Introduction to Data Science in Materials Science

Shyue Ping Ong

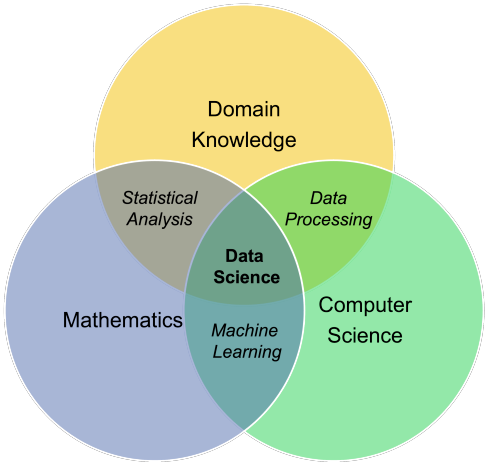
University of California, San Diego

Fall 2023

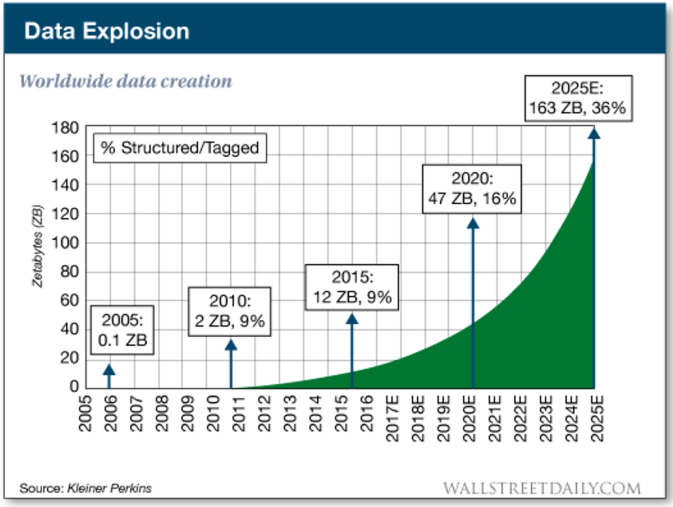
What is Data Science?

Data science is a multi-disciplinary field that uses scientific methods, processes, algorithms and systems to extract knowledge and insights from structured and unstructured data.

What is Data Science?



The Data Age


Data Explosion
Worldwide data creation
Zetabytes (ZB)
% Structured/Tagged
2005: 0.1 ZB
2010: 2 ZB, 9%
2015: 12 ZB, 9%
2020: 47 ZB, 16%
2025E: 163 ZB, 36%
0
20
40
60
80
100
120
140
160
180
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017E
2018E
2019E
2020E
2021E
2022E
2023E
2024E
2025E
Source: Kleiner Perkins
WALLSTREETDAILY.COM

Growth of Materials Data (as of Jan 1 2020)

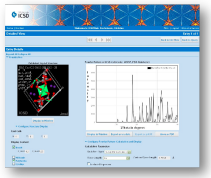


Figure: ICSD: ∼200,000 crystals

Figure: COD: ∼400,000 crystals

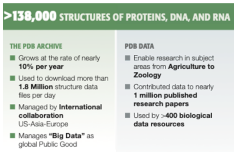


Figure: Protein data bank

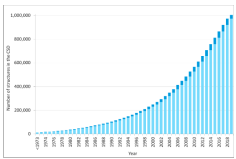


Figure: Cambridge structural database (small-molecule organic crystal structures)

But Quantity and Quality Lags Many Other Fields

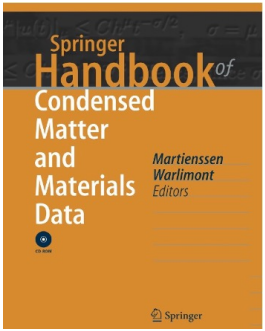


Figure: One of the most comprehensive handbooks on materials data: Density, thermal and electrical conductivity, melting and boiling points, etc., but O(100) binaries and limited ternaries...

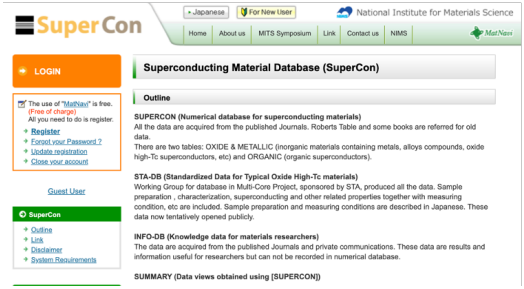
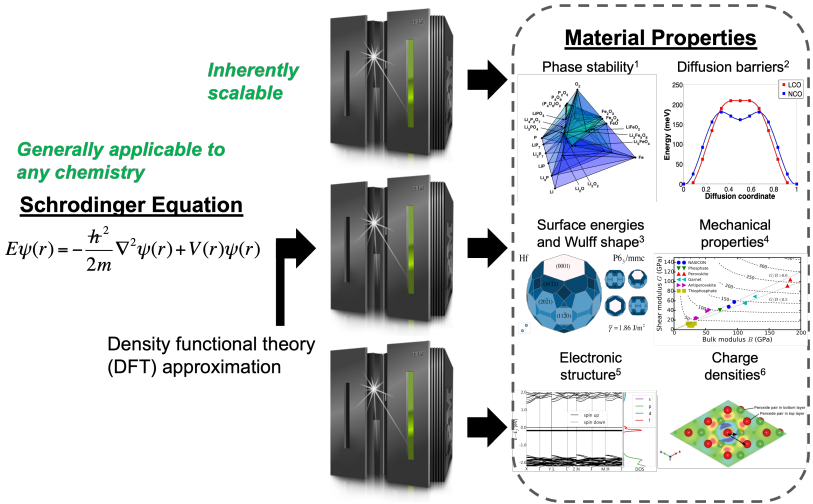
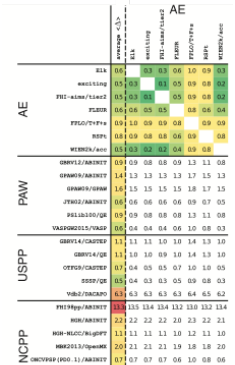


