IMPERIAL
Department of Materials

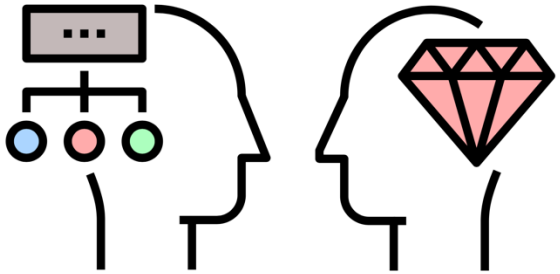
THE THOMAS YOUNG CENTRE TyC

Machine Learning for Materials

4. Crystal Representations

Aron Walsh

Department of Materials
Centre for Processable Electronics



Module Contents

1. Introduction
2. Machine Learning Basics
3. Materials Data
4. **Crystal Representations**
5. Classical Learning
6. Artificial Neural Networks
7. Building a Model from Scratch
8. Accelerated Discovery
9. Generative Artificial Intelligence
10. Recent Advances

Class Outline

Crystal Representations

A. Compositional

B. Structural

C. Graphs

Representation of Materials

Model performance depends on the choice of compositional and structural features

Minimal representation

Ab initio quantum mechanics (QM)

Input:
Atomic number, Z
Coordinates, **R**

$$\hat{\mathbf{H}}|\Psi\rangle = \mathrm{E}|\Psi\rangle$$

electronic wavefunction

Output:
Properties

Effective representation

Supervised machine learning (ML)

Input:
Feature vector, $\mathbf{X}$

$$y = f(\mathbf{X}, \boldsymbol{\Theta})$$

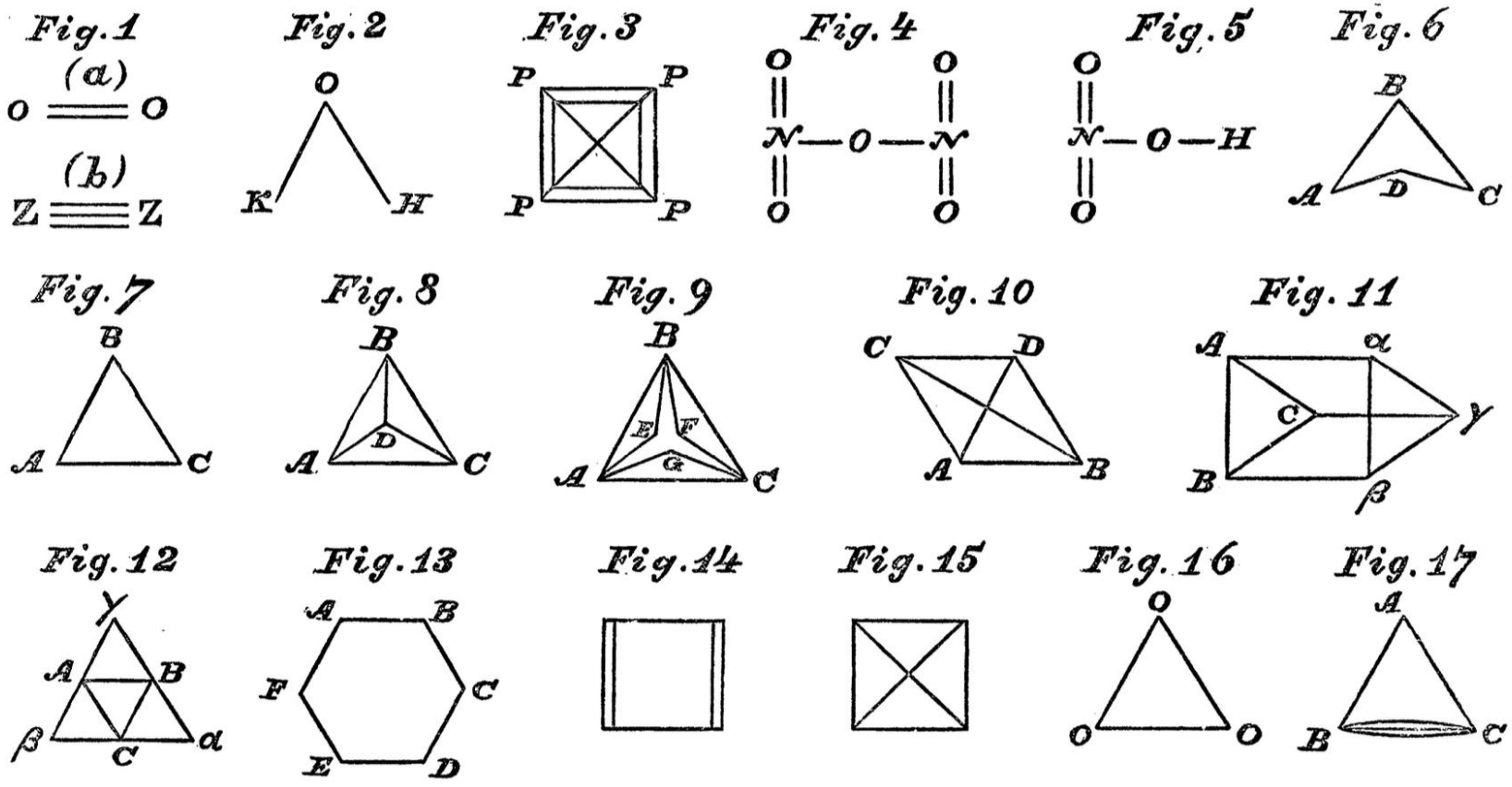
learned weights

Output:
Properties

K. T. Butler et al, Nature 559, 547 (2018)

How to Best Represent a Molecule?

