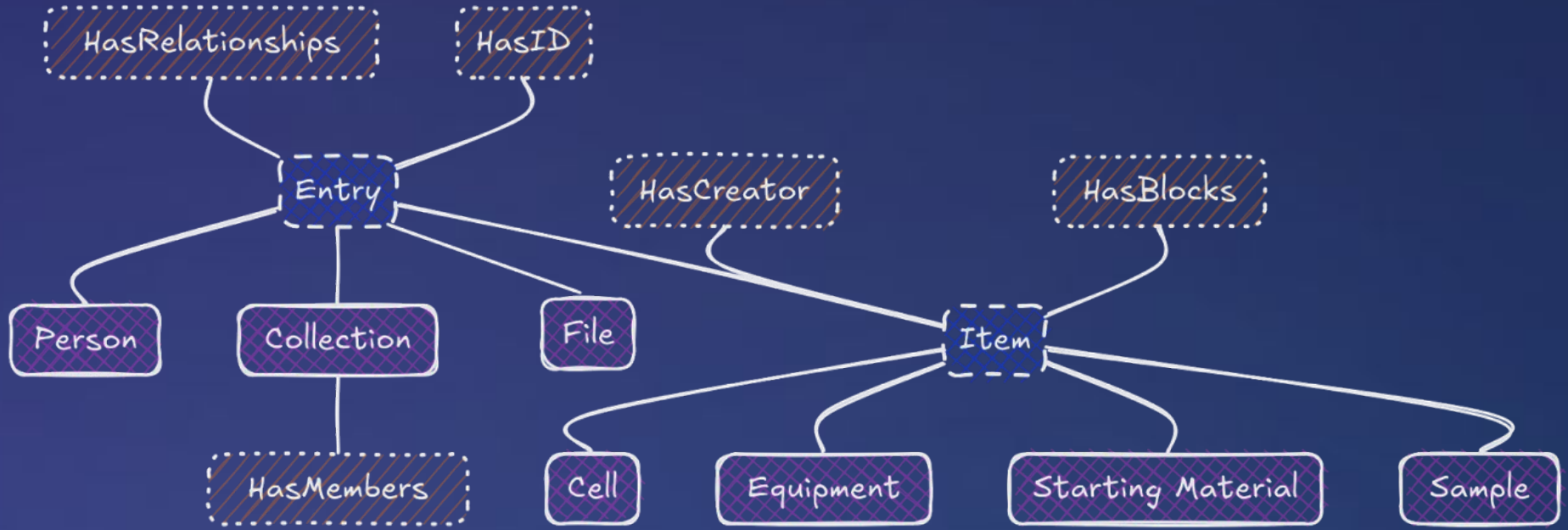
- Collaborative development is performed on GitHub using modern best practices for open source software development.
  - ~30,000 lines of code
  - 650+ merged PRs
- Automated CI with testing of server and GUI

- Simplified deployment via Docker
- Automated cloud deployments with Ansible & Terraform
- Decentralizing maintenance costs via the community
- Option to fund via consultancy

