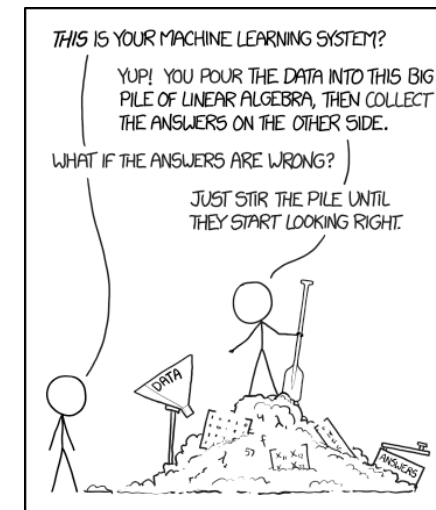
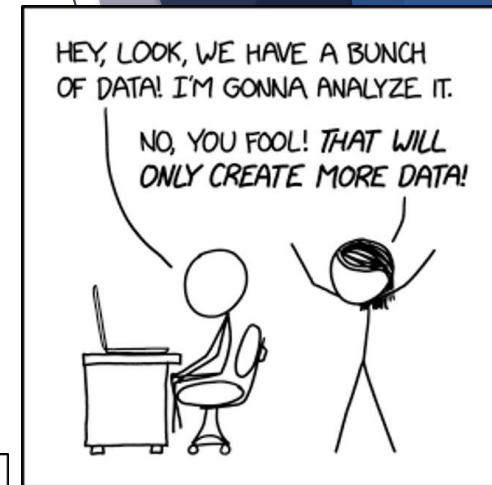


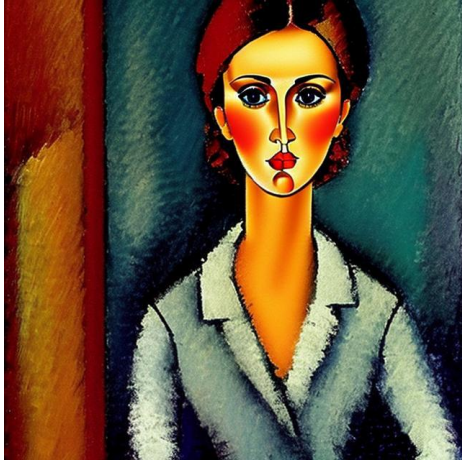
Data to knowledge[to data]

Alin M Elena

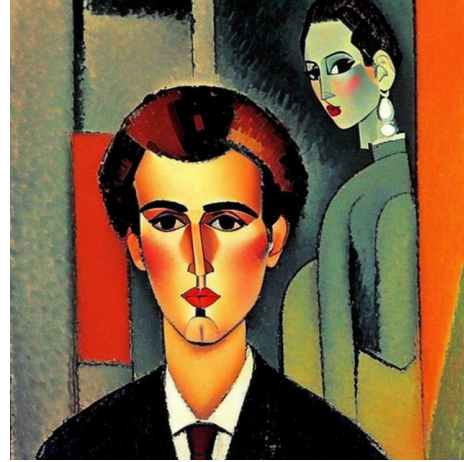
Elliott Kasoar, Federica Zanca



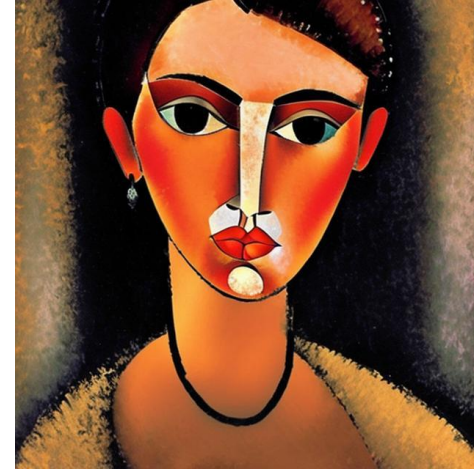
People



Alin Elena



Elliott Kasoar



Federica Zanca

Openart.ai



SCD organogram

PSDI

Facilities, Institutes & Hubs

Examples:

- Catalysis Hub
- CCFE
- Central Laser Facility
- Diamond
- Future Manufacturing Hub
- ISIS
- Royce Institute
- ATI

National Research Facilities

Examples:

- HarwellXPS
- NXCT
- NCS
- PSDS
- SuperSTEM
- UK High Field Solid-State NMR
- XMaS

Computational Initiatives

Examples:

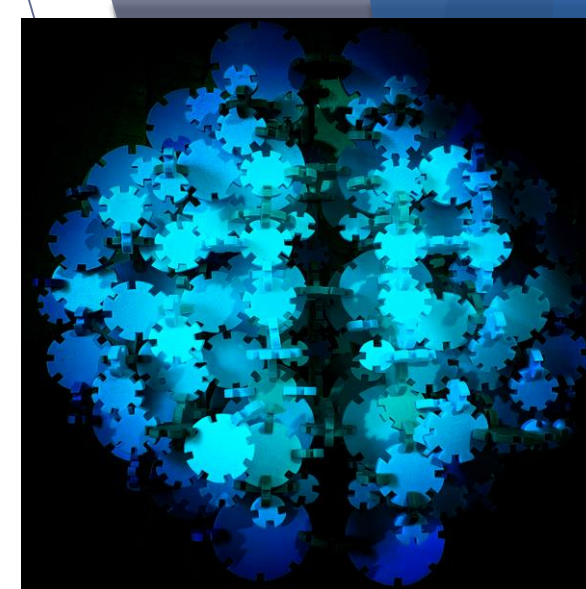
- CCP5++
- CCP9
- CCP Biosim
- CCP EngSci
- CCPi
- CCP NC
- CCP NTH
- CCP QC
- CCP SAS
- CCP Turbulence
- CCP WSI
- SSI
- UK society of RSE
- HEC Biosim
- HEC Plasma
- MCC
- UKCTRF
- UKCP
- UKTC
- CoSeC
- EPSRC Tier2
- ExCALIBUR
- STFC Hartree Centre
- ARCHER