Networks of atoms (nodes) connected by bonds (edges)

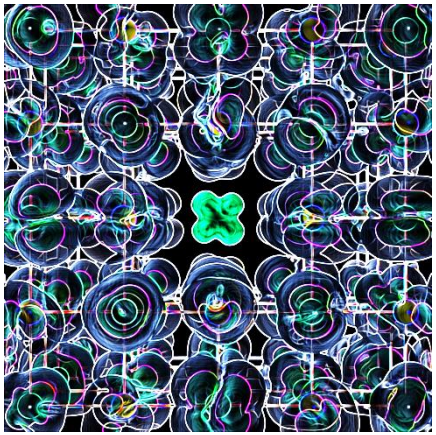


J. J. Sylvester, Am. J. Math. 1, 64 (1878)

How to Best Represent a Material?

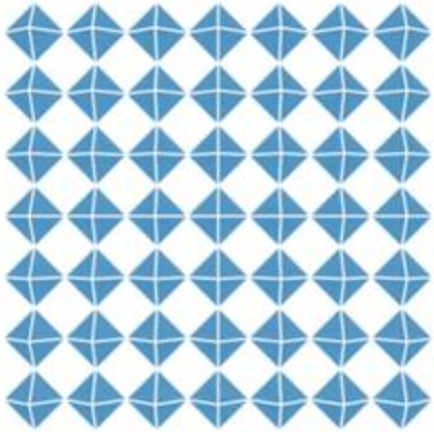
Many possible materials features
from atomistic to macroscopic length scales

Electronic



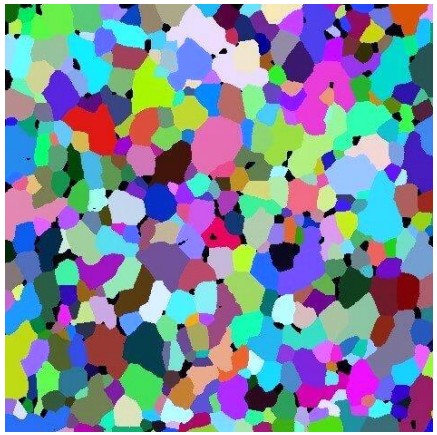
Wavefunctions
or electron density
(Å)

Atomic scale



Local atomic
connectivity
(nm)

Microstructure



Grain size and
orientation
(μm)

Macroscale

Shape
(cm)

Image after Taylor Sparks (University of Utah)

Hot Encoding

We can use an *n*-dimensional vector to categorise the atomic number of the elements in a compound

Element (One-hot)

[1000000000...]

H He Li Be B C N O F....

Compound (Multi-hot)

[0000010100...]

H He Li Be B C N O F....

'1' indicates the presence of that specific element and '0' for others

Hand-Built (Local) Representations

We can *define* elemental feature vectors based on standard properties of the elements

```
import elementembeddings

print(AtomEmbeds["magpie"].dim)

22

print(AtomEmbeds["magpie"].feature_labels)

['Number', 'MendeleevNumber', 'AtomicWeight', 'MeltingT', 'Column', 'Row', '
CovalentRadius', 'Electronegativity', 'NsValence', 'NpValence', 'NdValence',
 'NfValence', 'NValence', 'NsUnfilled', 'NpUnfilled', 'NdUnfilled', 'NfUnfilled',
'NUnfilled', 'GSvolume_pa', 'GSbandgap', 'GSmagmom', 'SpaceGroupNumber']
```

22 dimensional Magpie representation from

L. Ward et al, npj Comp. Mater. 2, 16028 (2016)

Hand-Built (Local) Representations

We can also *define* compound feature vectors based on standard properties of the elements

```
Fe203_magpie = CompositionalEmbedding("Fe203", "magpie")
```

$$\mathbf{X}(Fe_2O_3) = [2\mathbf{X}(Fe) + 3\mathbf{X}(O)]/5$$

	X_1	X_2	X_3	… X_n
Fe	0.52	0.11	0.01	0.80
O	0.32	0.23	0.14	0.64
Fe_2O_3	0.40	0.18	0.09	0.70

Different types of pooling is possible (e.g. max, min, mean)

Learned (Distributed) Representations

We can *learn* continuous feature vectors with elemental information as part of model training