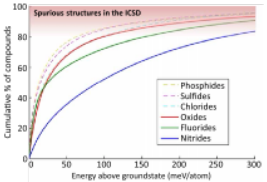
Figure: $\sim$1000+ superconductors (many minor composition modifications). Ref: https://supercon.nims.go.jp/

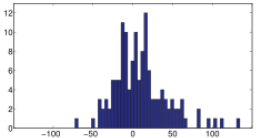
First Principles Materials Computations

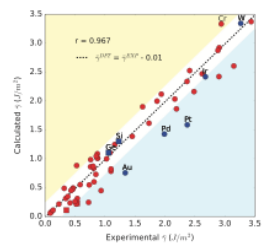


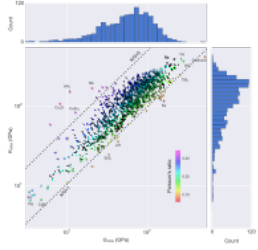
Electronic structure calculations are today *reliable* and *reasonably accurate*...









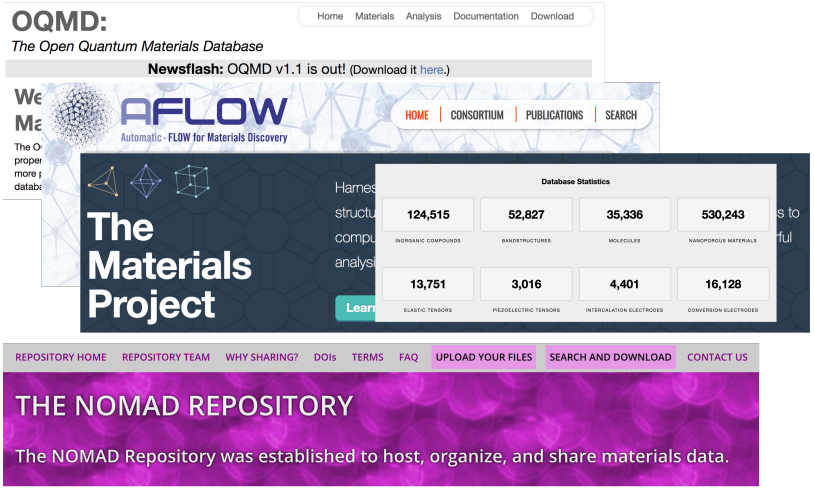


- (left) Modern electronic structure codes give relatively consistent equations of state.
- (right, clockwise from top left) Good predictions can be obtained for phase stability,[1] formation energies, surface energies,[2] and elastic constants[3].

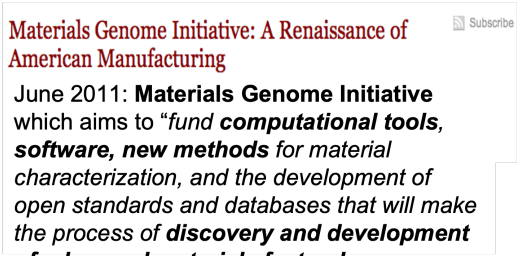
Software frameworks for high-throughput computational materials science

- Materials Project (https://materialsproject.org)[4]
  - Python Materials Genomics or pymatgen (https://pymatgen.org)[5]
  - Custodian (https://materialsproject.github.io/custodian/)
  - FireWorks [6]
- Atomic Simulation Environment (https://wiki.fysik.dtu.dk/ase)
- AFLOW (http://aflowlib.org)[7]
- AiiDa (http://www.aiida.net)

Computation + Automation → Large databases

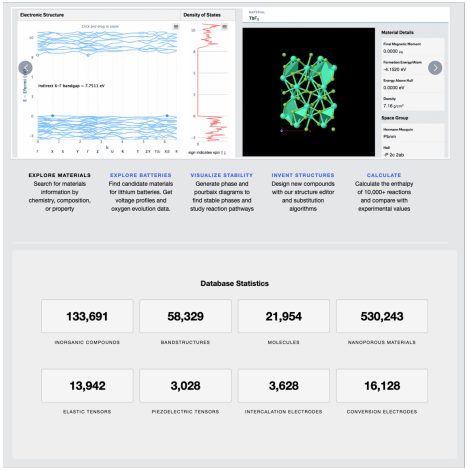


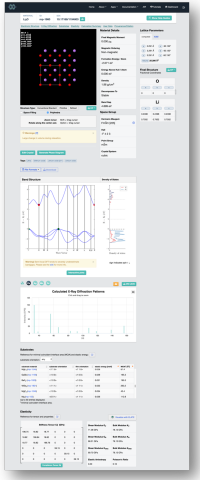
Google for Materials



The Materials Project is an open science project to make the computed properties of all known inorganic materials publicly available to all researchers to accelerate materials innovation.