Research Institutions, Groups and Laboratories

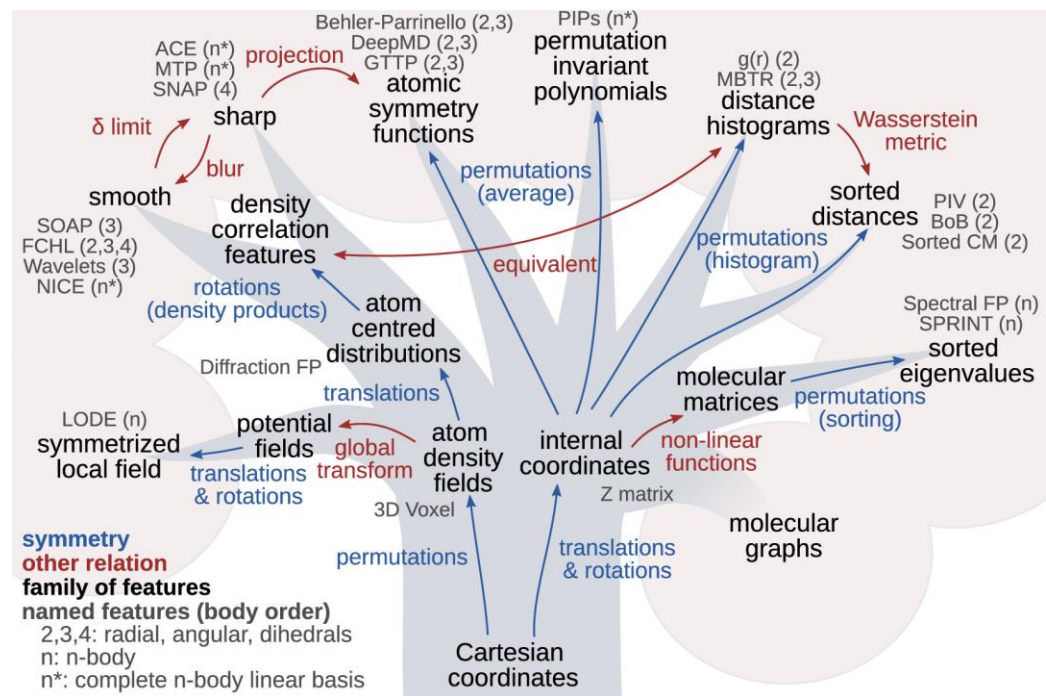
Examples:

- Equipment Infrastructures
- Equipment Facilities
- University Labs
- ELNs
- Repositories
- Local Computing Resources

PHYSICAL SCIENCES DATA INFRASTRUCTURE



Aim

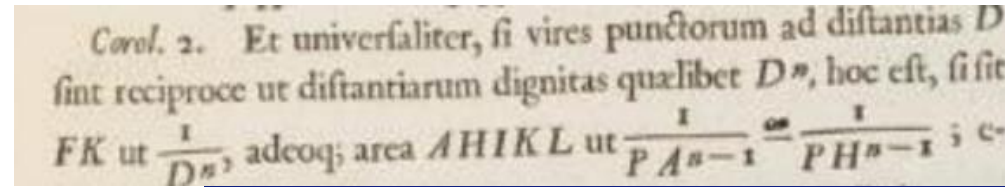


<https://doi.org/10.1021/acs.chemrev.1c00021>

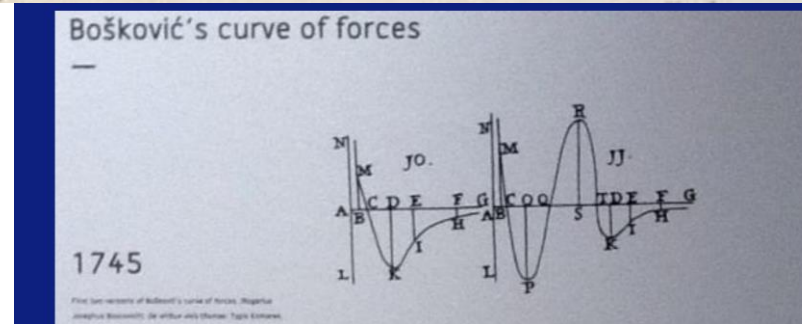
- ▶ design and deploy the hardware infrastructure to host both training data for machine learnt interatomic potentials – MLIP and potentials themselves
- ▶ **Features:** Curate, store, distribute and interrogate MLIPs and training data
- ▶ Visualisation and training workflows
- ▶ Benchmarking and validation
- ▶ Ideally, HPC integration

Science&history interlude

Newton (Principia... 1687 - $1/r^n$)



Roger Boscovich (De Viribus vivis, 1745)



