ME 5374-ST

# Machine Learning for Materials Science and Discovery

Fall 2025

Asst. Prof. Peter Schindler

## Lecture 9 – Data-Types and Databases in Materials Science

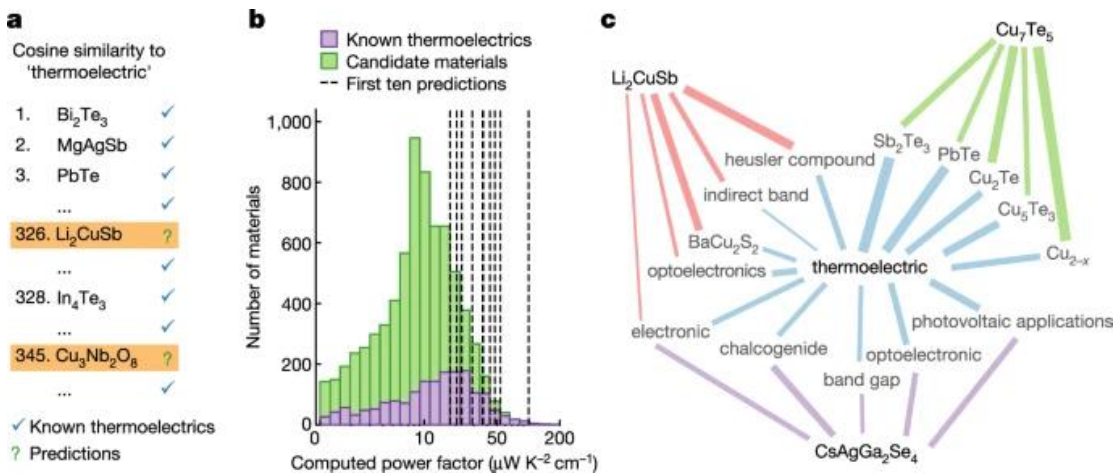
- Data-Types in Materials Science
- Atomistic File Types
- Application Programming Interfaces (APIs)
- Materials Databases
- FAIR Data Principles

# Types of Materials Data

Text

Scientific Literature

NLP and LLMs




Tshitoyan, V., *et al. Nature* **571**, 95–98 (2019)

# Types of Materials Data

**Experimental**

**Spectra, Images**

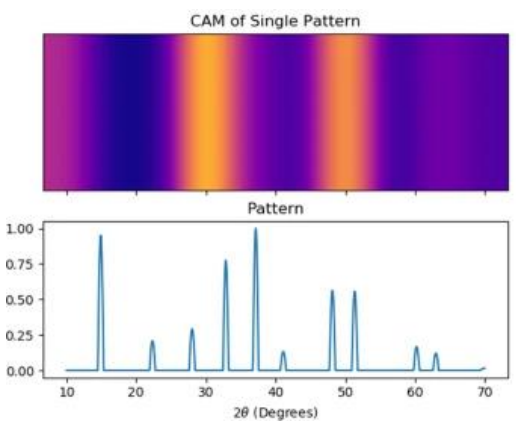
**Text**

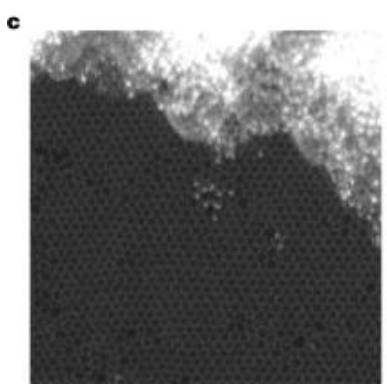
Micrographs

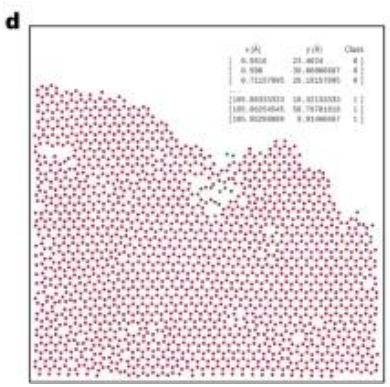
XRD, TEM, etc.