Google for Materials





Materials Application Programming Interface (API)[8]

- An open platform for accessing Materials Project data based on REpresentational State Transfer (REST) principles.
- *Flexible and scalable* to cater to large number of users, with different access privileges.
- Simple to use and code agnostic.
- Requires an API key, available at: https://www.materialsproject.org/dashboard
- Documentation: https://api.materialsproject.org/docs

RESTful API

A REST API maps a URL to a resource.

Example

GET https://api.dropbox.com/1/account/info

Returns information about a user's account.
Methods: GET, POST, PUT, DELETE, etc.
Response: Usually JSON or XML or both

Materials API Example

URL

https://api.materialsproject.org/summary/?formula=Fe2O3&_fields=formation_energy_per_atom

Example response:


```
{
    "data":[
        {"_id":"61a2dcaa2c86325a0218b5ef","formation_energy_per_atom":-1.6299189062500006},
        {"_id":"61a2dcb52c86325a021af9bd","formation_energy_per_atom":-1.4175868379999996},
        ...
    ],
    "meta":{
        "api_version":"0.48.0",
        "time_stamp":"2022-09-19T13:17:11.321756",
        "total_doc":26,
        "max_limit":1000,
        "default_fields":["material_id"]
    }
}
```


- Intuitive response format.
- Machine-readable (JSON parsers available for most programming languages).
- Metadata provides provenance for tracking.

Types of Materials Data

Qualitative data

- Nominal measurement.
- E.g., Metal/Insulator, Stable/Unstable.
- No rank or order.

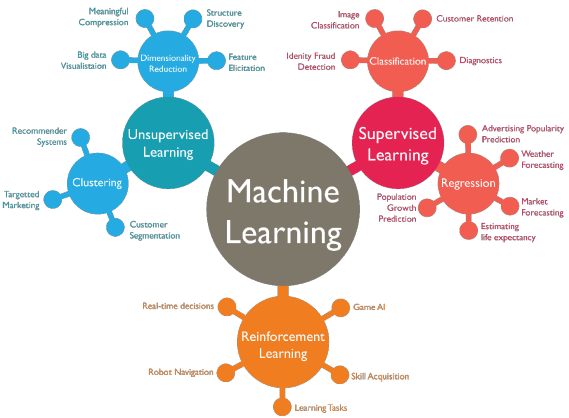
Ranked data

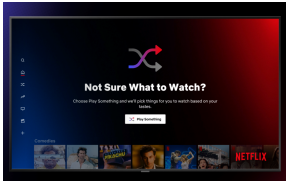
- Ordinal measurement (ordered).
- E.g., Insulator/ semiconductor/ conductor.
- Does not indicate distance between ranks.

Quantitative Data

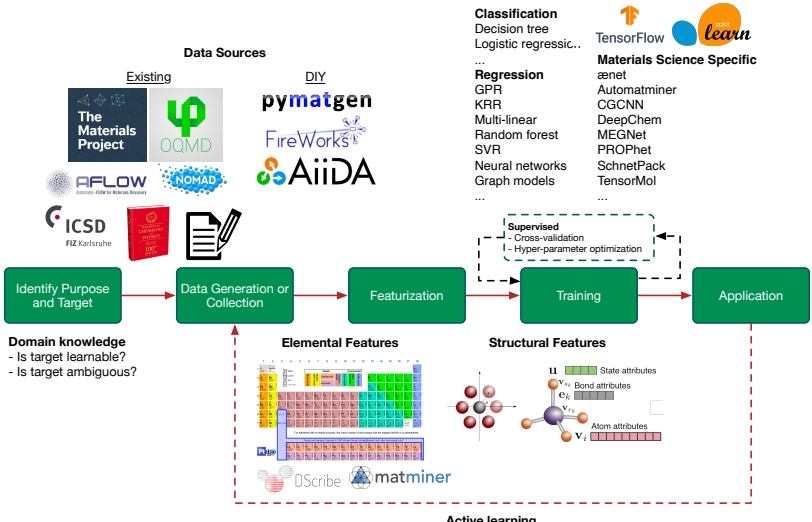
- Interval/ratio measurement (equal intervals and true 0).
- E.g., melting point, elastic constant, electrical/ionic conductivity.
- Considerable information and permits meaningful arithmetic operations.

What is Machine Learning?



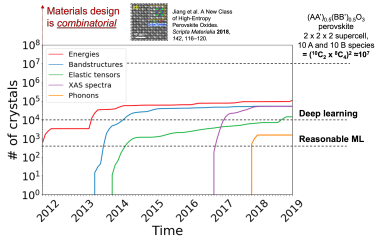


Materials ML Workflow

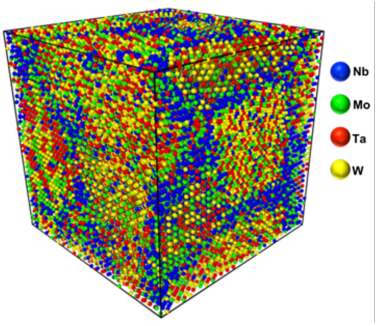


Where is ML valuable in Materials Science?