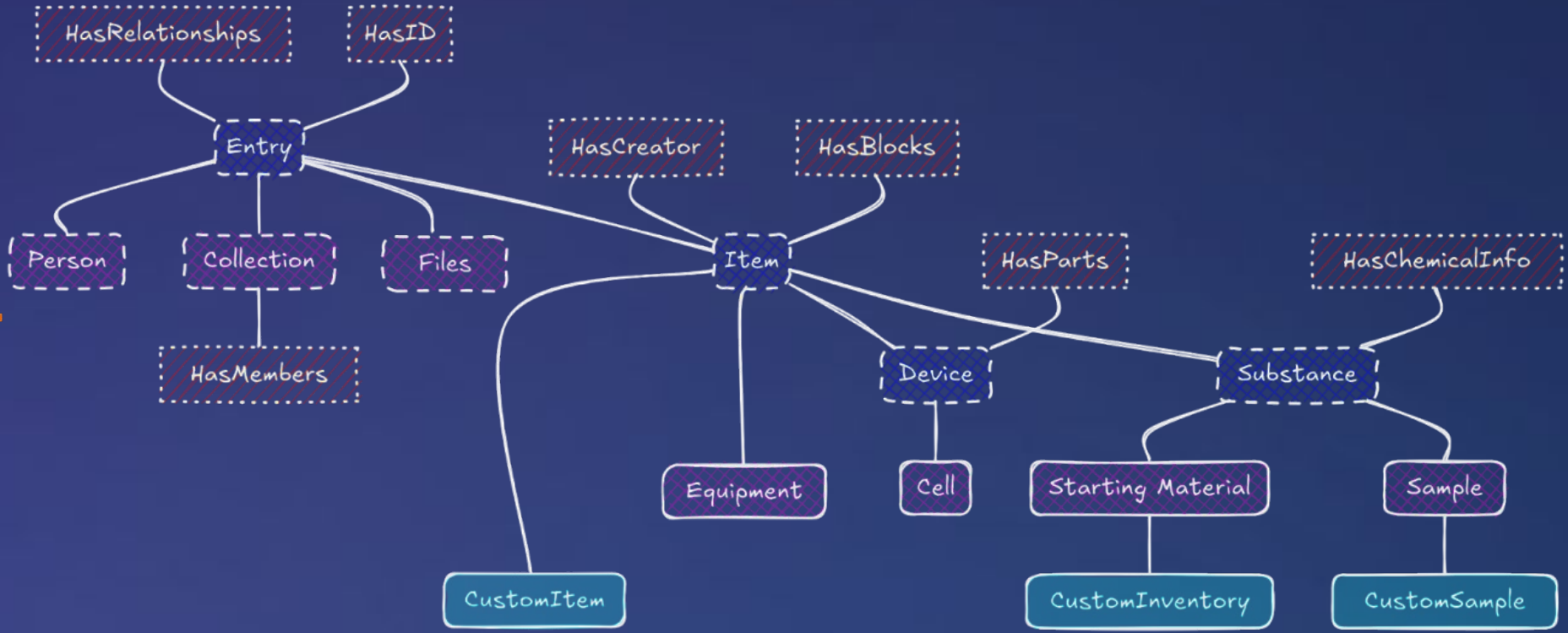


github.com/datalab-org/datalab

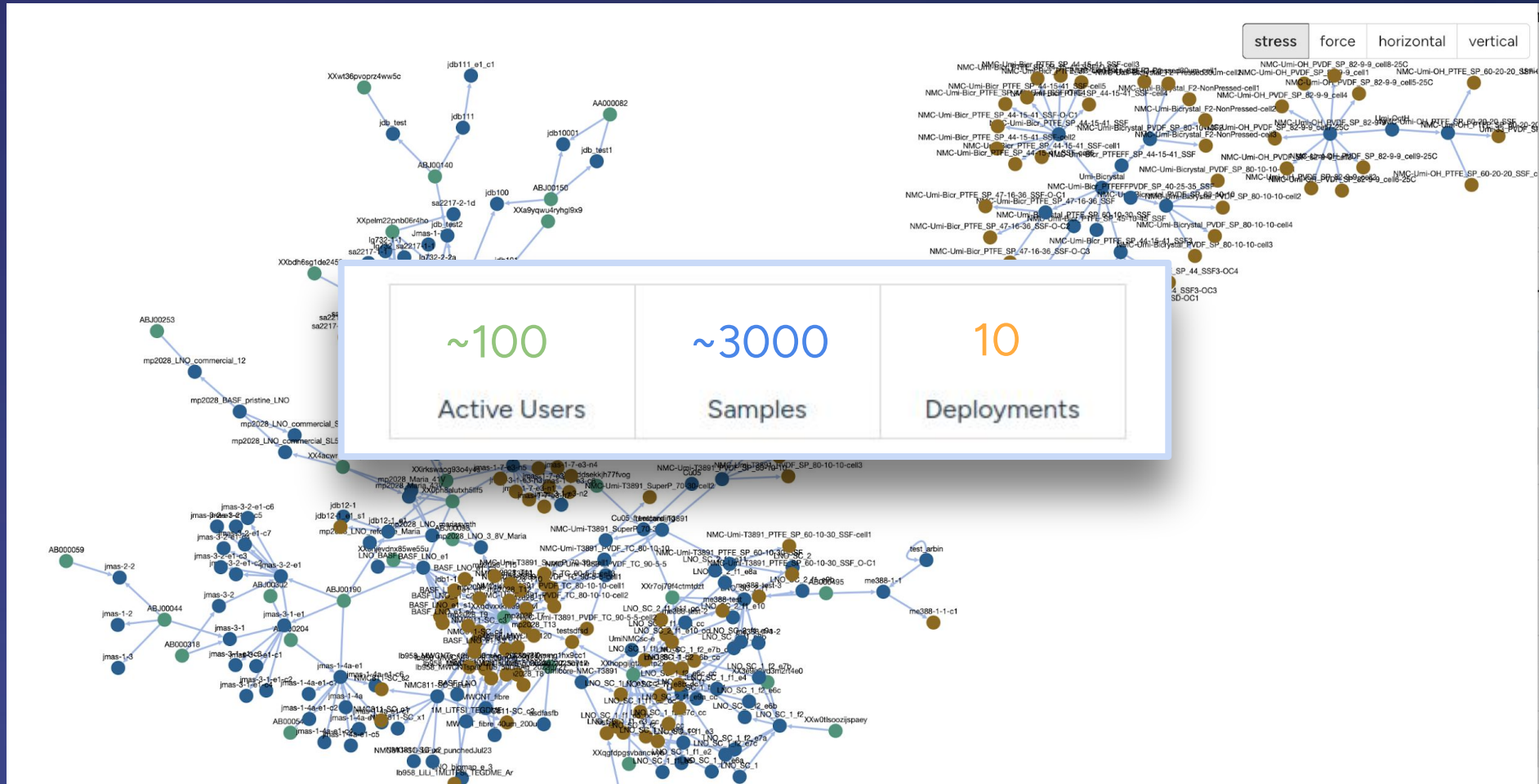
datalab data model: traits-based



datalab data model: custom schemas