Scientific Literature

NLP and LLMs



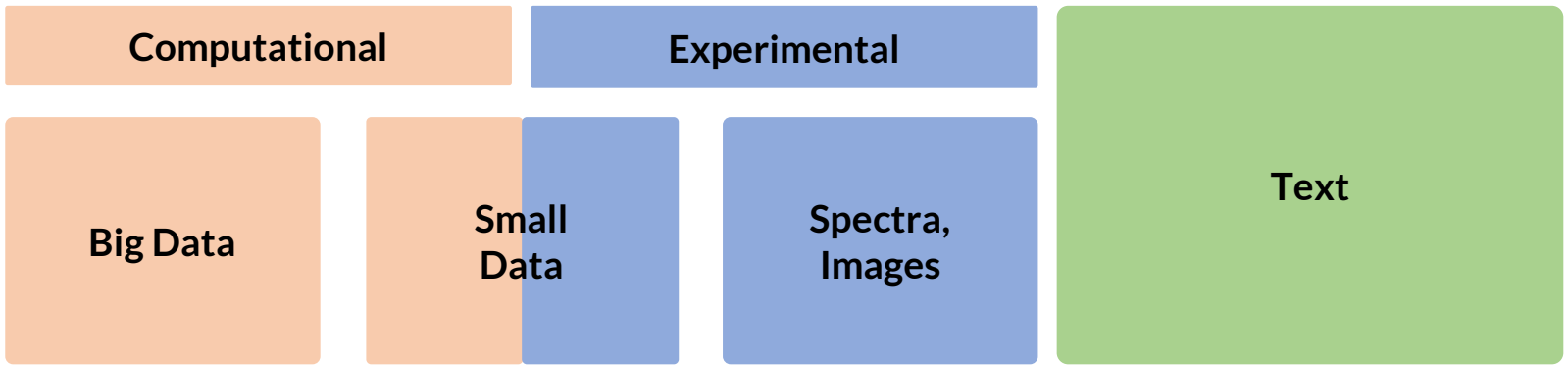






Oviedo, F., et al. *Comput Mater* **5**, 60 (2019).

Ziatdinov, M., et al. *Nat Mach Intell* **4**, (2022).

# Types of Materials Data




The Materials Project

Open Catalyst Project
Using AI to model and discover new catalysts to address the energy challenges posed by climate change.

NIST JARVIS DFT

OPTIMADE
Open Databases Integration for Materials Design

# ML Paradigms in Materials Science

**Model-centric AI**
How change the model/architecture to improve performance?

**Data-centric AI**
How systematically change data (x/y) to improve performance?

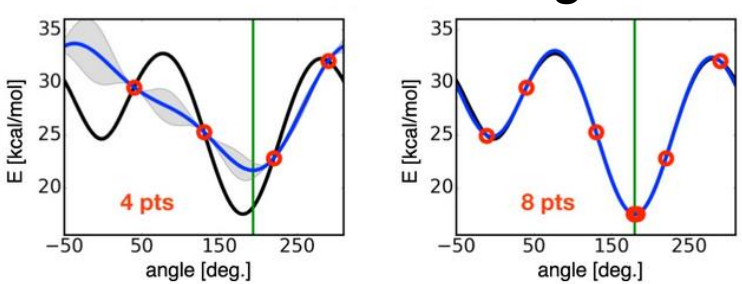
Big Data → "Good Data"

## Shallow ML + feature engineering

Computational | Experimental

Big Data | Small Data | Spectra, Images

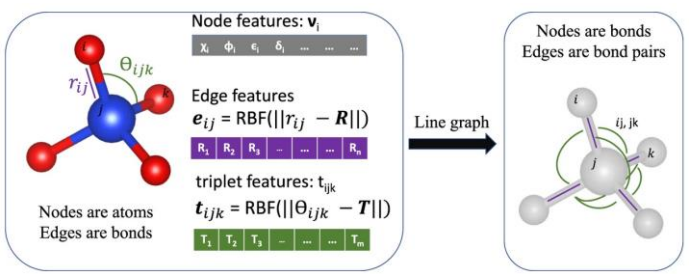
## Active Learning




Computational | Experimental

Big Data | Small Data | Spectra, Images

## Deep Learning




Computational | Experimental

Big Data | Small Data | Spectra, Images

## Transfer Learning

Computational | Experimental

Big Data | Small Data | Spectra, Images



Todorović, *et al. npj Comput Mater* **5**, 35 (2019)

Choudhary & DeCost, *npj Comput Mater* **7**, 185 (2021)