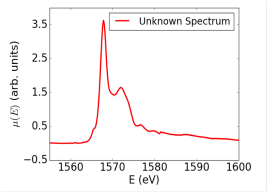
Too many to compute

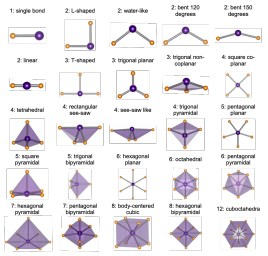


Too big to compute

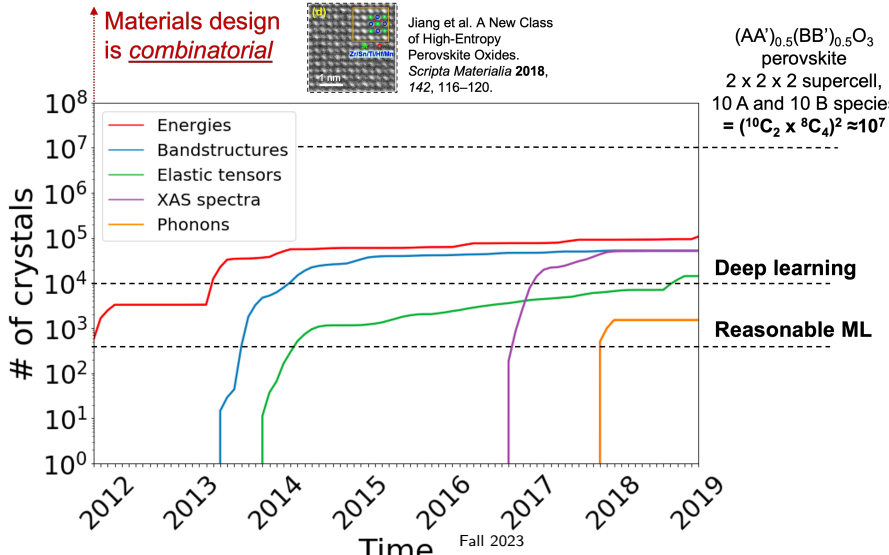


Too complex to understand.





Data History of the Materials Project



Surrogate models for "instant" property predictions

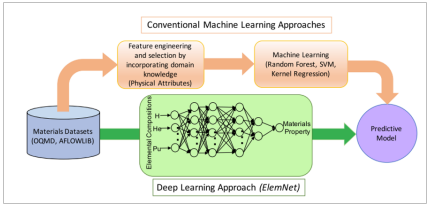
$$\text{Property} = f(\text{Composition}, \text{Structure})$$

- In ML terms, the material property, e.g., energetic (formation, energy above hull, reaction, binding, etc.), electronic (band gaps, DOS), mechanical, functional (e.g., ionic conductivity) is called the **"target"**.
- Composition and Structure are called the **"descriptors"** or **"features"**.
- Examples of compositonal features: stoichiometric attributes, e.g., number and ratio of elements, etc. elemental properties, e.g., mean, range, min, max of atomic number, electronegativity, row, group, atomic radii, etc., electronic structure, e.g., number of valence electrons, shells, etc.
- Examples of structural features: crystal/molecular symmetry, lattice parameters, atomic coordinates, connectivity / bonding between atoms.

Compositional features

- **Average atomic mass**: Composition-weighted average of the atomic masses of the elements in the compound. *Value for FeO: 0.5 x 55.845 + 0.5 x 15.999 = 35.92.*
- **Average column on periodic table**: Composition-weighted average of the columns of the elements in the compound. *Value for FeO: 0.5 x 8 + 0.5 x 16 = 12.0.*
- **Average row on the periodic table**: Composition-weighted average of the rows of the elements in the compound. *Value for FeO: 0.5 x 4 + 0.5 x 2 = 3.0.*
- **Maximum difference in atomic number**: Largest atomic number in the composition less the smallest. *Value for FeO: 26 – 8 = 18.*
- **Average atomic number**: Composition-weighted average of the atomic numbers of the elements in the compound. *Value for FeO: 0.5 x 26 + 0.5 x 8 = 17.0.*
- **Maximum difference in atomic radii**: Largest atomic radius in the composition less the smallest (in pm). *Value for FeO: 140 – 60 = 80.*
- **Average atomic radius**: Composition-weighted average of the atomic radii of the elements in the compound. *Value for FeO: 0.5 x 140 + 0.5 x 60 = 100.0.*
- **Maximum difference in electronegativity**: Largest electronegativity in the composition less the smallest. *Value for FeO: 3.44 – 1.83 = 1.61.*
- **Average electronegativity**: Composition-weighted average of the electronegativities of the elements in the compound. *Value for FeO: 0.5 x 3.44 + 0.5 x 1.83 = 2.635.*
- **Average number of *s* valence electrons**: Composition-weighted average of the number of *s* valence electrons associated with the elements in the compound. *Value for FeO: 0.5 x 4 + 0.5 x 2 = 3.0.*
- **Average number of *p* valence electrons**: Analogous to above, but for *p* electrons. *Value for FeO: 0.5 x 0 + 0.5 x 4 = 2.0.*
- **Average number of *d* valence electrons**: Analogous to above, but for *d* electrons. *Value for FeO: 0.5 x 6 + 0.5 x 0 = 3.0.*
- **Average number of *f* valence electrons**: Analogous to above, but for *f* electrons. *Value for FeO: 0.5 x 0 + 0.5 x 0 = 0.0.*
- ***s* fraction of valence electrons**: Composition-weighted fraction of all valence electrons in the compound that represent *s* states. *Value for FeO: 3.0 / (3.0 + 2.0 + 3.0 + 0.0) = 0.375.*
- ***p* fraction of valence electrons**: Analogous to above, but for *p* electrons. *Value for FeO: 2.0 / (3.0 + 2.0 + 3.0 + 0.0) = 0.25.*
- ***d* fraction of valence electrons**: Analogous to above, but for *d* electrons. *Value for FeO: 3.0 / (3.0 + 2.0 + 3.0 + 0.0) = 0.375.*
- ***f* fraction of valence electrons**: Analogous to above, but for *f* electrons. *Value for FeO: 0.0 / (3.0 + 2.0 + 3.0 + 0.0) = 0.0.*