SkipSpecies
200 D
Structure graph pooling

APL Machine Learning

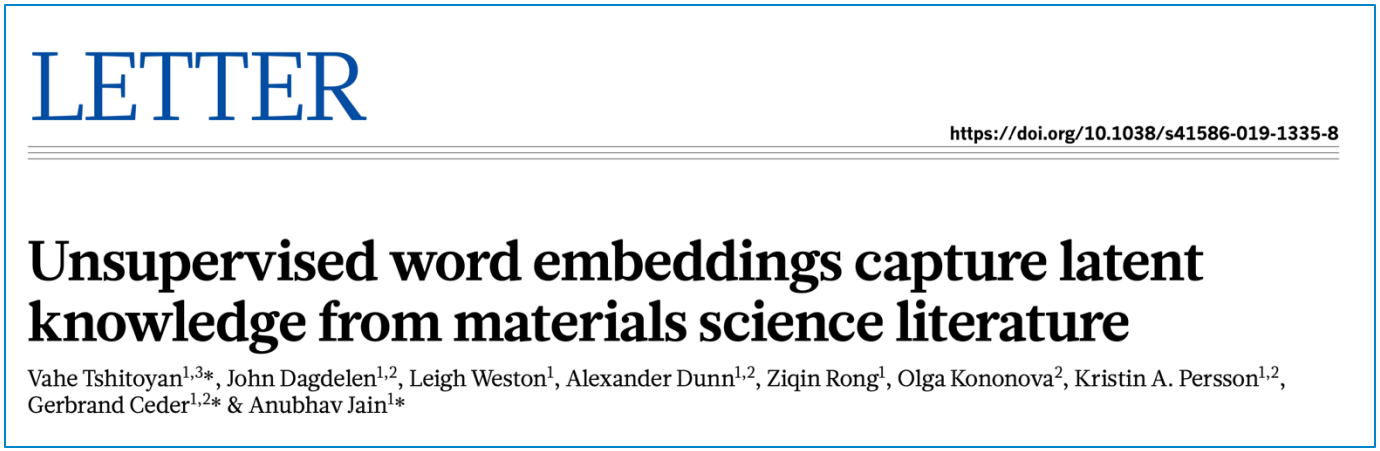
Ionic species representations for materials informatics

Cite as: APL Mach. Learn. 2, 036112 (2024); doi: 10.1063/5.0227009
Submitted: 5 July 2024 • Accepted: 28 August 2024 •
Published Online: 19 September 2024

View Online Export Citation CrossMark

Anthony Onwuli,[1] Keith T. Butler,[2,a)] and Aron Walsh[1,b)]

Mat2Vec
200 D
Literature word embedding



LETTER

https://doi.org/10.1038/s41586-019-1335-8

Unsupervised word embeddings capture latent knowledge from materials science literature

Vahe Tshitoyan[1,3]*, John Dagdelen[1,2], Leigh Weston[1], Alexander Dunn[1,2], Ziqin Rong[1], Olga Kononova[2], Kristin A. Persson[1,2], Gerbrand Ceder[1,2]* & Anubhav Jain[1]*

https://github.com/WMD-group/ElementEmbeddings

Element Embeddings

Toolkit to access and modify elemental and compositional representations for machine learning

ElementEmbeddings

Made with Python | License MIT | code style black | open issues 6 | ElementEmbeddings CI passing

codecov 71% | DOI 10.5281/zenodo.8117601 | pypi v0.1.1 | docs mkdocs material | python 3.8 | 3.9 | 3.10

The **Element Embeddings** package provides high-level tools for analysing elemental embeddings data. This primarily involves visualising the correlation between embedding schemes using different statistical measures.

Motivation

Machine learning approaches for materials informatics have become increasingly widespread. Some of these involve the use of deep learning techniques where the representation of the elements is learned rather than specified by the user of the model. While an important goal of machine learning training is to minimise the chosen error function to make more accurate predictions, it is also important for us material scientists to be able to interpret these models. As such, we aim to evaluate and compare different atomic embedding schemes in a consistent framework.

Getting started

ElementEmbeddings's main feature, the Embedding class is accessible by importing the class.