# Crystal Information File (CIF)

```
# generated using pymatgen
data_Mo2NO
_symmetry_space_group_name_H-M   'P 1'
_cell_length_a   3.10000000
_cell_length_b   6.20000000
_cell_length_c   3.10000000
_cell_angle_alpha   90.00000000
_cell_angle_beta   90.00000000
_cell_angle_gamma   90.00000000
_symmetry_Int_Tables_number   1
_chemical_formula_structural   Mo2NO
_chemical_formula_sum   'Mo2 N1 O1'
_cell_volume   59.58200000
_cell_formula_units_Z   1
loop_
 _symmetry_equiv_pos_site_id
 _symmetry_equiv_pos_as_xyz
  1  'x, y, z'
loop_
 _atom_site_type_symbol
 _atom_site_label
 _atom_site_symmetry_multiplicity
 _atom_site_fract_x
 _atom_site_fract_y
 _atom_site_fract_z
 _atom_site_occupancy
  Mo  Mo0  1  0.00000000  0.00000000  0.00000000  1
  N  Mo  1  0.00000000  0.50000000  0.00000000  1
  Mo  Mo2  1  0.50000000  0.70000000  0.70000000  1
  O  O3  1  0.00000000  0.25000000  0.25000000  1
```

No explicit symmetry

Lattice

Symmetry/Equivalent Positions

Atomic sites (species, occupation, coordinates,…)

```
_space_group_IT_number           29
_symmetry_space_group_name_Hall  'P 2c -2ac'
_symmetry_space_group_name_H-M   'P c a 21'
_cell_angle_alpha                90
_cell_angle_beta                 90
_cell_angle_gamma                90
_cell_length_a                   5.5833
_cell_length_b                   5.5892
_cell_length_c                   5.5812
_cell_formula_units_Z            4
_cell_volume                     174.168
_database_code_amcsd             0005243
_exptl_crystal_density_diffrn    6.328
_cod_original_formula_sum        'Co As S'
_cod_database_code               9004218
loop_
_space_group_symop_operation_xyz
x,y,z
1/2+x,-y,z
1/2-x,y,1/2+z
-x,-y,1/2+z
```

More details on syntax:
https://www.iucr.org/resources/cif/spec/version1.1/cifsyntax

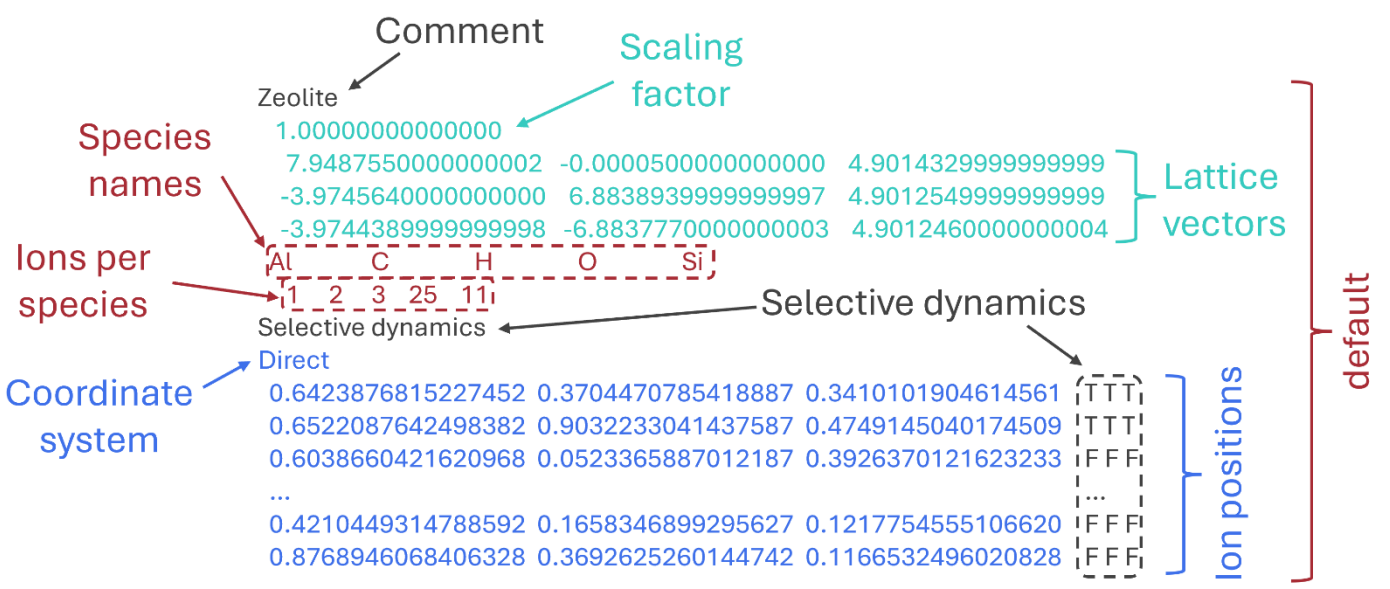
# XYZ File

```
<number of atoms>
comment line
<element> <X> <Y> <Z>
...
```