Figure: Meredig et al. (2014) Phys. Rev. B89, 094104



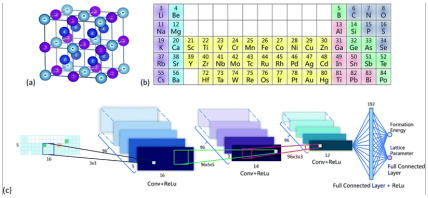


Figure: Jha et al. (2018) Sci. Rep., 8(1), 17593., Zheng, X., et al (2018). Chem. Sci., 9(44), 8426-8432.

Structural features

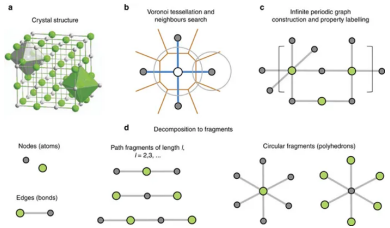


Figure: Property-labelled materials fragments + gradient boosting decision tree.[9]

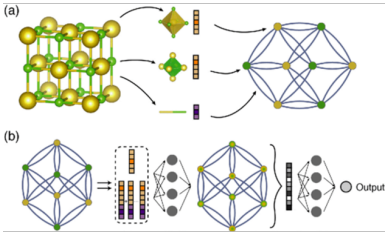


Figure: Crystal graph + graph convolutional neural networks

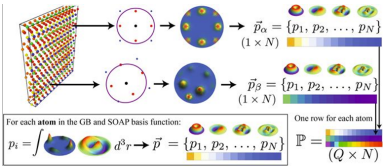


Figure: Smooth overlap of atom positions (SOAP).[10]

Example: Graph-based representations

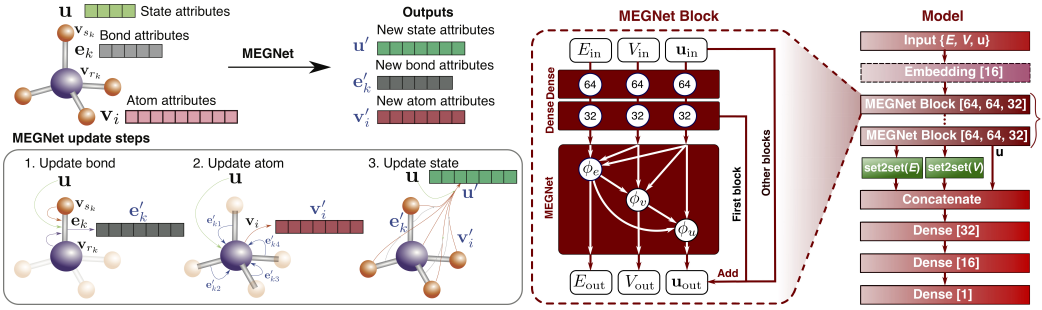


Figure: MatErials Graph Networks (MEGNet).[11]

MEGNet Performance Benchmarks

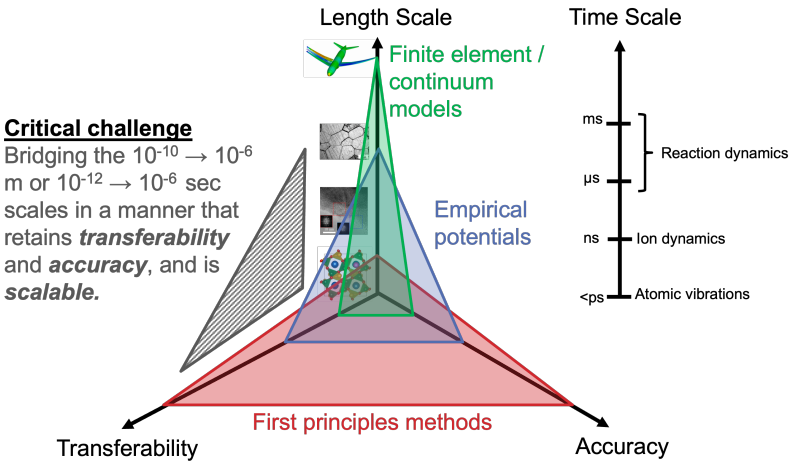
	MEGNET	MEGNET-Simple	SchNet	"Chemical Accuracy"
U0 (meV)	9	12	14	43
G (meV)	10	12	14	43
ε_{HOMO} (eV)	0.038	0.043	0.041	0.043
ε_{LUMO} (eV)	0.031	0.044	0.034	0.043
C_v (cal/molK)	0.030	0.029	0.033	0.05