Dr Anthony Onwuli

Latest embeddings

CrystaLLM

SkipSpecies

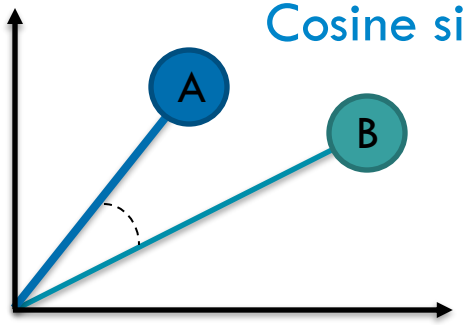
CGNF

https://github.com/WMD-group/ElementEmbeddings

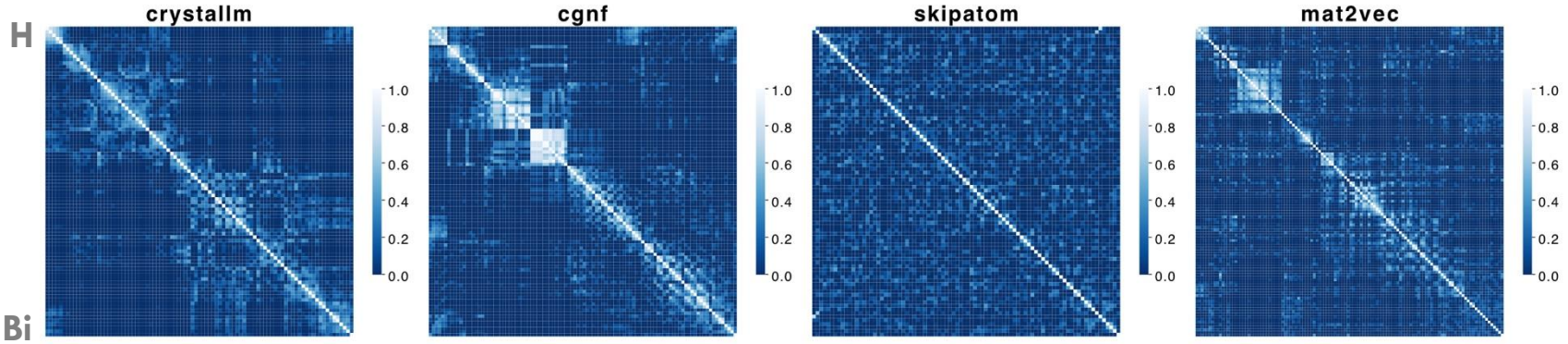
Learned Chemical Similarity

Quantify with distance (e.g. Chebyshev), similarity (e.g. Cosine), or correlation (e.g. Pearson) metrics

Name	Dimension	Type
Magpie	22	Element properties
Mat2Vec	200	Chemical abstracts
Skipatom	200	Crystal structure graphs
MegNet	16	Graph neural network
CrystaLLM	512	Crystal structure text



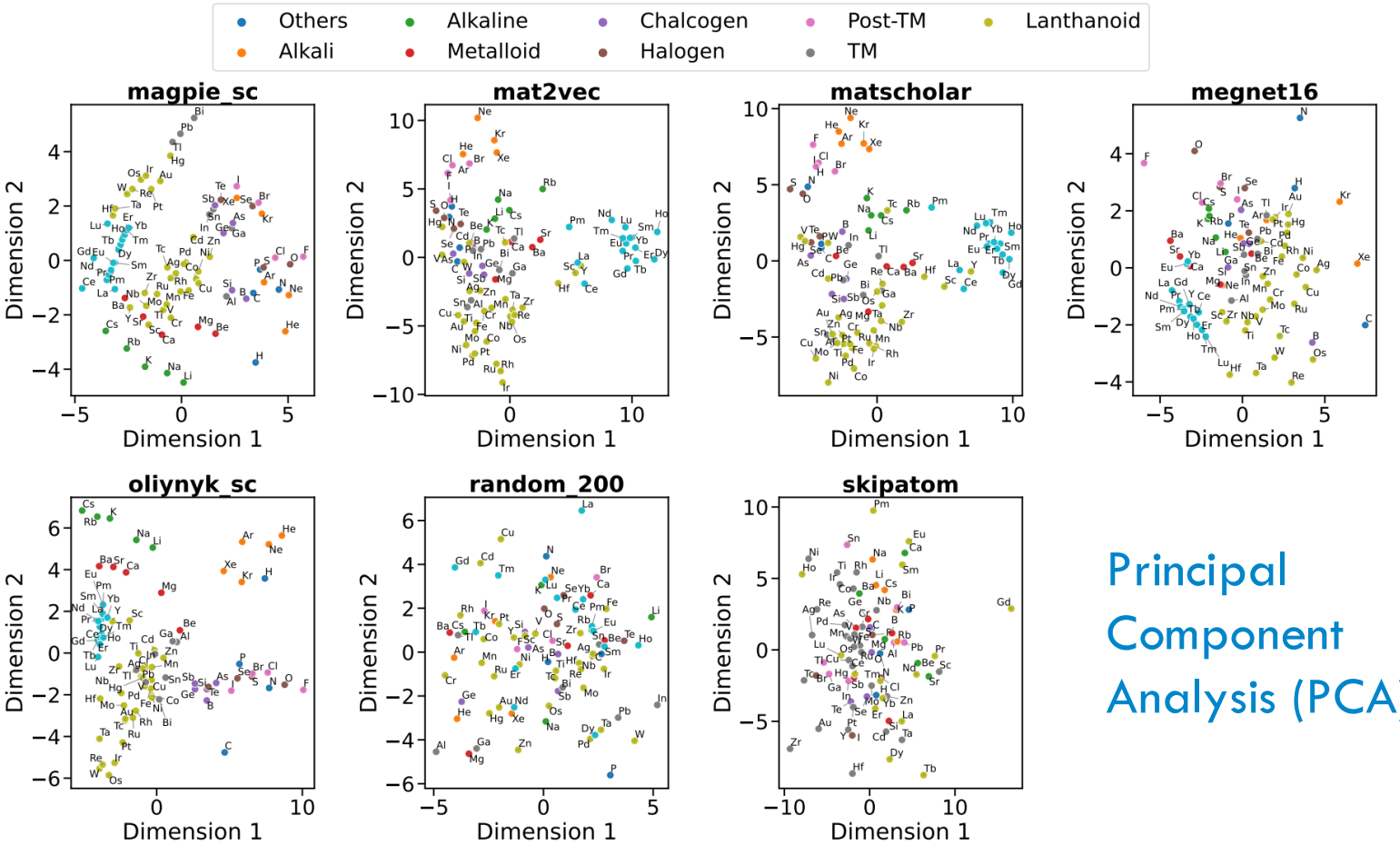
$$\cos(\theta) = \frac{\boldsymbol{A} \cdot \boldsymbol{B}}{\|\boldsymbol{A}\| \|\boldsymbol{B}\|}$$



Anthony Onwuli et al, Digital Discovery 2, 1558 (2023)