- Stores Cartesian coordinates (in Angstrom) and atom types of a molecule
- Extended XYZ format allows definition of lattice (and other information) in the comment line

```
8
Lattice="5.44 0.0 0.0 0.0 5.44 0.0 0.0 0.0 5.44" Properties=species:S:1:pos:R:3 Time=0.0
Si        0.00000000      0.00000000      0.00000000
Si        1.36000000      1.36000000      1.36000000
Si        2.72000000      2.72000000      0.00000000
Si        4.08000000      4.08000000      1.36000000
Si        2.72000000      0.00000000      2.72000000
Si        4.08000000      1.36000000      4.08000000
Si        0.00000000      2.72000000      2.72000000
Si        1.36000000      4.08000000      4.08000000
```

# POSCAR File (VASP)



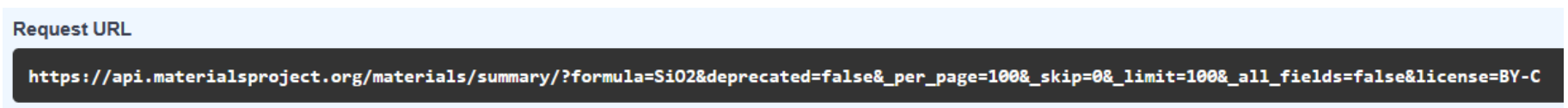

- Crystal structure input file for VASP (Density Functional Theory)
- Coordinate systems: “Direct” (=Fractional) or “Cartesian”
- More details: https://www.vasp.at/wiki/index.php/POSCAR

# Application Programming Interfaces (APIs)

API is a set of defined functions, procedures, methods, or classes which enable a **structured way of exchanging data** between programs. *It facilitates…*

- uploading, examining, and downloading of data (without manual steps)
- selectively querying (fetch data that adhere to constraints)
- keeping track of versions
- standardization of interface

*Representational State Transfer* (REST API or RESTful interface) is the most commonly used framework that interfaces via stateless endpoints through request URLs.
(see here for details: https://aws.amazon.com/what-is/restful-api/ )




https://api.materialsproject.org/docs

## OPTIMADE

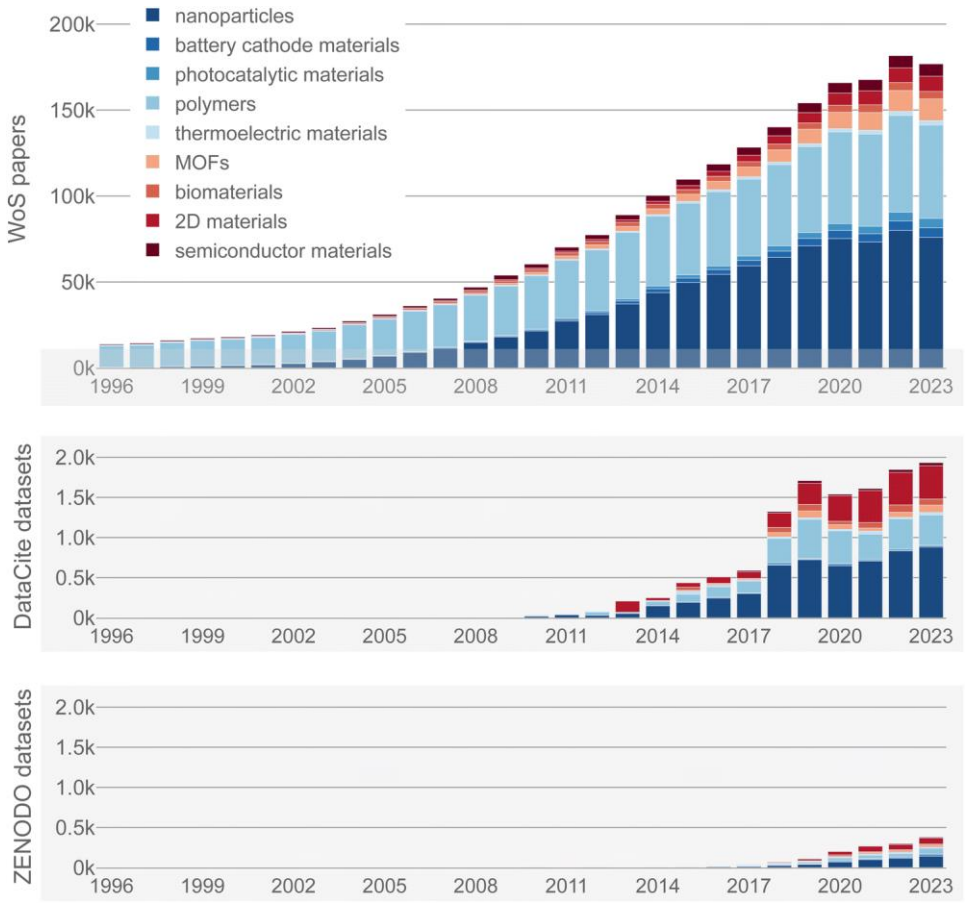
OPTIMADE
Open Databases Integration for Materials Design