Table: 130,462 QM9 molecules

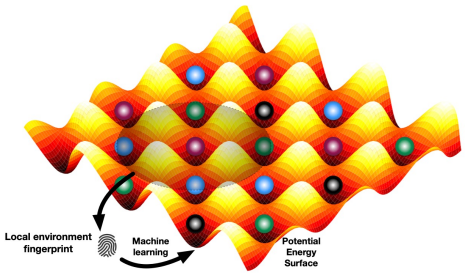
Property	MEGNet	SchNet	CGCNN
Formation energy E_f (meV/atom)	28 (60,000)	35	39 (28,046)
Band gap E_g (eV)	0.330 (36,720)	-	0.388 (16,485)
$\log_{10} K_{VRH}$ (GPa)	0.050 (4,664)	-	0.054 (2,041)
$\log_{10} G_{VRH}$ (GPa)	0.079 (4,664)	-	0.087 (2,041)
Metal classifier	78.9% (55,391)	-	80% (28,046)
Non-metal classifier	90.6% (55,391)	-	95% (28,046)

Table: Materials Project Crystals. Brackets indicate number of data points.

Scale Challenge in Materials Science



ML as a solution to the scale challenge



- Examples: Neural Network Potential (NNP)[12], Gaussian Approximation Potential (GAP)[13], moment tensor potential (MTP)[14], spectral neighbor analysis potential,[15], atomic cluster expansion[16], etc.
- ML models: Linear regression, Gaussian kernels, neural networks, etc.
- Local environment descriptors:

$$G_i^{\mathrm{atom,rad}} = \sum_{j \neq i}^{N_{\mathrm{atom}}} e^{-\eta(R_{ij}-R_s)^2} \cdot f_c(R_{ij}),$$

$$G_i^{\mathrm{atom,ang}} = 2^{1-\zeta} \sum_{j,k \neq i}^{N_{\mathrm{atom}}} (1+\lambda\cos\theta_{ijk})^{\zeta} \cdot e^{-\eta'(R_{ij}^2+R_{ik}^2+R_{jk}^2)} \cdot f_c(R_{ij}) \cdot f_c(R_{ik})$$

$$\rho_i(\boldsymbol{R}) = \sum_j f_c(R_{ij}) \cdot \exp(-\frac{|\boldsymbol{R}-\boldsymbol{R}_{ij}|^2}{2\sigma_{\mathrm{atom}}^2}),$$

Automatable workflows for ML Interatomic Potential Construction

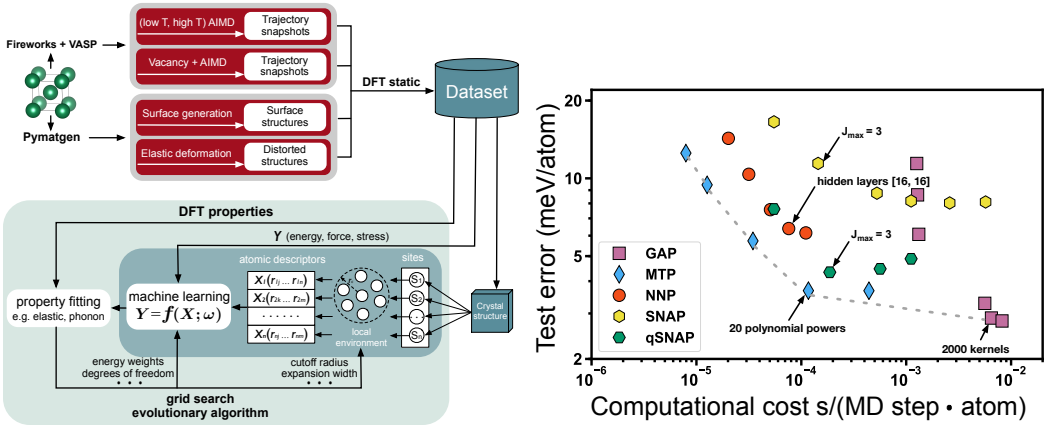


Figure: Automatic workflow for ML-IAP construction and performance benchmarks.[17]

Example: Ni-Mo

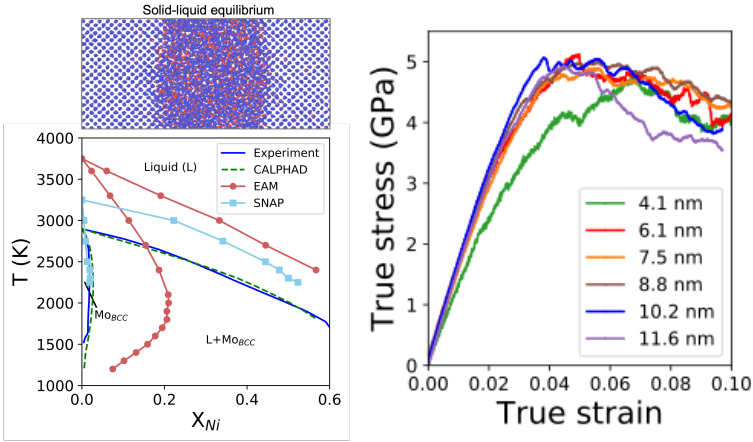
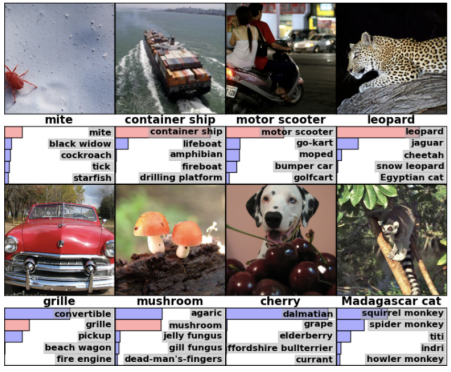