Learned Chemical Similarity

Dimensionality reduction confirms a natural clustering of elements into "groups"



Principal Component Analysis (PCA)

Anthony Onwuli et al, Digital Discovery 2, 1558 (2023)

Class Outline

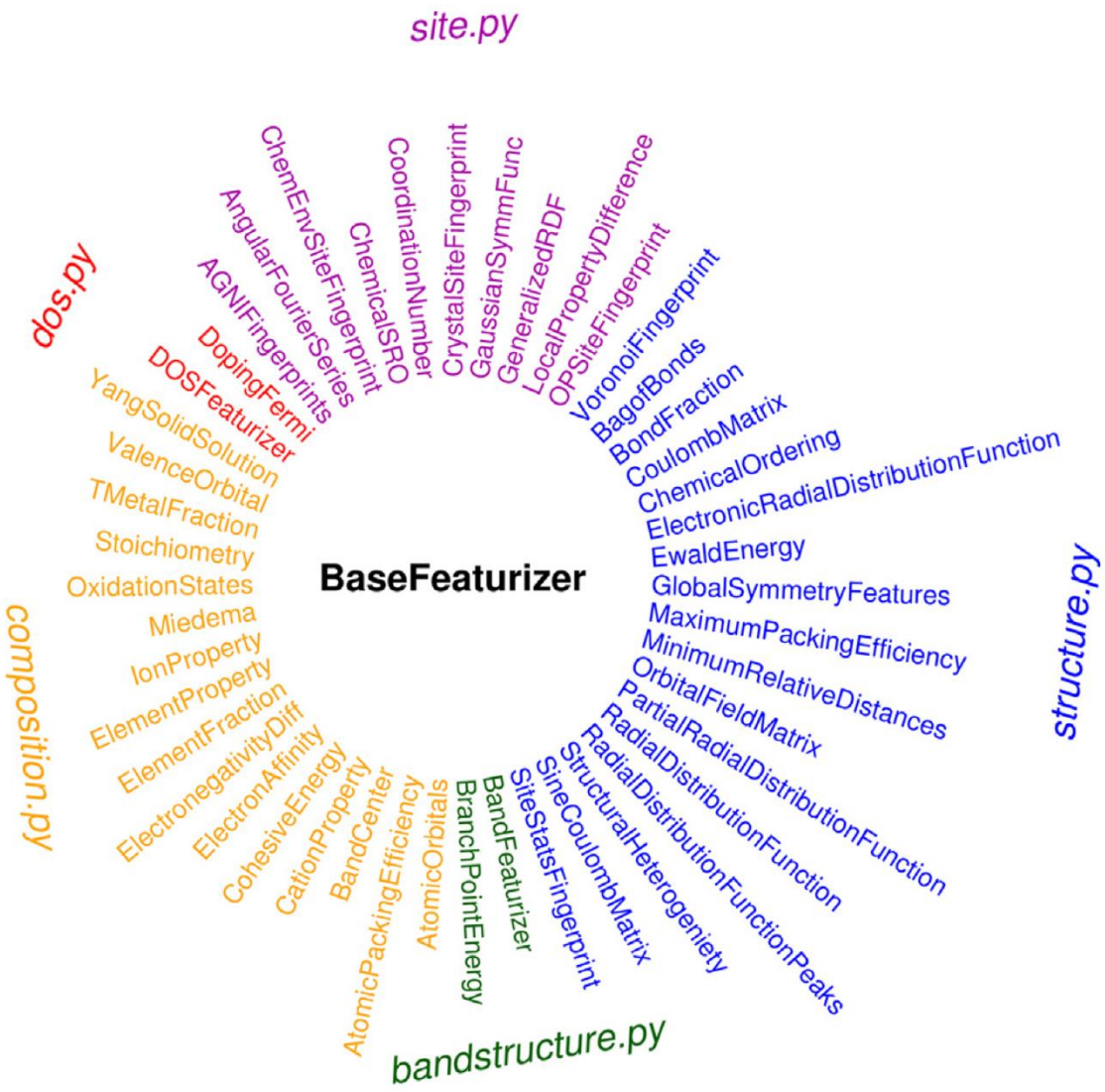
Crystal Representations

A. Compositional

B. Structural

C. Graphs

Many Possible Materials Features



Implemented in https://github.com/hackingmaterials/matminer

Learn from Crystallography

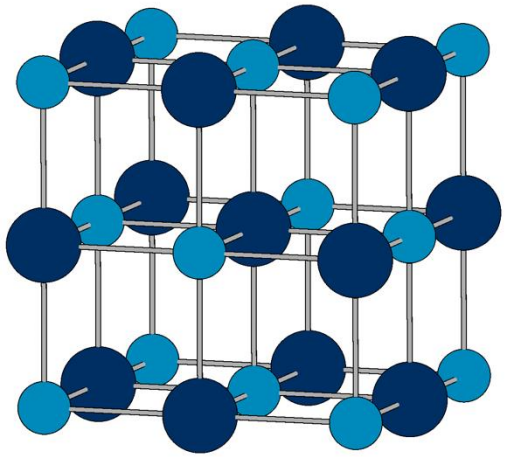
High symmetry crystal:

MgO

Cubic

8 atom unit cell

$a = b = c$



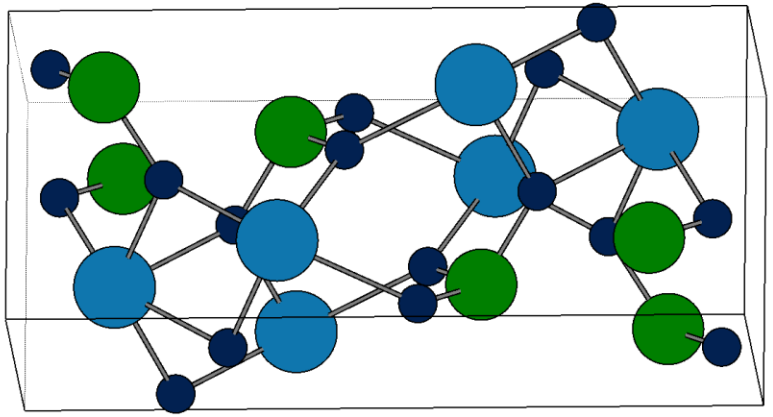
Low symmetry crystal:

$BiVO_4$

Monoclinic

24 atom unit cell

$a \neq b \neq c$



Learn from Crystallography

7 crystal systems, 14 Bravais lattices,
230 space groups, 10^3 prototype structures

Conventional description

Unit cell ($\mathcal{L}$)

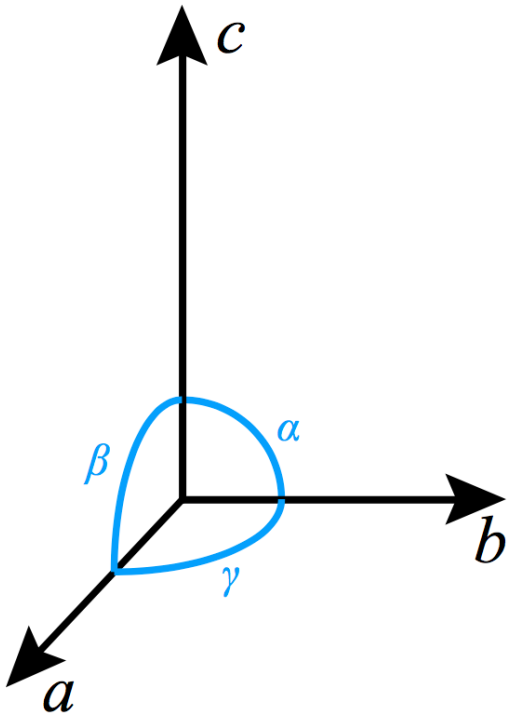
$a, b, c, \alpha, \beta, \gamma$

Fractional coordinates ($\mathcal{X}$)

$(x_1, y_1, z_1)\ldots$

Atom types ($\mathcal{A}$)

Sn, Ti, O…



Problem for ML: conventional description lacks invariance*

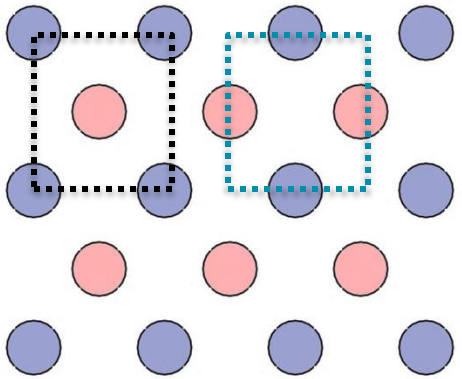
*with respect to atomic permutation, unit cell rotations, and translations

Unit Cell Transformations

The same structure is described in each case

Two-atom orthorhombic unit cell

$$\begin{bmatrix} a & b & c \\ x_1 & y_1 & z_1 \\ x_2 & y_2 & z_2 \end{bmatrix} \quad \begin{bmatrix} 4 & 5 & 6 \\ 0 & 0 & 0 \\ 0.5 & 0.5 & 0.5 \end{bmatrix}$$