API standard developed for materials science and adopted by many large-scale databases

# Data Sources

| Source type | Access methods | Examples | Access type | Drawbacks |
|---|---|---|---|---|
| **Papers & textbooks** | Hand extract (copy/paste)<br>Download PDF/LaTeX/XML<br>Text-mine with NLP/LLMs | Journal articles, Landolt-Börnstein (https://materials.springer.com/bookshelf) | Manual (unless text-mining with NLP/LLMs) | Time-consuming; error-prone; parsing PDFs difficult; coverage limited |
| **Public websites (no API)** | Manual download, copy/paste<br>Web crawling (legal grayzone) | University pages, project web pages, vendor datasheets | Manual | Data formats inconsistent; scraping may be legally/ethically restricted; fragile if site changes |
| **Public repositories (no API)** | Manual download of files or bulk repositories | Figshare, Zenodo, NIMS (https://mits.nims.go.jp/) | Manual | No programmatic access; updates require re-download; can be very large |
| **Public databases (with API)** | Programmatic access (REST/GraphQL APIs) | Materials Project, PubChem, OQMD | API | Rate limits (sometimes)<br>Learning curve for API queries |
| **Commercial databases** | API + web portals; licensed access | Proprietary vendor/industry databases<br>ICSD, Pauling File | API (licensed) | Expensive; restrictive licenses; not open for redistribution |
| **Automated experiments (own data)** | Instrument/robot control, high-throughput workflows | Autonomous labs; Laboratory Information Management Systems (LIMS)-managed workflows | Manual (experiment) + API (instrument control) | Very costly; needs infrastructure; time-intensive |

# Potential Value of Text-Mining






***Chem. Soc. Rev.***, 2025, **54**, 1125-1150

# Recap: Database File Formats

| Format | Structure | Human-readable | Loading speed | Strengths | Limitations |
|---|---|---|---|---|---|
| **CSV** (Comma-Separated Values) | Rows and columns separated by delimiters (flat, tabular) | ✅ Yes | ⚡ Medium | Simple, lightweight, works well with spreadsheets | No nested structures, no data types (all text) |
| **JSON** (JavaScript Object Notation) | Key–value pairs, lists, nested objects (hierarchical) | ✅ Yes | 🐢 Slow | Handles complex data, preserves data types, widely used in APIs (Application Programming Interfaces) | More verbose, harder to edit manually than CSV |
| **HDF5** (Hierarchical Data Format) | Binary format with hierarchical groups and datasets | ❌ No | ⚡ Fast | Efficient for very large datasets, supports metadata | Requires special libraries (e.g., h5py, PyTables) |
| **Pickle** | Python-specific serialized objects | ❌ No | ⚡ Fast | Can store almost any Python object easily | Not portable outside Python, unsafe if source is untrusted |
| **YAML** (YAML Ain't Markup Language) | Human-readable key–value and nested structure | ✅ Yes | 🐢 Slow | More flexible and readable than JSON, allows comments | Less standardized than JSON |