Figure: ML-IAP results on Ni-Mo. (left) Ni-Mo phase diagram. (right) Stress-strain curves as a function of grain size[17]

Modelling complex relationships



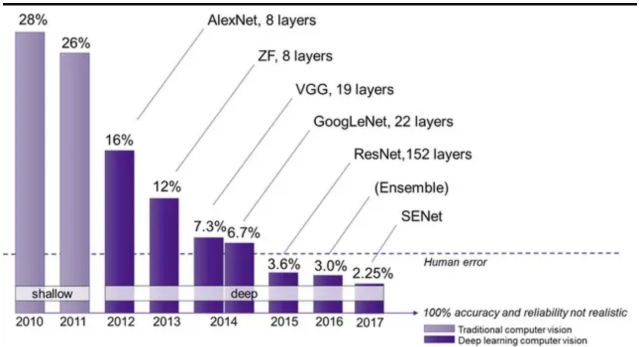


Figure: ImageNet (https://www.image-net.org/)

Example: Coordination environment from X-ray Absorption Spectra

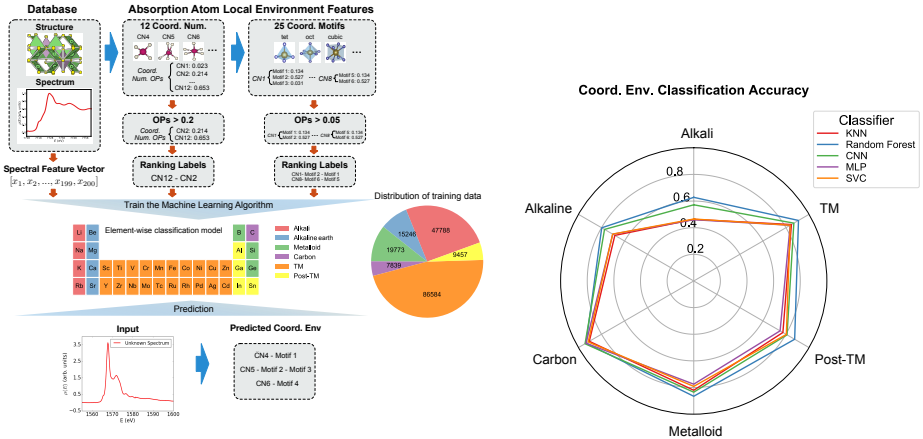


Figure: Random Forest Coordination Environment Classification[18]

Other examples

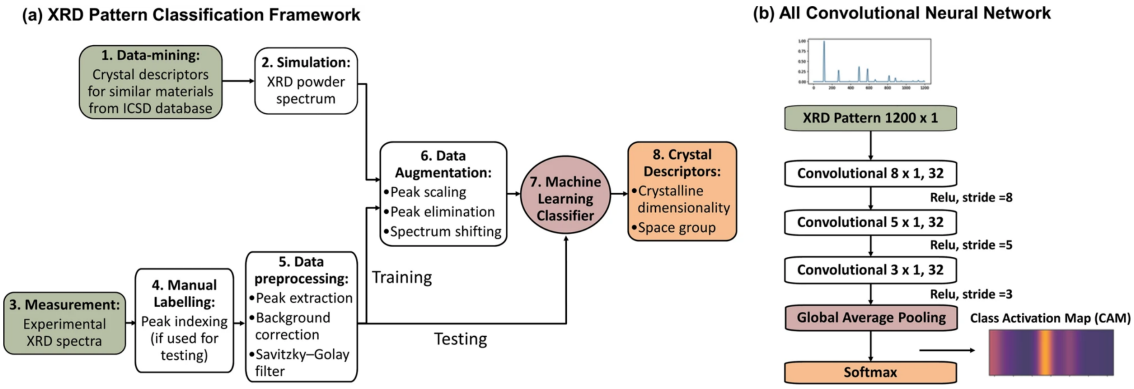


Figure: X-ray diffraction data classification with CNNs[19]

Bibliography I

Wenhao Sun, Stephen T Dacek, Shyue Ping Ong, Geoffroy Hautier, Anubhav Jain, William D Richards, Anthony C Gamst, Kristin A Persson, and Gerbrand Ceder.
The thermodynamic scale of inorganic crystalline metastability.
Science Advances, 2(11):e1600225–e1600225, November 2016.

Richard Tran, Zihan Xu, Balachandran Radhakrishnan, Donald Winston, Wenhao Sun, Kristin A. Persson, and Shyue Ping Ong.
Surface energies of elemental crystals.
Scientific Data, 3:160080, September 2016.

Maarten de Jong, Wei Chen, Thomas Angsten, Anubhav Jain, Randy Notestine, Anthony Gamst, Marcel Sluiter, Chaitanya Krishna Ande, Sybrand van der Zwaag, Jose J Plata, Cormac Toher, Stefano Curtarolo, Gerbrand Ceder, Kristin A Persson, and Mark Asta.
Charting the complete elastic properties of inorganic crystalline compounds.
Scientific Data, 2:150009, March 2015.