Gustav Mie (1903)

$$\Phi_{12}(r) = \left(\frac{n}{n-m}\right) \left(\frac{n}{m}\right)^{m/(n-m)} \epsilon \left[\left(\frac{\sigma}{r}\right)^n - \left(\frac{\sigma}{r}\right)^m \right]$$

Edgar Grüneisen (1912)

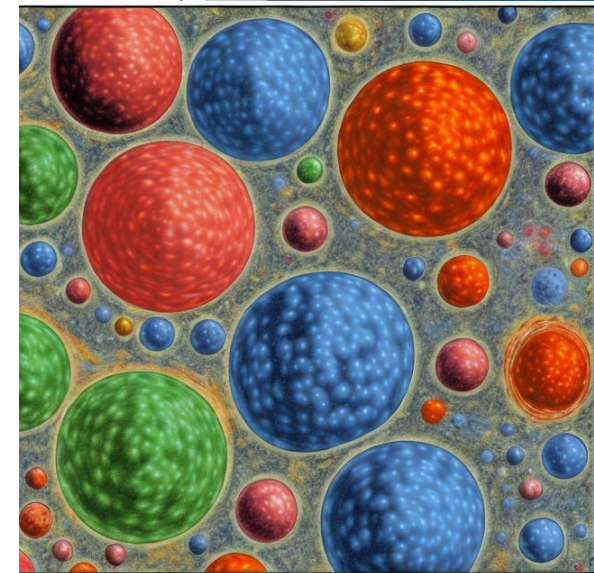
$$-\frac{a}{r^x} + \frac{b}{r^y}$$

Lennard-Jones (1924-1931)

$$f(r) = \frac{\lambda_n}{r^n} - \frac{\lambda_m}{r^m},$$

Fowler&Bernal (1933) $n=12$

$$V = -F(r) - C/r^6 + B/r^n,$$



199x-201x



Source: openart.ai prompt purgatory by hieronymus bosch

MACE

$$h_{i,k00}^{(0)} = \sum_z W_{kz} \delta_{zz_i}, \quad (1)$$

$$\bar{h}_{i,kl_2m_2}^{(s)} = \sum_{\tilde{k}} W_{k\tilde{k}l_2}^{(s)} h_{i,\tilde{k}l_2m_2}^{(s)}, \quad (2)$$

$$j_0^n(r_{ij}) = \sqrt{\frac{2}{r_{\text{cut}}}} \frac{\sin\left(n\pi \frac{r_{ij}}{r_{\text{cut}}}\right)}{r_{ij}} f_{\text{cut}}(r_{ij}), \quad (3)$$

$$R_{k\eta_1l_1l_2l_3}^{(s)}(r_{ij}) = \text{MLP}\left(\{j_0^n(r_{ij})\}_n\right), \quad (4)$$

$$\begin{aligned} \phi_{ij,k\eta_1l_3m_3}^{(s)} &= \sum_{l_1l_2m_1m_2} C_{\eta_1,l_1m_1l_2m_2}^{l_3m_3} R_{k\eta_1l_1l_2l_3}^{(s)}(r_{ij}) \\ &\quad \times Y_{l_1}^{m_1}(\hat{\mathbf{r}}_{ij}) \bar{h}_{j,kl_2m_2}^{(s)}, \end{aligned} \quad (5)$$

$$A_{i,kl_3m_3}^{(s)} = \sum_{\tilde{k},\eta_1} W_{k\tilde{k}\eta_1l_3}^{(s)} \sum_{j \in \mathcal{N}(i)} \phi_{ij,\tilde{k}\eta_1l_3m_3}^{(s)}, \quad (6)$$

$$\mathbf{A}_{i,klm}^{(s),\nu} = \prod_{\xi=1}^{\nu} A_{i,kl_{\xi}m_{\xi}}^{(s)}, \quad (7)$$

$$\mathbf{B}_{i,\eta_{\nu}kLM}^{(s),\nu} = \sum_{lm} \mathcal{C}_{\eta_{\nu}lm}^{LM} \mathbf{A}_{i,klm}^{(s),\nu}, \quad (8)$$

$$m_{i,kLM}^{(s)} = \sum_{\nu} \sum_{\eta_{\nu}} W_{z_i\eta_{\nu}kL}^{(s),\nu} \mathbf{B}_{i,\eta_{\nu}kLM}^{(s),\nu}, \quad (9)$$