# Examples: Property-Focused Materials Databases (mostly computational)

| name | structure information | mechanical properties | thermal properties | electronic properties | API[a] | data license | refs |
|---|---|---|---|---|---|---|---|
| Materials Project | Y | Y | Y | Y | Y | CC BY 4.0 | 85 |
| Open Quantum Materials Database | Y | N | Y | Y | Y | CC BY 4.0 | 86 |
| AFLOW for Materials Discovery | Y | Y | Y | Y | Y | *b* | 87 |
| Novel Materials Discovery (NOMAD) | Y | Y | Y | Y | Y | CC BY 4.0 | 88 |
| Open Materials Database | Y | N | Y | Y | Y | CC BY 4.0 | 89 |
| Citrine Informatics | Y | Y | Y | Y | Y | CC BY | 90 |
| Materials Platform for Data Science (MPDS) | Y | Y | Y | Y | Y | CC BY 4.0 | 91 |
| AiiDA/Materials Cloud | Y | Y | Y | Y | Y | Varies | 92, 93 |
| NREL MatDB | Y | N | Y | Y | N | Own license | 94 |
| NIST TRC Alloy Data | N | N | Y | N | On request | Free | 95 |
| NIST TRC ThermoData | N | N | Y | N | N | NIST SRD | 96 |
| NIST JARVIS-DFT/-ML Database | Y | Y | Y | Y | Y | Public domain | 97, 98 |
| MatWeb | N | Y | Y | N | N | Paid | 99 |
| Total Materia | N | Y | Y | N | N | Paid | 100 |
| Ansys Granta (MaterialUniverse repository) | N | Y | Y | N | N | Paid | 101 |
| MATDAT | N | Y | Y | N | N | Paid | 102 |

*Machine Learning for Materials Scientists: An Introductory Guide toward Best Practices*,
A. Y.-T. Wang, R. J. Murdock, S. K. Kauwe, A. O. Oliynyk, A. Gurlo, J. Brgoch, K. A. Persson, and T. D. Sparks, *Chemistry of Materials* 2020 *32* (12), 4954-4965

# Examples: Structure-Focused Materials Databases (mostly experimental)

| name | no. records[a] | API | Data license | ref |
|---|---|---|---|---|
| Cambridge Structural Database (CSD) | 1,055,780 | Y | Paid | 103 |
| Inorganic Crystal Structure Database (ICSD) | 216,302 | N | Paid | 104 |
| Pearson's Crystal Data (PCD) | 335,000 | N | Paid | 105 |
| International Centre for Diffraction Data (ICDD) | 1,004,568 | N | Paid | 106 |
| Crystallography Open Database (COD) | 455,714 | Y | Open-access | 107 |
| Pauling File | 357,612 | Y | Paid | 108 |
| CrystMet database | 160,000 | N | Paid | 109 |

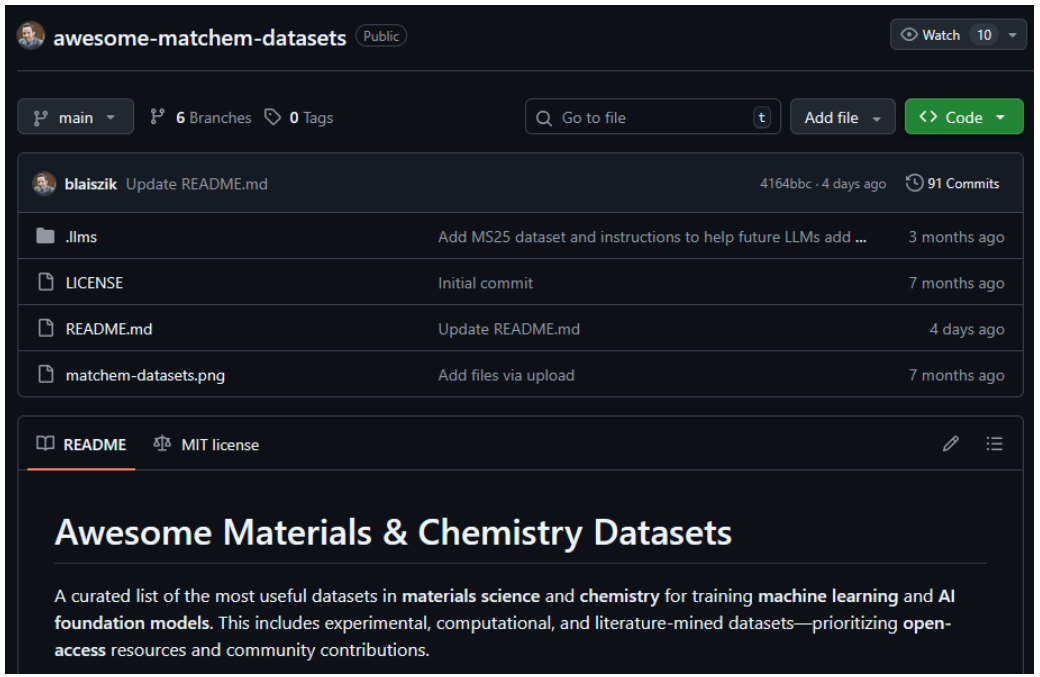
[a]Note: values for number of records were updated as of the submission date (May 2020).