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



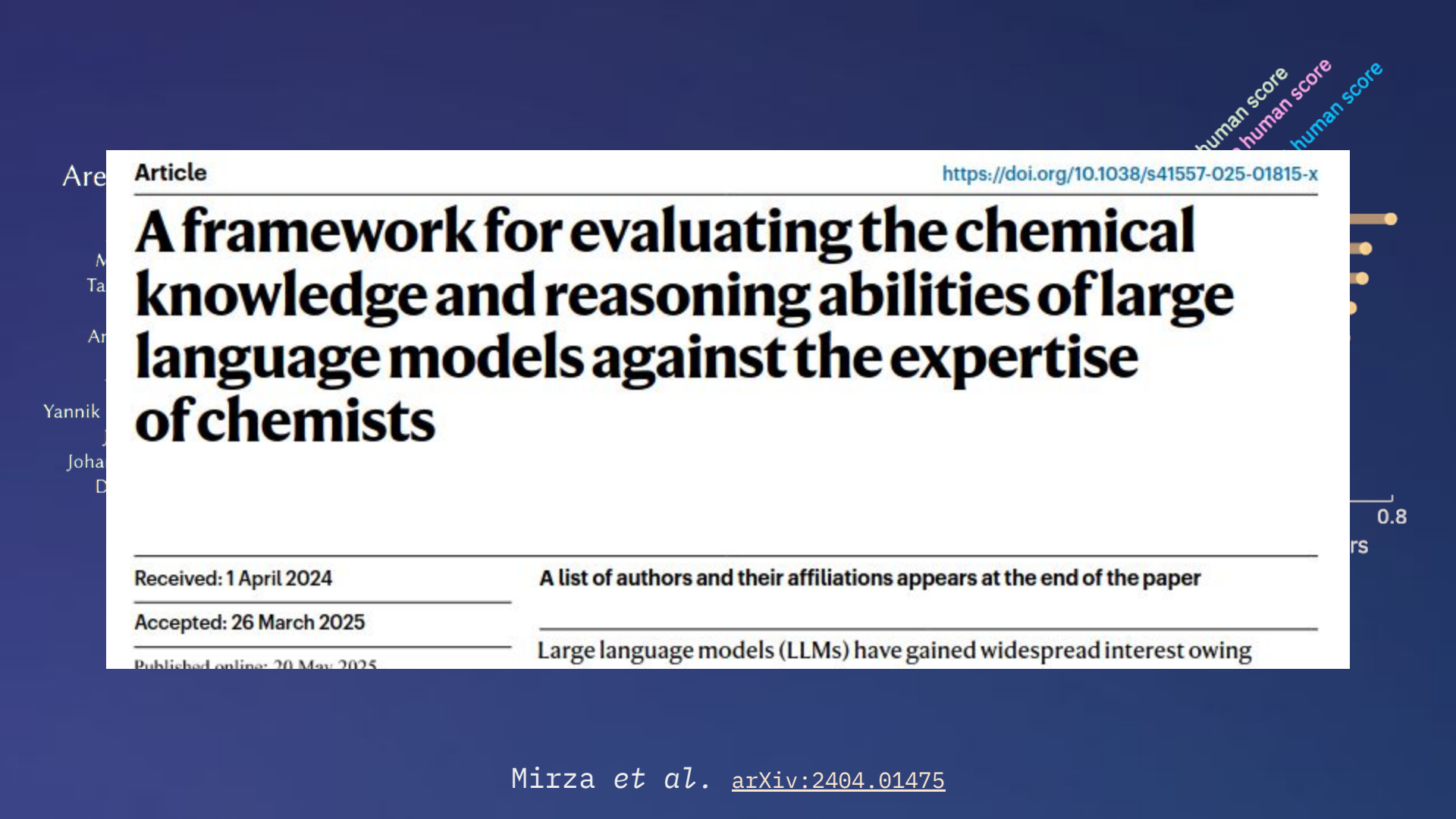
Growth of the Bocarsly group data

Prof Joshua Bocarsly (Cambridge ✈️ UHouston)