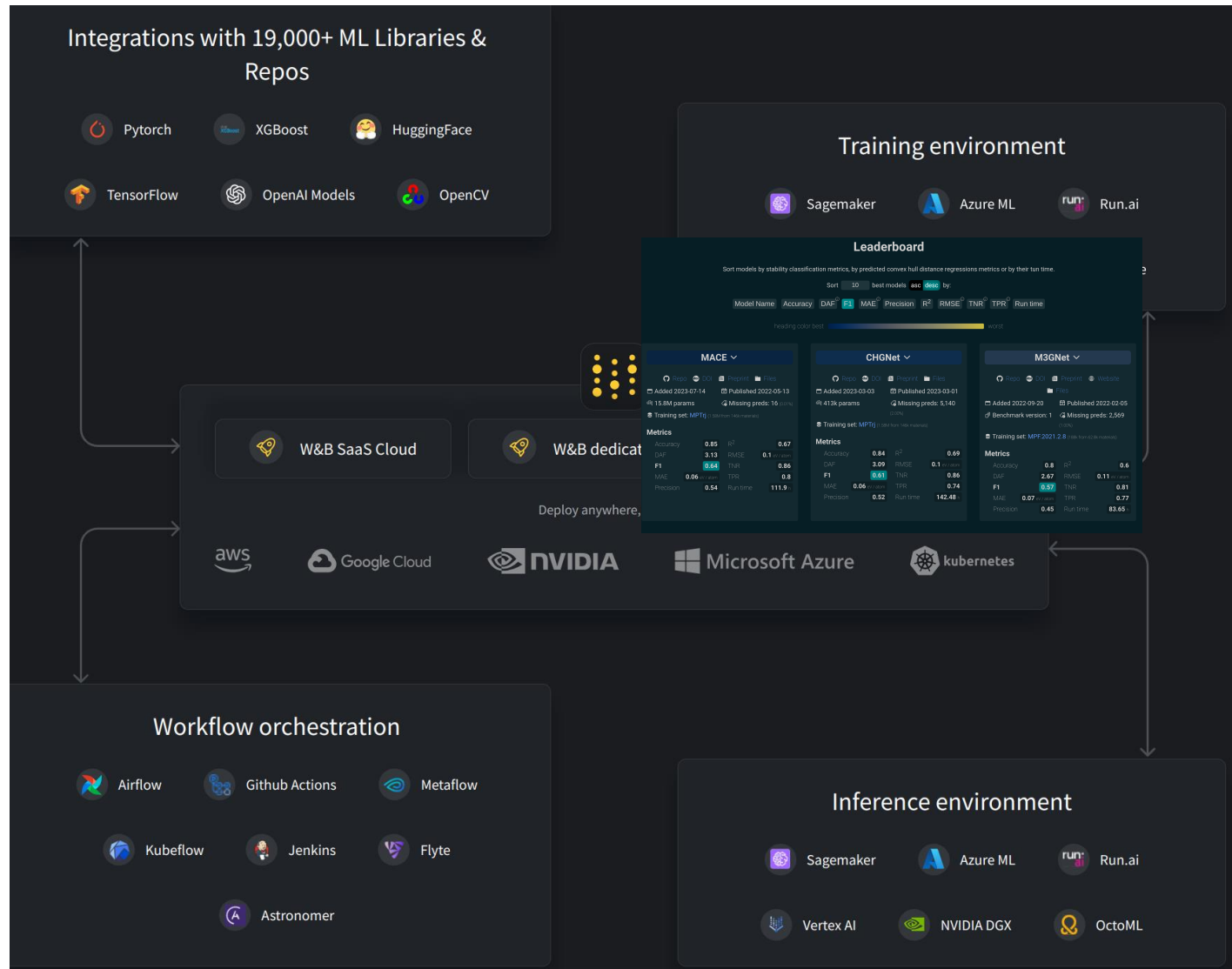
$$h_{i,kLM}^{(s+1)} = \sum_{\tilde{k}} W_{kL,\tilde{k}}^{(s)} m_{i,\tilde{k}LM}^{(s)} + \sum_{\tilde{k}} W_{kz_iL,\tilde{k}}^{(s)} h_{i,\tilde{k}LM}^{(s)}. \quad (10)$$

$$E_i = \sum_{s=1}^S E_i^{(s)} = \sum_{s=1}^S \mathcal{R}^{(s)}\left(\mathbf{h}_i^{(s)}\right),$$

$$\mathcal{R}^{(s)}\left(\mathbf{h}_i^{(s)}\right) = \begin{cases} \sum_k W_k^{(s)} h_{i,k00}^{(s)} & \text{if } s < S, \\ \text{MLP}\left(\{h_{i,k00}^{(s)}\}_k\right) & \text{if } s = S. \end{cases}$$