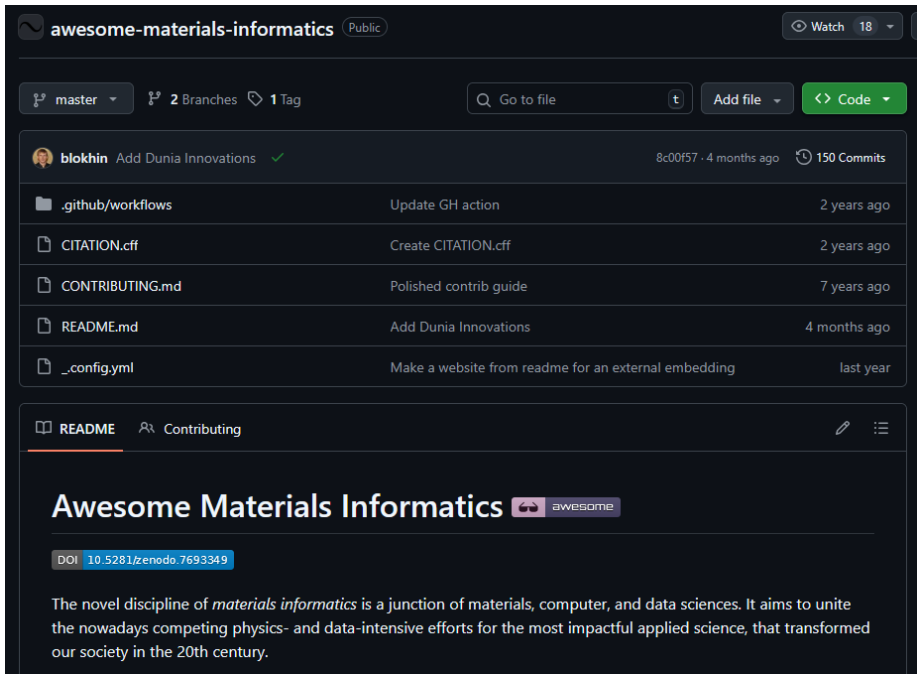
*Machine Learning for Materials Scientists: An Introductory Guide toward Best Practices*,
A. Y.-T. Wang, R. J. Murdock, S. K. Kauwe, A. O. Oliynyk, A. Gurlo, J. Brgoch, K. A. Persson, and T. D. Sparks, *Chemistry of Materials* 2020 *32* (12), 4954-4965