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)
- → AIChemy & DigiBat(Imperial)
- → ConCat lab (TUBerlin)
- → PSDI Polymers (UK)
- → 20+ person startup (USA)
- → 100+ person scale-up (UK)
- → Menkin group (Cambridge)
- → Solid state chemistry group (Edinburgh)
- → Materials Research Lab (UCSB)



Article

https://doi.org/10.1038/s41557-025-01815-x

A framework for evaluating the chemical knowledge and reasoning abilities of large language models against the expertise of chemists

Received: 1 April 2024

Accepted: 26 March 2025

Published online: 20 May 2025

A list of authors and their affiliations appears at the end of the paper

Large language models (LLMs) have gained widespread interest owing

Mirza *et al.* arXiv:2404.01475

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**

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**

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 ($NaCoO_2$) 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:

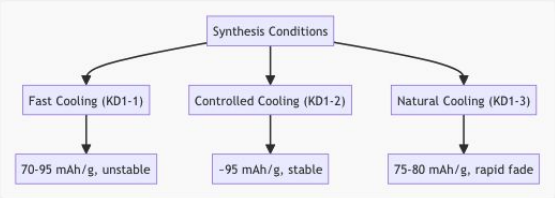
The samples (KD1-1, KD1-2, KD1-3) were all synthesized similarly at 850°C under O_2, but with some key differences:

- 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 (with Josh Bocarsly & Yue Wu)

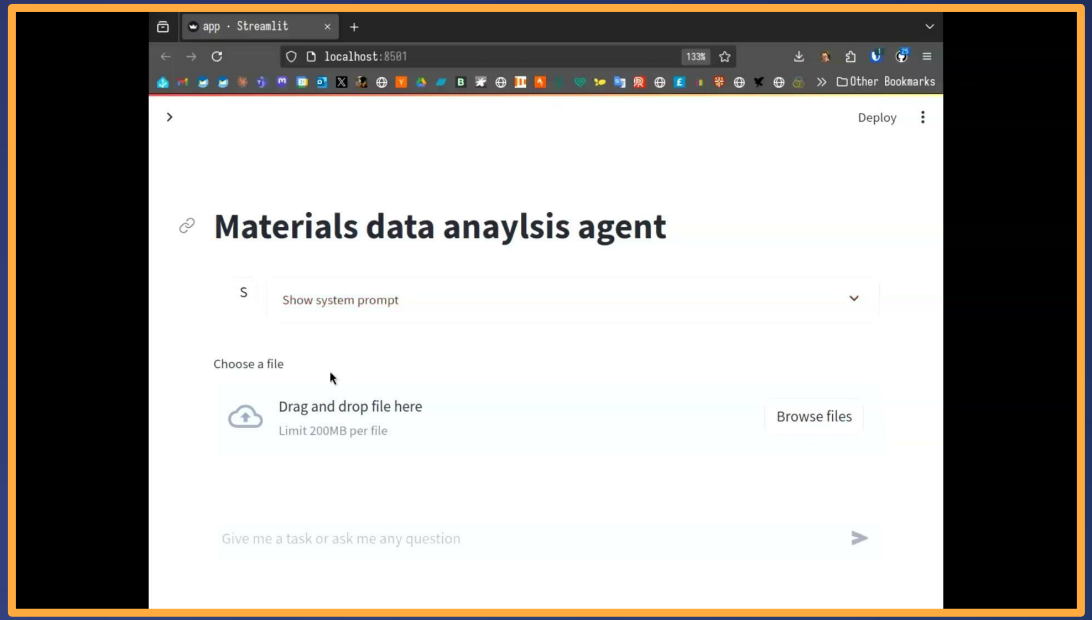
- Agentic interface to *datalab* -- still chat driven, but now uses our Python API directly (LLM hackathon 2024)
- Can automate complicated (not complex) data analysis tasks, e.g.
  - Search/RAG
  - Writing tools that we don't have time for
  - Multimodal/historic data input from spreadsheets, physical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

Zimmerman *et al* arXiv:2411.15221 (2024)

YeLLowhaMMer

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
- Saves us writing UI workflows for complicated features!