Atomic permutation

$$\begin{bmatrix} 4 & 5 & 6 \\ 0.5 & 0.5 & 0.5 \\ 0 & 0 & 0 \end{bmatrix}$$

Crystal rotation

$$\begin{bmatrix} 5 & 4 & 6 \\ 0.5 & 0.5 & 0.5 \\ 0 & 0 & 0 \end{bmatrix}$$

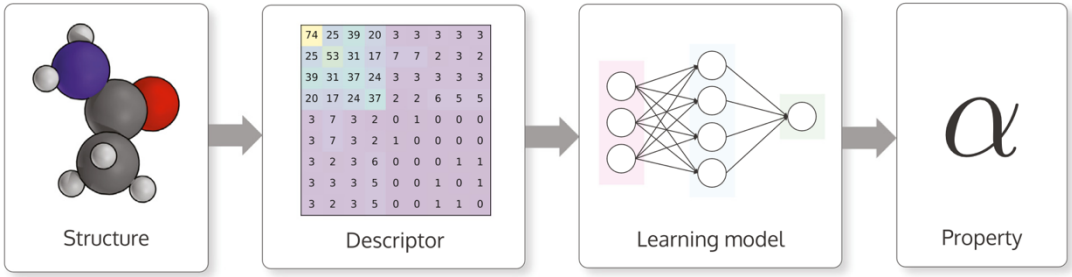
Unit cell translation

$$\begin{bmatrix} 4 & 5 & 6 \\ 0.0 & 0.5 & 0.5 \\ 0.5 & 0 & 0 \end{bmatrix}$$

ML models based on variant representations struggle to generalise

Structural Representations

Many structural descriptors have been developed

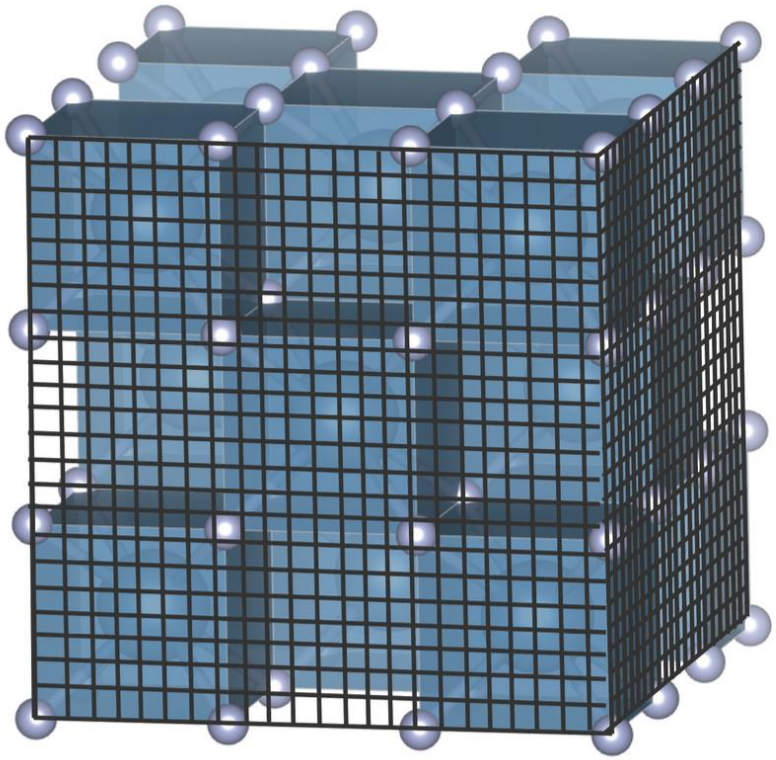


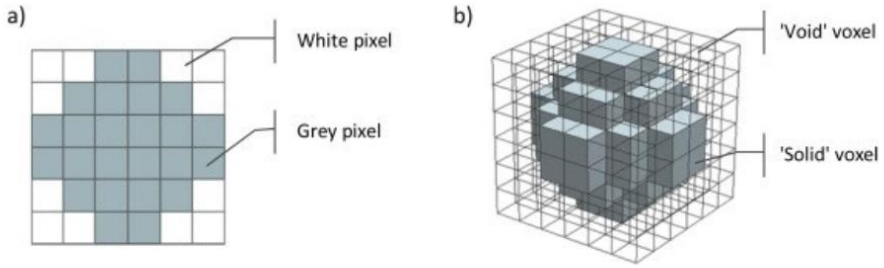
- **Coulomb Matrix** (Rupp et al, 2012)
  - mimics electrostatic interactions ($q_i q_j / r_{ij}$)
- **Atom-Centered Symmetry Functions** (Behler, 2011)
  - site expansion of radial and angular terms
- **Many Body Tensor Representation** (Huo et al, 2017)
  - distribution of local structural motifs
- **Atomic Cluster Expansion** (Drautz, 2019)
  - high body-order expansion of atomic environments

Several are implemented in https://singroup.github.io/dscribe

Real Space Grid

Voxels (three-dimensional pixels) used in computer graphics can describe a unit cell





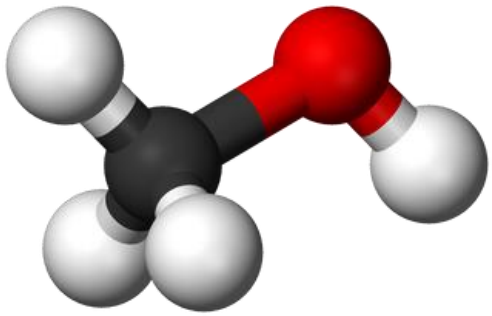
Used in early materials ML, but not recommended for structure

Image courtesy of Taylor Sparks (University of Utah)

Pairwise Interatomic Distances

Coulomb matrix is a global descriptor that mimics the electrostatic interaction between nuclei