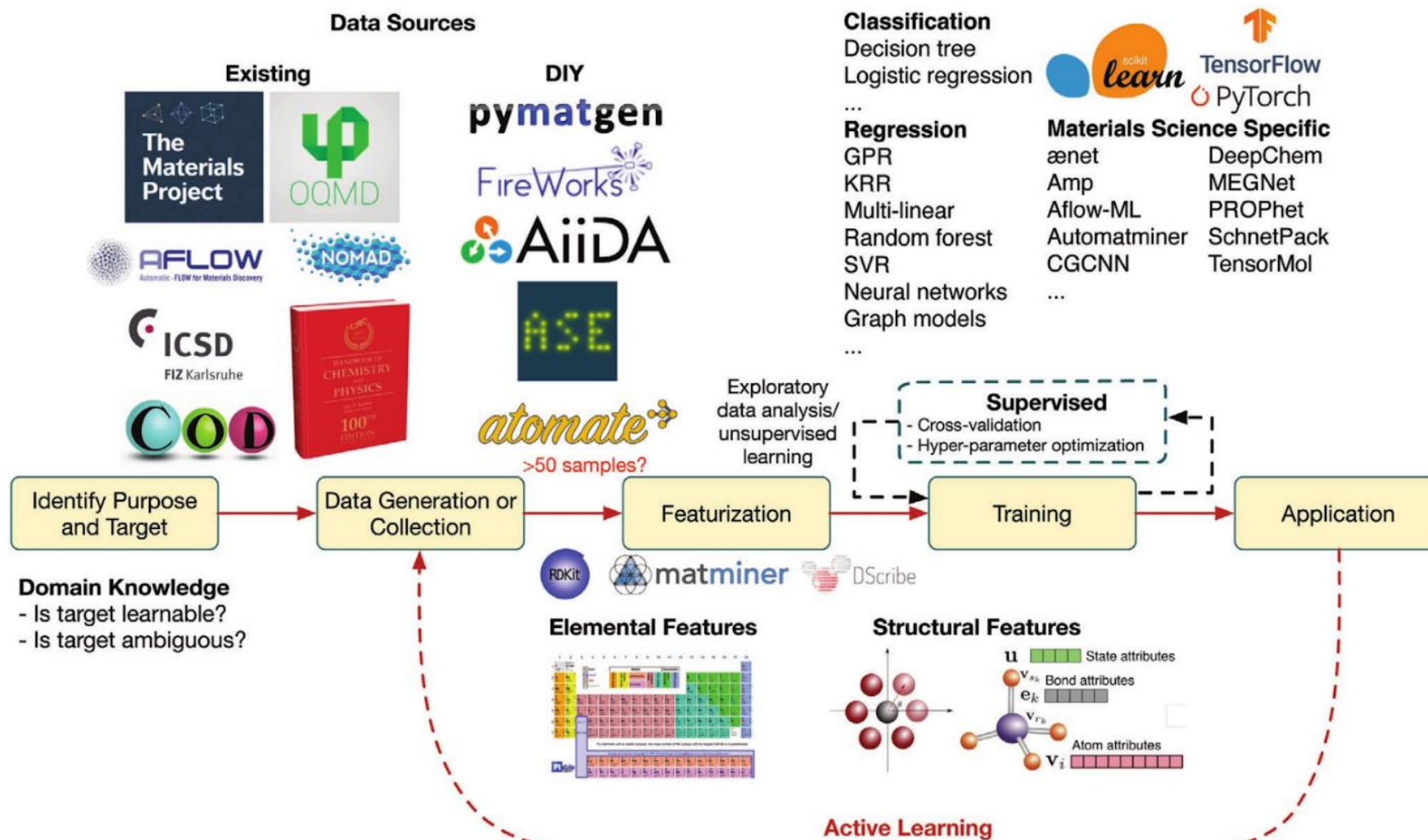
$$\mathbf{F} = -\nabla \sum_i E_i.$$

Workflows – AI community view

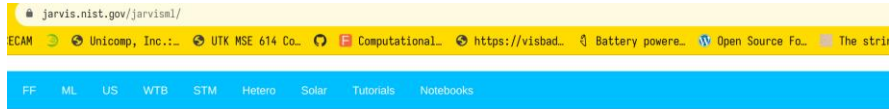


source: wandb

Workflows– science view



Current databases example



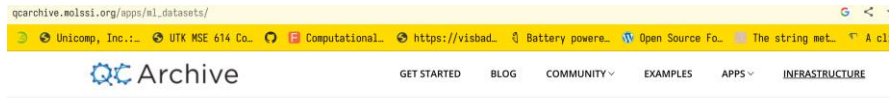
First app

This app predicts property of an input material given in POSCAR format using JARVIS-ML model for several pr

Submit

```

Mo1 Se2
1.0
1.661759 -2.878250 0.000000
1.661759 2.878250 0.000000
0.000000 0.000000 35.451423
Mo Se
1 2
direct
0.666667 0.333333 0.326886 Mo
0.333333 0.666667 0.374080 Se
    
```

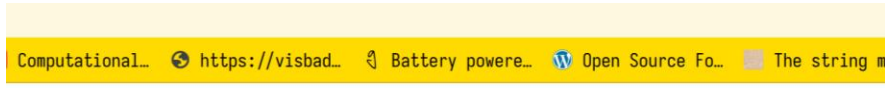
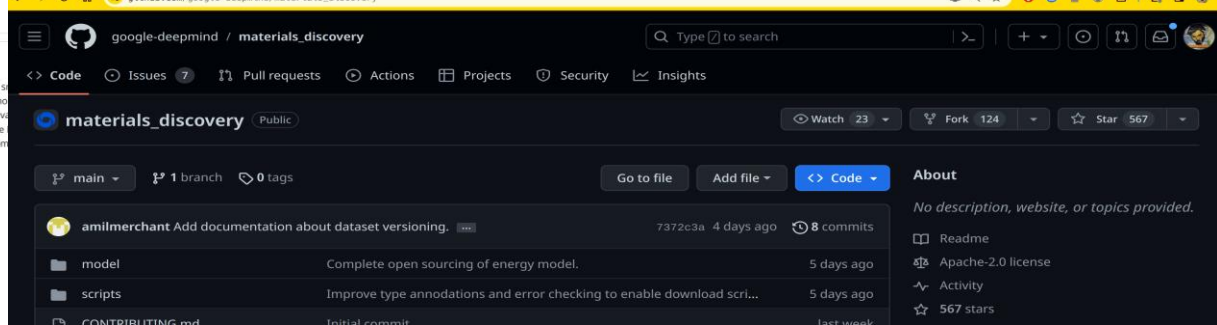


Search:

Add your Dataset License

Name	Quality	Data Points	Elements	Sampling	Download
GDML					

Description
Molecular dynamics trajectories of salicylic acid malonate dihydrate, ethanol, ethanol. Four levels of theory are available: CCSD(T)/cc-pVTZ. Trajectories at the wavefunction level contain 10³ geom



species. You can narrow the selection to models that support multiple species after you click.