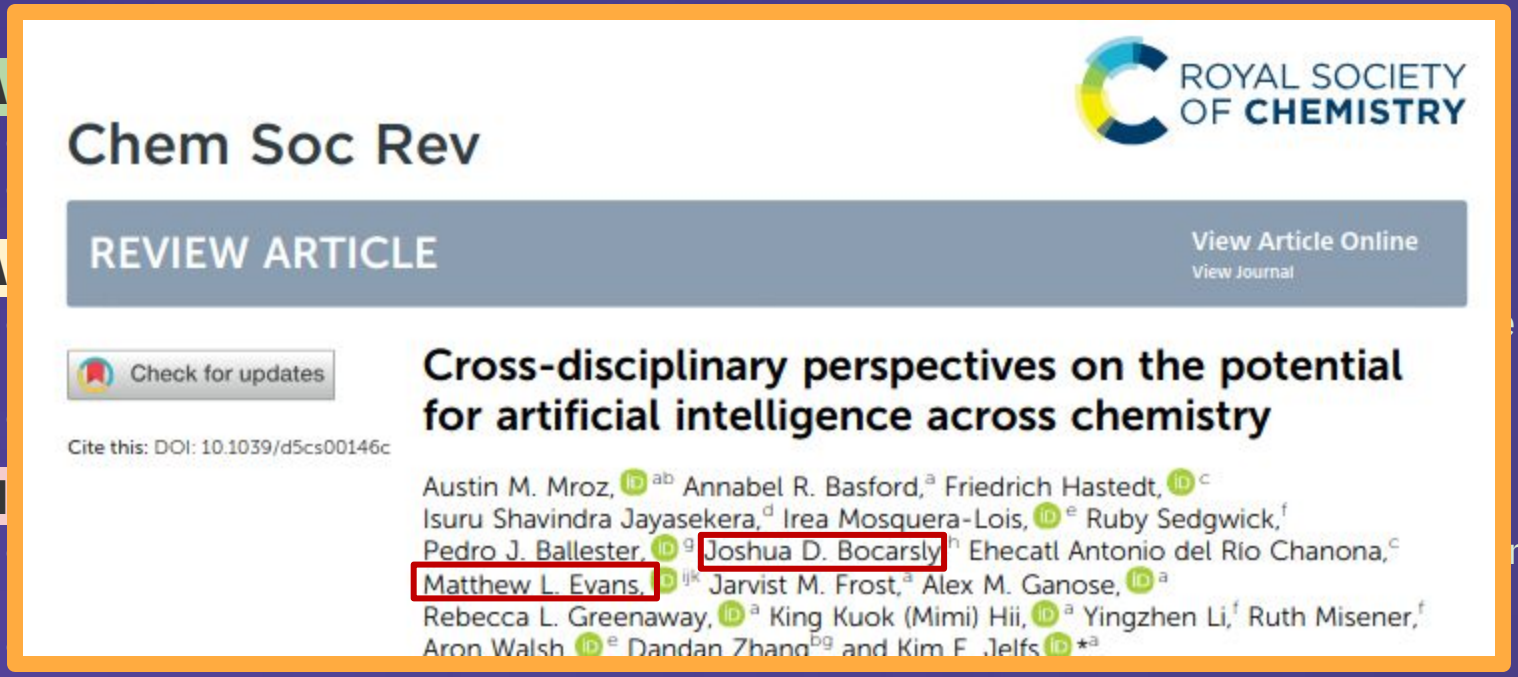


https://yellowhammer.streamlit.app

Zimmerman *et al* arXiv:2411.15221 (2024)

AI + *datalab*

- **A**
- **A**
- **M**



Chem Soc Rev

ROYAL SOCIETY OF CHEMISTRY

REVIEW ARTICLE

View Article Online
View Journal

Check for updates

Cite this: DOI: 10.1039/d5cs00146c

Cross-disciplinary perspectives on the potential for artificial intelligence across chemistry