# Compilation of Databases and Tools in Materials Informatics

https://github.com/blaiszik/awesome-matchem-datasets

https://github.com/tilde-lab/awesome-materials-informatics



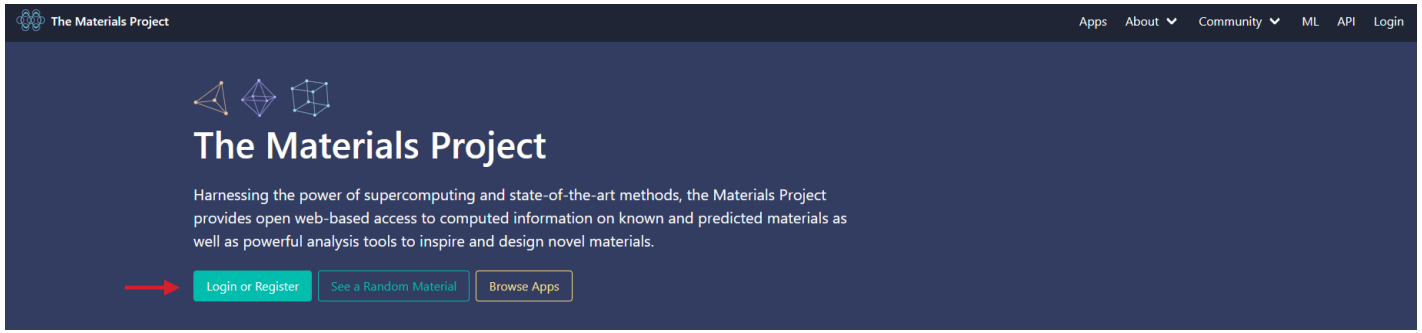



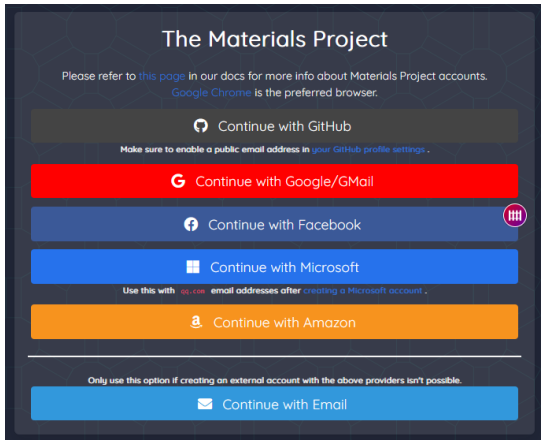
# FAIR Data Principles

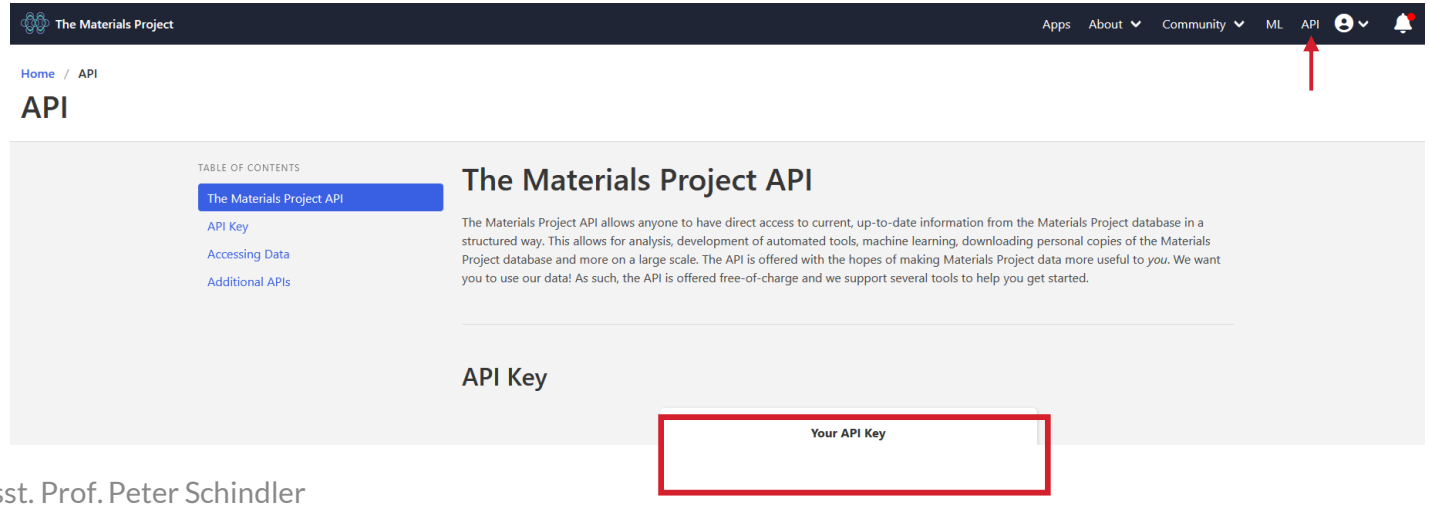
| Letter / Term | Purpose | Example Solution |
|---|---|---|
| **F – Findable** | Ensure data can be located easily by humans/machines | Assign persistent identifiers (DOIs)<br>Provide searchable metadata in repositories |
| **A – Accessible** | Make data retrievable via standard protocols | Open web protocols (HTTP/HTTPS) or APIs,<br>Ensure data can be accessed with authentication if needed (e.g., PubChem API) |
| **I – Interoperable** | Enable data integration across platforms and tools | Store crystallographic data in CIF format, adopt common ontologies<br>(e.g., OPTIMADE for materials APIs) |
| **R – Reusable** | Facilitate long-term use and reproducibility | Provide detailed metadata (methods, units), apply open licenses (e.g., CC-BY), deposit in repositories like Zenodo or Figshare |

# Materials Project: Obtaining API Key

Go to: https://materialsproject.org



# Lecture Feedback

Please, scan the QR code and take a minute to let me know how the lecture was and mention any **feedback/questions**

This form is **anonymous!**