							B	C	N
							Al	Si	P
Mn	Fe	Co	Ni	Cu	Zn	Ga	Ge	As	
Tc	Ru	Rh	Pd	Ag	Cd	In	Sn	Sb	
Re	Os	Ir	Pt	Au	Hg	Tl	Pb	Bi	
Bh	Hs	Mt	Ds	Rg	Cn	Nh	Fl	Mc	
Sm	Eu	Gd	Tb	Dy	Ho	Er	Tm		
Pu	Am	Cm	Bk	Cf	Es	Fm	Mn		

Accessibility

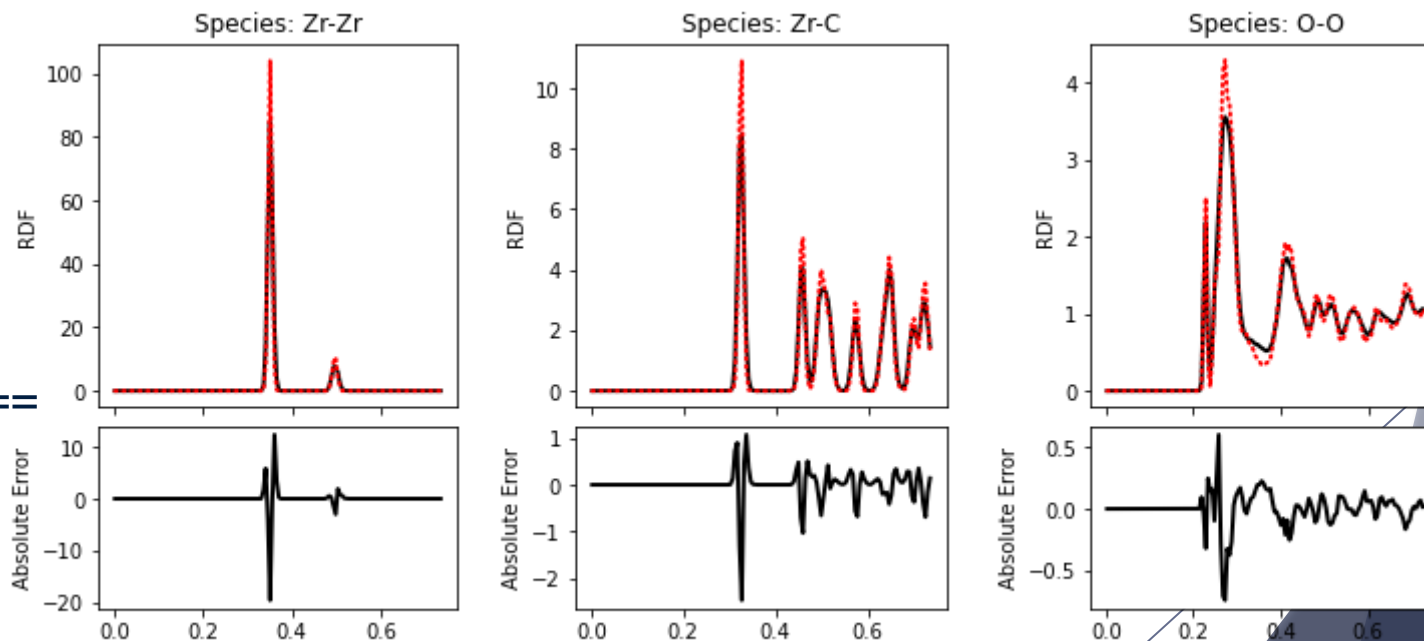
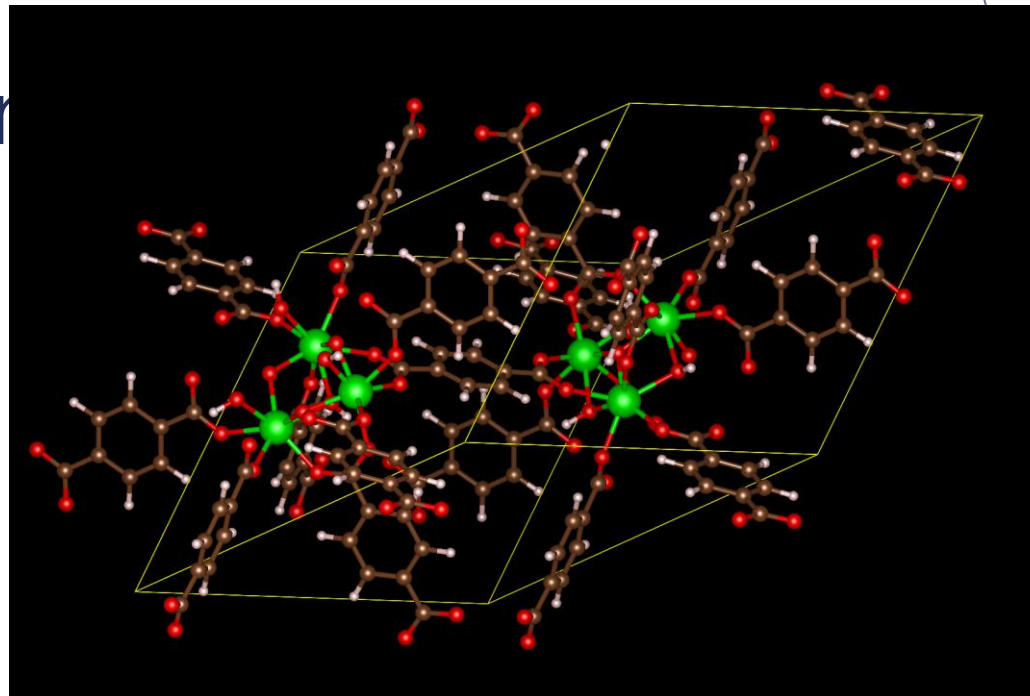
unique KIM potential is Interatomic potential directly with many r