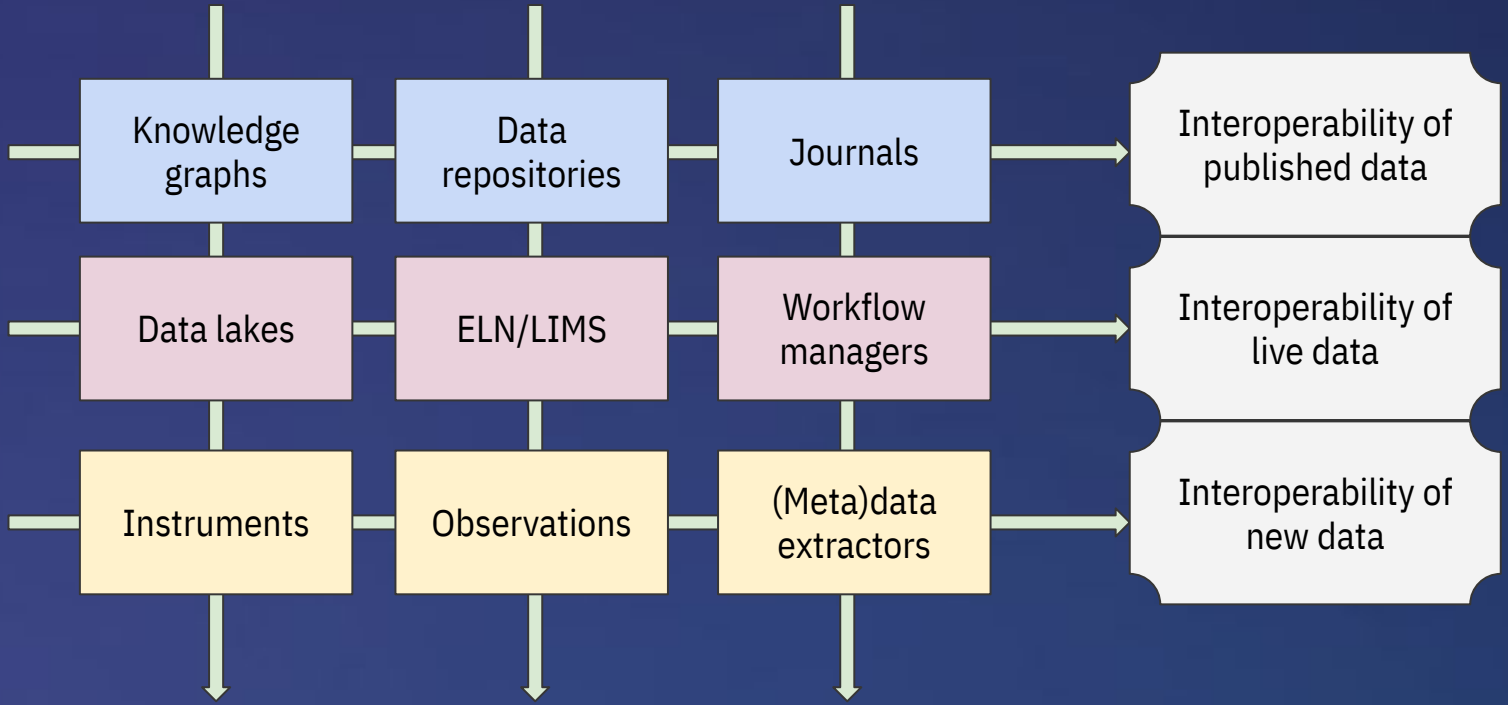
Austin M. Mroz, [ab] Annabel R. Basford,[a] Friedrich Hastedt, [c] Isuru Shavindra Jayasekera,[d] Irea Mosquera-Lois, [e] Ruby Sedgwick,[f] Pedro J. Ballester, [g] Joshua D. Bocarsly [h] Ehecatl Antonio del Rio Chanona,[c] Matthew L. Evans, [ijk] Jarvist M. Frost,[a] Alex M. Ganose, [a] Rebecca L. Greenaway, [a] King Kuok (Mimi) Hii, [a] Yingzhen Li,[f] Ruth Misener,[f] Aron Walsh [e] Dandan Zhang[bg] and Kim E. Jelfs [*a]

data,

n an

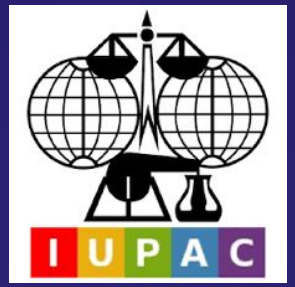
- Access across other platforms, ELNs, autonomous labs etc.
- Now technically feasible, but limited fit with current scientific problems

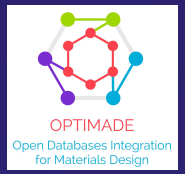
Interoperability hierarchy

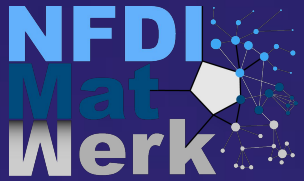


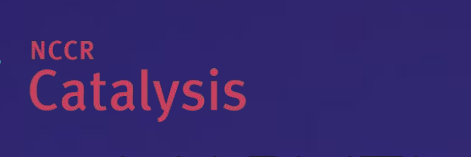
Machine-Actionable Data Interoperability for Chemical Sciences (MADICES)

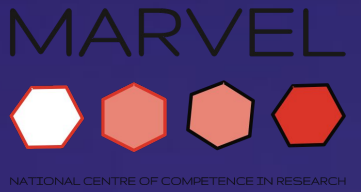
- CECAM workshop series: 2022 and 2024, upcoming in Oct 2025
- Practicable standards for interoperability across:











Datatractor (MaRDA extractors working group)