Bibliography II

Anubhav Jain, Shyue Ping Ong, Geoffroy Hautier, Wei Chen, William Davidson Richards, Stephen Dacek, Shreyas Cholia, Dan Gunter, David Skinner, Gerbrand Ceder, and Kristin A. Persson.
Commentary: The Materials Project: A materials genome approach to accelerating materials innovation.
APL Materials, 1(1):011002, 2013.

Shyue Ping Ong, William Davidson Richards, Anubhav Jain, Geoffroy Hautier, Michael Kocher, Shreyas Cholia, Dan Gunter, Vincent L. Chevrier, Kristin A. Persson, and Gerbrand Ceder.
Python Materials Genomics (pymatgen): A robust, open-source python library for materials analysis.
Computational Materials Science, 68:314–319, February 2013.

Bibliography III


Anubhav Jain, Shyue Ping Ong, Wei Chen, Bharat Medasani, Xiaohui Qu, Michael Kocher, Miriam Brafman, Guido Petretto, Gian-Marco Rignanese, Geoffroy Hautier, Daniel Gunter, and Kristin A. Persson.
FireWorks: A dynamic workflow system designed for high-throughput applications.
Concurrency and Computation: Practice and Experience, 27(17):5037–5059, December 2015.


Stefano Curtarolo, Wahyu Setyawan, Shidong Wang, Junkai Xue, Kesong Yang, Richard H. Taylor, Lance J. Nelson, Gus L.W. Hart, Stefano Sanvito, Marco Buongiorno-Nardelli, Natalio Mingo, and Ohad Levy.
AFLOWLIB.ORG: A distributed materials properties repository from high-throughput ab initio calculations.
Computational Materials Science, 58:227–235, June 2012.

Bibliography IV

Shyue Ping Ong, Shreyas Cholia, Anubhav Jain, Miriam Brafman, Dan Gunter, Gerbrand Ceder, and Kristin a. Persson.
The Materials Application Programming Interface (API): A simple, flexible and efficient API for materials data based on REpresentational State Transfer (REST) principles.
Computational Materials Science, 97:209–215, February 2015.

Olexandr Isayev, Corey Oses, Cormac Toher, Eric Gossett, Stefano Curtarolo, and Alexander Tropsha.
Universal Fragment Descriptors for Predicting Electronic Properties of Inorganic Crystals.
Nature Communications, 8:1–12, 2016.

Conrad W Rosenbrock, Eric R Homer, Gábor Csányi, and Gus L W Hart.
Discovering the building blocks of atomic systems using machine learning: Application to grain boundaries.
npj Computational Materials, 3(1):29, December 2017.

Bibliography V


Chi Chen, Weike Ye, Yunxing Zuo, Chen Zheng, and Shyue Ping Ong.
Graph Networks as a Universal Machine Learning Framework for Molecules and Crystals.
Chemistry of Materials, 31(9):3564–3572, May 2019.


Jörg Behler.
High-Dimensional Neural Network Potentials for Complex Systems.
Angewandte Chemie International Edition, pages n/a–n/a.


Albert P Bartók, Mike C Payne, Risi Kondor, and Gábor Csányi.
Gaussian approximation potentials: The accuracy of quantum mechanics, without the electrons.
Physical Review Letters, 104:136403, 2010.

Bibliography VI


Ivan S Novikov, Konstantin Gubaev, Evgeny V Podryabinkin, and Alexander V Shapeev.
The MLIP package: Moment tensor potentials with MPI and active learning.
Machine Learning: Science and Technology, 2(2):025002, January 2021.


A.P. Thompson, L.P. Swiler, C.R. Trott, S.M. Foiles, and G.J. Tucker.
Spectral neighbor analysis method for automated generation of quantum-accurate interatomic potentials.
Journal of Computational Physics, 285:316–330, March 2015.


Ralf Drautz.
Atomic cluster expansion of scalar, vectorial, and tensorial properties including magnetism and charge transfer.
Physical Review B, 102(2):024104, July 2020.

Bibliography VII

Yunxing Zuo, Chi Chen, Xiangguo Li, Zhi Deng, Yiming Chen, Jörg Behler, Gábor Csányi, Alexander V. Shapeev, Aidan P. Thompson, Mitchell A. Wood, and Shyue Ping Ong.
Performance and Cost Assessment of Machine Learning Interatomic Potentials.
The Journal of Physical Chemistry A, 124(4):731–745, January 2020.

Chen Zheng, Chi Chen, Yiming Chen, and Shyue Ping Ong.
Random Forest Models for Accurate Identification of Coordination Environments from X-Ray Absorption Near-Edge Structure.
Patterns, page 100013, April 2020.

Bibliography VIII


Felipe Oviedo, Zekun Ren, Shijing Sun, Charles Settens, Zhe Liu, Noor Titan Putri Hartono, Savitha Ramasamy, Brian L. DeCost, Siyu I. P. Tian, Giuseppe Romano, Aaron Gilad Kusne, and Tonio Buonassisi.
Fast and interpretable classification of small X-ray diffraction datasets using data augmentation and deep neural networks.
npj Computational Materials, 5(1):60, December 2019.

The End