UiO66 HDNNP potential

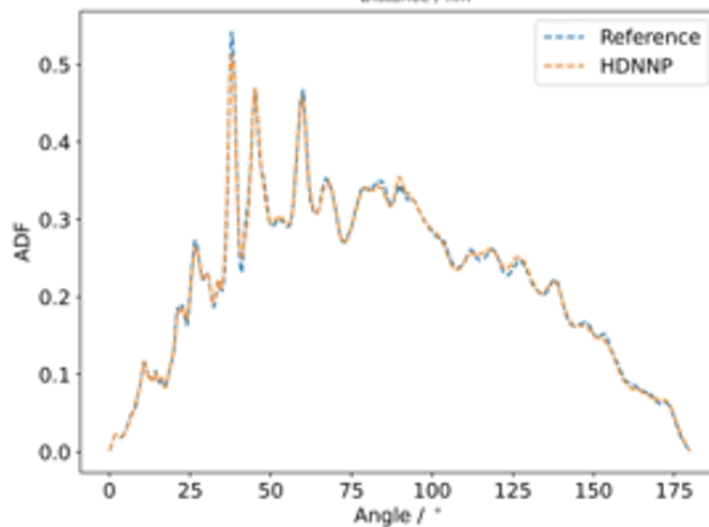
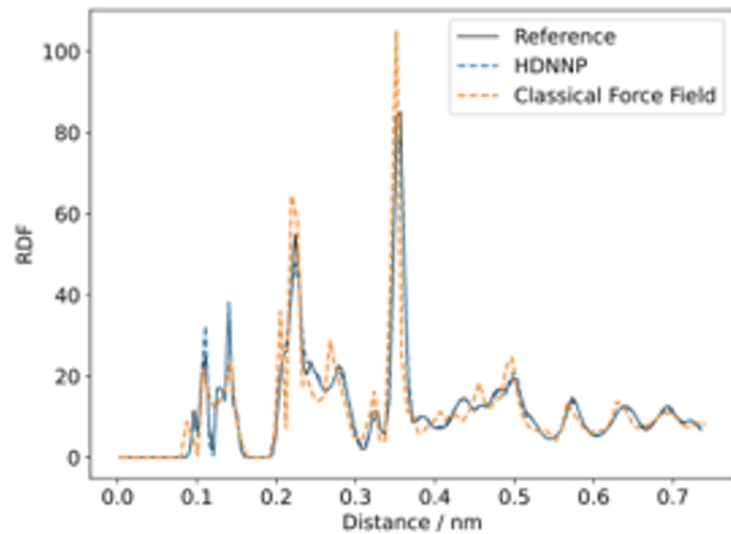
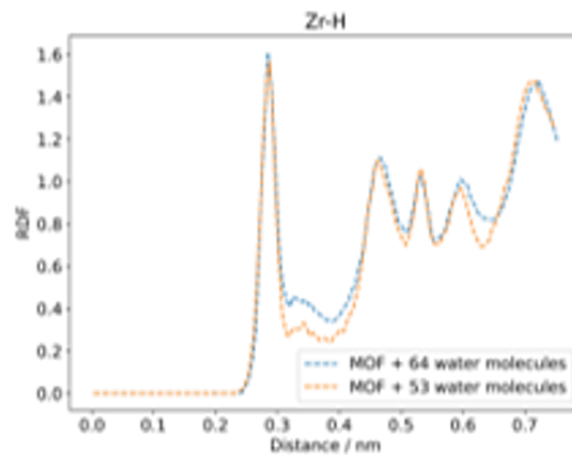
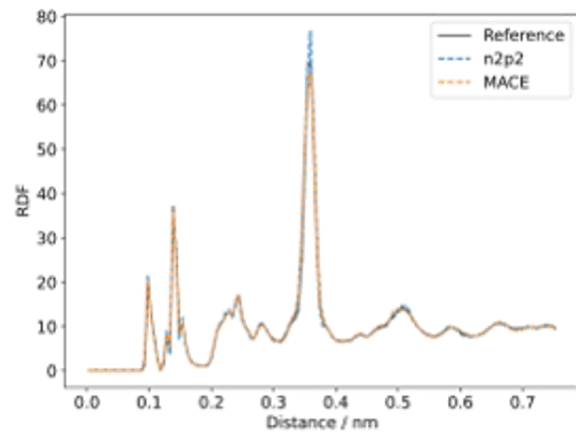
Score Summary

Label | Accuracy [%]

Zr-Zr	89.34
Zr-O	89.79
Zr-C	91.12
Zr-H	93.88
O-O	95.27
O-C	96.94
O-H	95.08
C-C	90.84
C-H	94.97
H-H	96.06
Mean	93.33