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

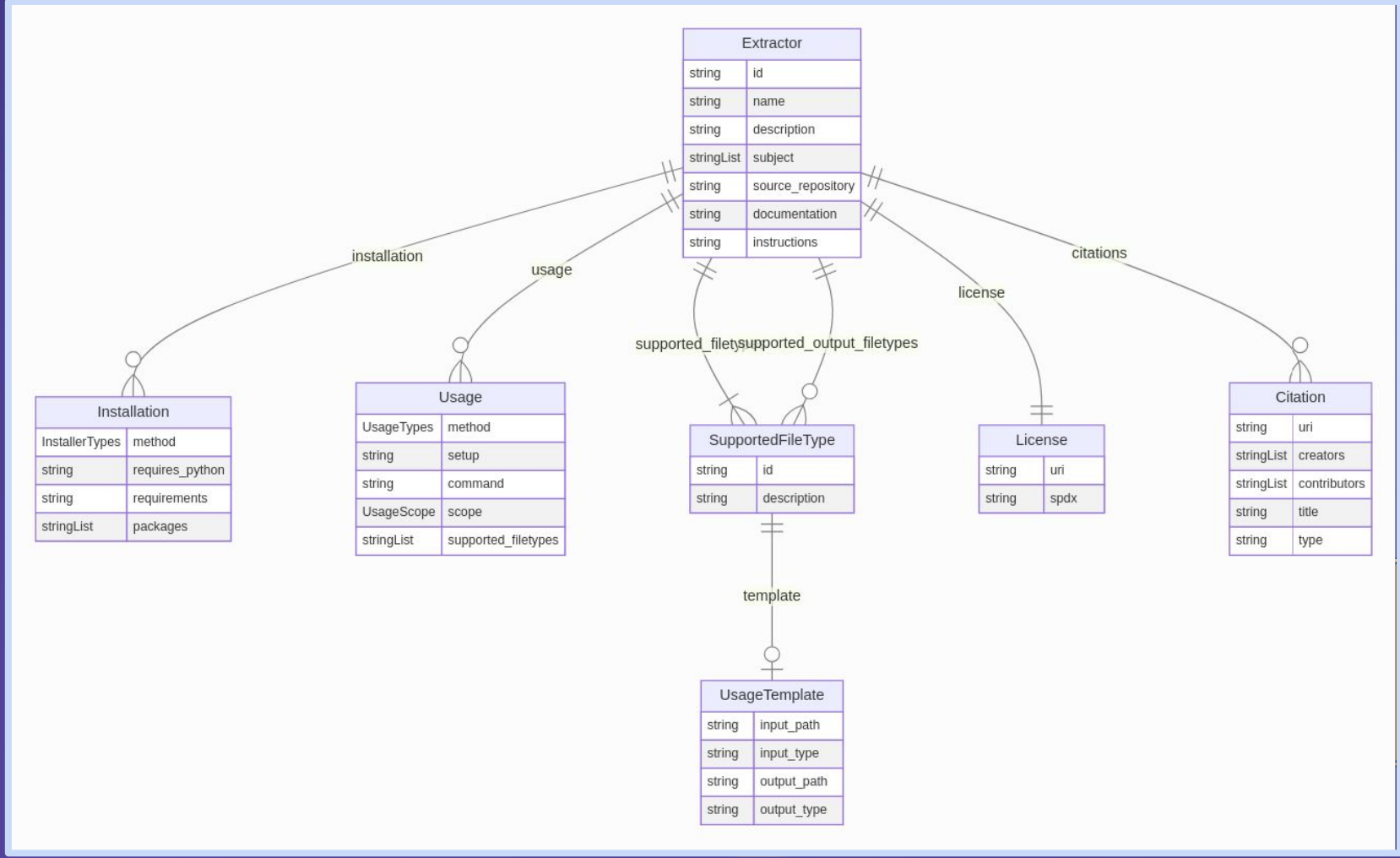
Supporting all possible instrument file format types in a single platform/framework is not possible