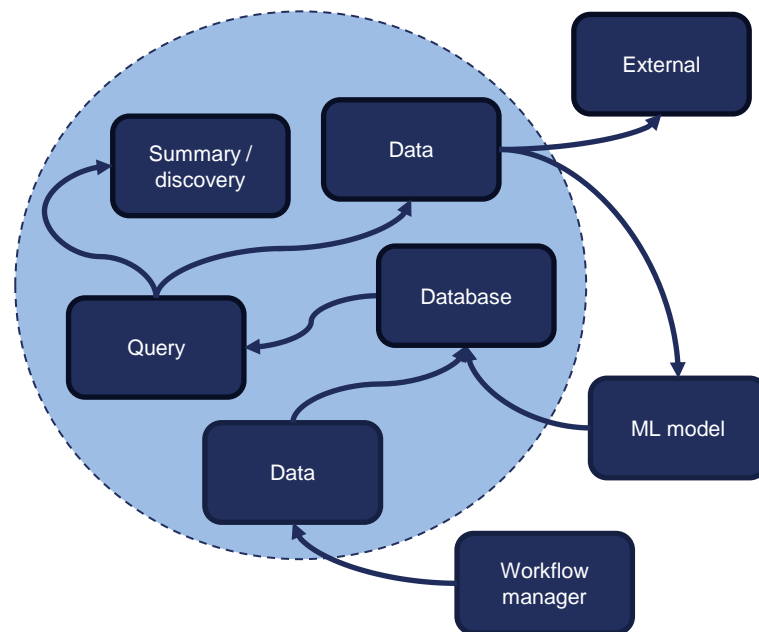


MACE



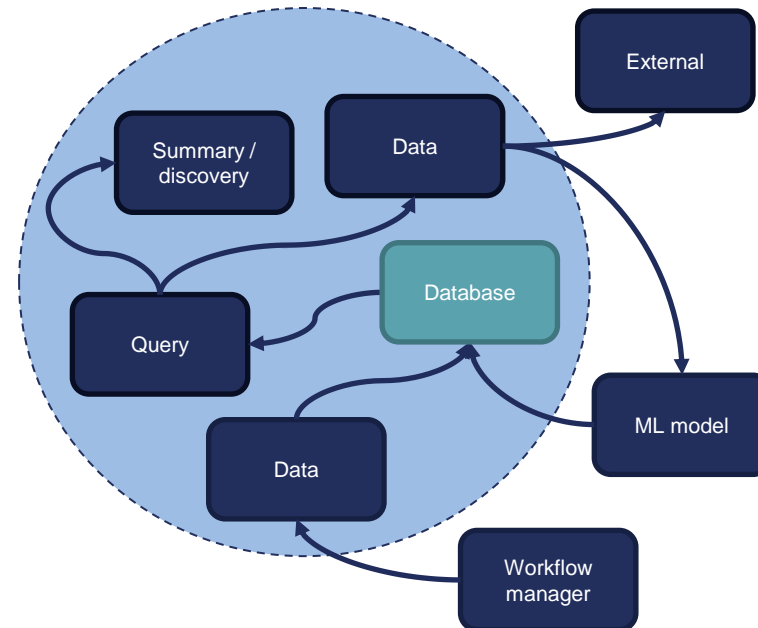
Database features

- ▶ Proof of concept database&app:
 - installable centrally within PSDI
 - local for user - blueprints
 - users can both interrogate, download and deposit data
 - reconfigurable database structure
 - API and web interface
- ▶ Advanced search features: opensearch techniques
- ▶ Integrate into ML workflows
- ▶ Reproduceable and Reusable
- ▶ Proof of concept in collaboration with Gábor Csányi@University of Cambridge



Database implementation

- ▶ JSON/dictionary data input
 - ▶ Atomistic data e.g. elements, positions, energies, forces
 - ▶ Metadata e.g. units, provenance
 - ▶ User-defined keys
- ▶ API access via Python client and CLI initially
- ▶ Complex, scalable queries via OpenSearch
- ▶ Data discovery due to unstructured data
 - ▶ List/frequency of keys
 - ▶ Visualisation
- ▶ Filtered data used to train ML models
 - ▶ Generate and store new structures and force fields in database



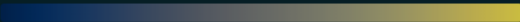
but

Leaderboard

Sort models by stability classification metrics, by predicted convex hull distance regressions metrics or by their tun time.

Sort best models asc desc by:

Model Name Accuracy DAF[⊕] **F1** MAE[⊕] Precision R² RMSE[⊕] TNR[⊕] TPR[⊕] Run time

heading color best  worst

MACE

Repo DOI Preprint Files

Added 2023-07-14 Published 2022-05-13

15.8M params Missing preds: 16 (0.01%)

Training set: [MPTrj](#) (1.58M from 146k materials)

Metrics

Accuracy	0.85	R ²	0.67
DAF	3.13	RMSE	0.1 eV / atom
F1	0.64	TNR	0.86
MAE	0.06 eV / atom	TPR	0.8
Precision	0.54	Run time	111.9 h

CHGNet

Repo DOI Preprint Files

Added 2023-03-03 Published 2023-03-01

413k params Missing preds: 5,140 (2.00%)

Training set: [MPTrj](#) (1.58M from 146k materials)

Metrics

Accuracy	0.84	R ²	0.69
DAF	3.09	RMSE	0.1 eV / atom
F1	0.61	TNR	0.86
MAE	0.06 eV / atom	TPR	0.74
Precision	0.52	Run time	142.48 h

M3GNet

Repo DOI Preprint Website

Files

Added 2022-09-20 Published 2022-02-05

Benchmark version: 1 Missing preds: 2,569 (1.00%)

Training set: [MPF.2021.2.8](#) (188k from 62.8k materials)

Metrics

Accuracy	0.8	R ²	0.6
DAF	2.67	RMSE	0.11 eV / atom
F1	0.57	TNR	0.81
MAE	0.07 eV / atom	TPR	0.77
Precision	0.45	Run time	83.65 h

<https://matbench-discovery.materialsproject.org/models>

Q&A