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

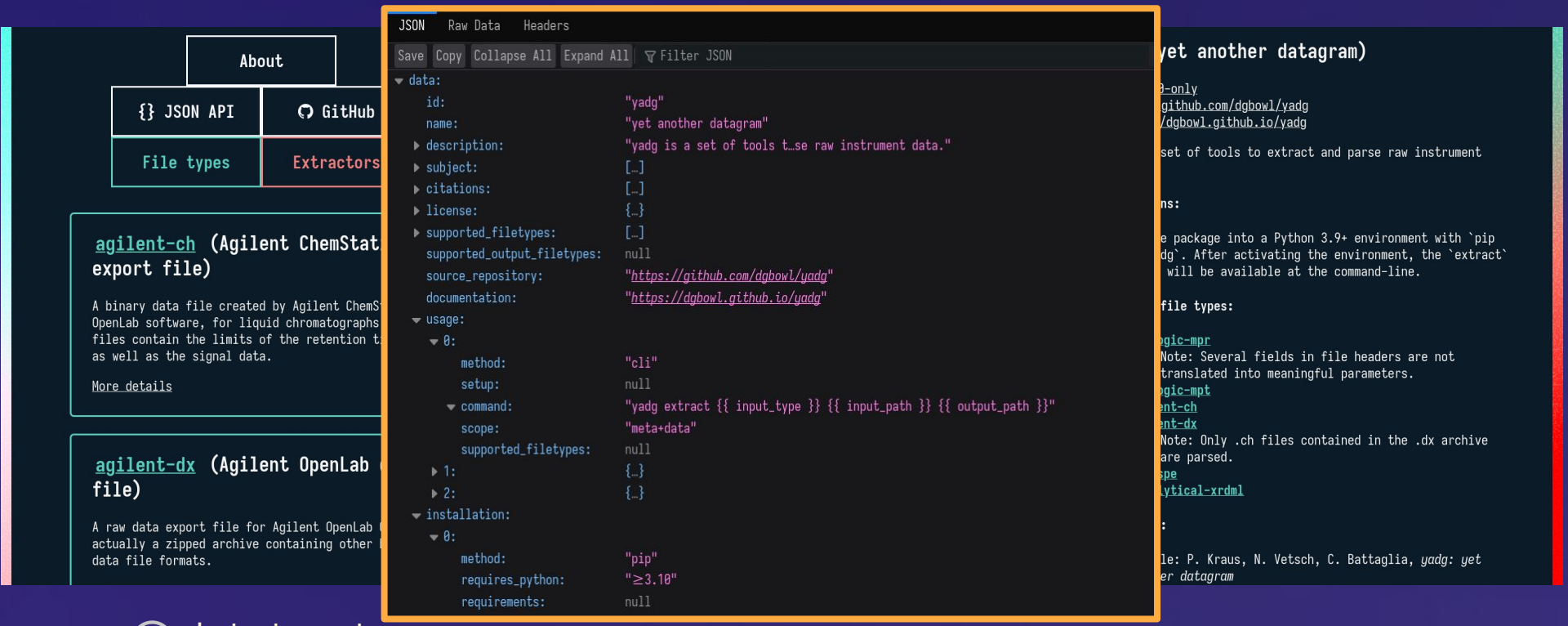
datatractor

Evans *et al, MRS Bulletin (in press)* arXiv:2410.18839 (2025)


Extractor
string id
string name
string description
stringList subject
string source_repository
string documentation
string instructions
installation
usage
supported_filetypes
supported_output_filetypes
license
citations
Installation
InstallerTypes method
string requires_python
string requirements
stringList packages
Usage
UsageTypes method
string setup
string command
UsageScope scope
stringList supported_filetypes
SupportedFileType
string id
string description
template
UsageTemplate
string input_path
string input_type
string output_path
string output_type
License
string uri
string spdx
Citation
string uri
stringList creators
stringList contributors
string title
string type

Datatractor (MaRDA extractors working group)

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



datatractor

Evans *et al*, *MRS Bulletin (in press)* arXiv:2410.18839 (2025)

Datatractor (MaRDA extractors working group)

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

```python
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 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 2
    ox or red                  (uts) int64 6kB 1 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 0
    control changes            (uts) int64 6kB 1 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 0
    counter inc.               (uts) int64 6kB 0 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

Evans *et al, MRS Bulletin (in press)* arXiv:2410.18839 (2025)

Call to arms

- **Dealing with tricky interconnected data?**
- **Interested groups/companies**
- **Individual researchers doing this work already, plugin or schema development?**

+ Monthly *datalab* meetings
+ Communal Slack (see invite link on GitHub)
+ Demo server demo.datalab-org.io
+ *datalab-org* and *datatractor* on GitHub
+ MADICES workshop October 2025 at Paul Scherrer Insitut, Switzerland (https://madices.github.io)

Future

- (Just!) funded for 3 years in Cambridge starting October, focusing on roadmap for *datalab* v1:
  - OPTIMADE for search and shared JSON-LD context
  - Schema customisation and extension
  - Easy publishing into repositories with full context
  - Broader and narrower plugins
- Continued commercial support from https://datalab.industries and https://matgenix.com

Acknowledgements

datalab development team:

Prof Josh Bocarsly (UHouston)

Ben Charmes

Dr Yue Wu

Dr Ben Smith (artist's impression)

Current funding with:

- Prof Gian-Marco Rignanese (UCLouvain)
- Dr David Waroquiers (Matgenix)

Plus close collaborators:

- Prof Clare Grey & the Grey Group (Cambridge)
- Prof Matt Cliffe (Nottingham)
- Dr Peter Kraus (TUBerlin)
- *datalab* federation
- PSDI
- Faraday Institution (CATMAT)