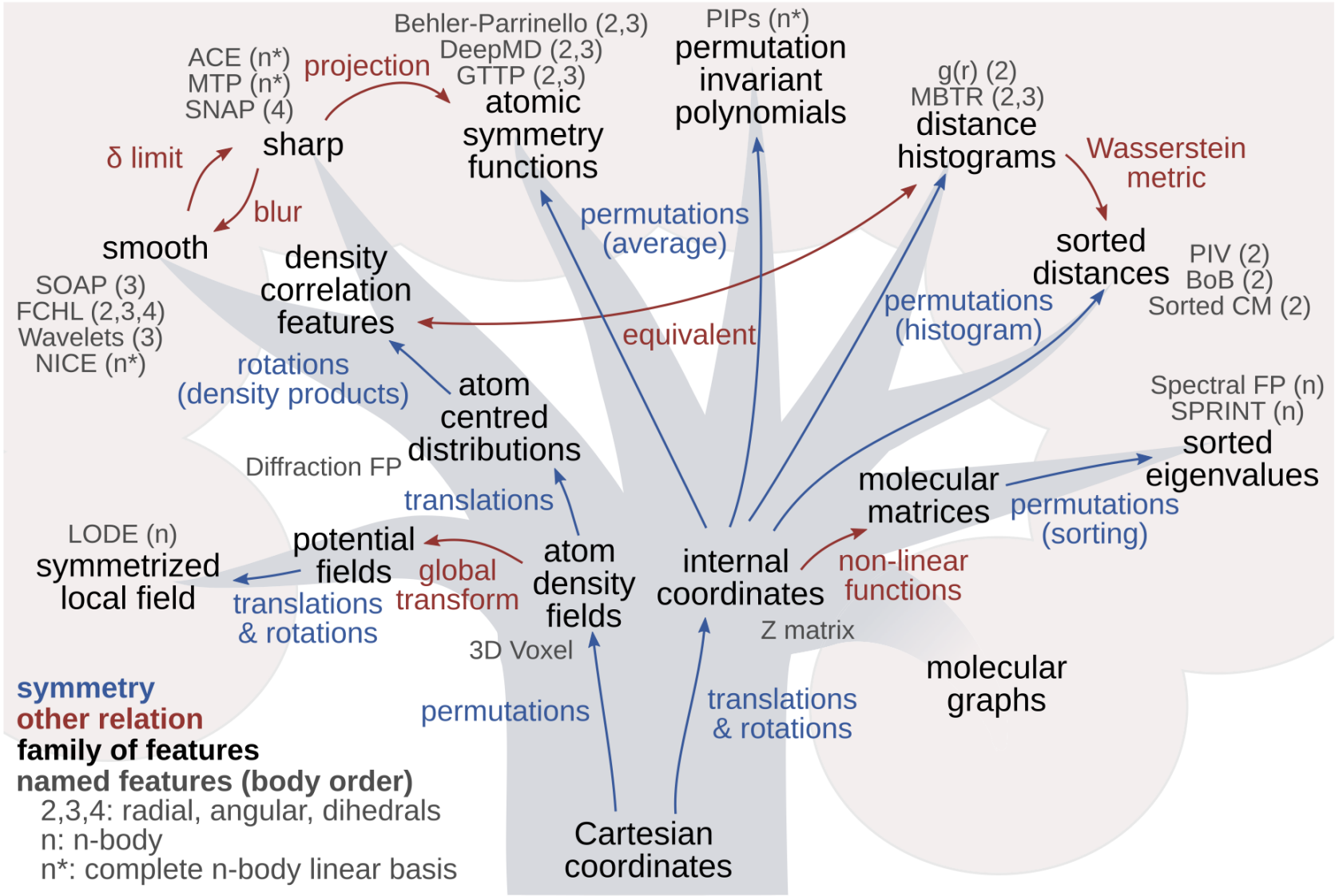
$$M_{ij}^{\text{Coulomb}} = \begin{cases} 0.5 Z_i^{2.4} & \text{for } i = j \\ \frac{Z_i Z_j}{R_{ij}} & \text{for } i \neq j \end{cases}$$



$$\begin{bmatrix} 36.9 & 33.7 & 5.5 & 3.1 & 5.5 & 5.5 \\ 33.7 & 73.5 & 4.0 & 8.2 & 3.8 & 3.8 \\ 5.5 & 4.0 & 0.5 & 0.35 & 0.56 & 0.56 \\ 3.1 & 8.2 & 0.35 & 0.5 & 0.43 & 0.43 \\ 5.5 & 3.8 & 0.56 & 0.43 & 0.5 & 0.56 \\ 5.5 & 3.8 & 0.56 & 0.43 & 0.56 & 0.5 \end{bmatrix}$$

Sine matrix is a modification that accounts for periodicity

Invariant Structural Representations



Comprehensive review: F. Musil et al, Chem. Rev. 121, 9759 (2021)

Invariant Structural Representations

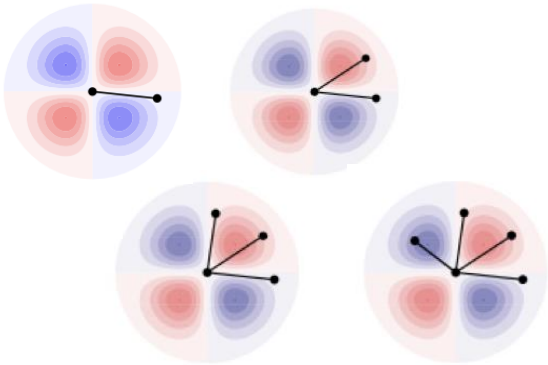
Atomic Cluster Expansion (ACE) provides a systematic representation of atomic environments through radial (R) and angular (Y) terms

Site basis function $\phi(r) = R_l \; Y_l^m$

Permutation invariance $\boldsymbol{A_i} = \sum_{neighbours} \phi(r)$

Rotation (Q) invariance $\boldsymbol{B_i} = \int \boldsymbol{A_i} \; dQ$

Product basis ***B*** forms a body-order expansion



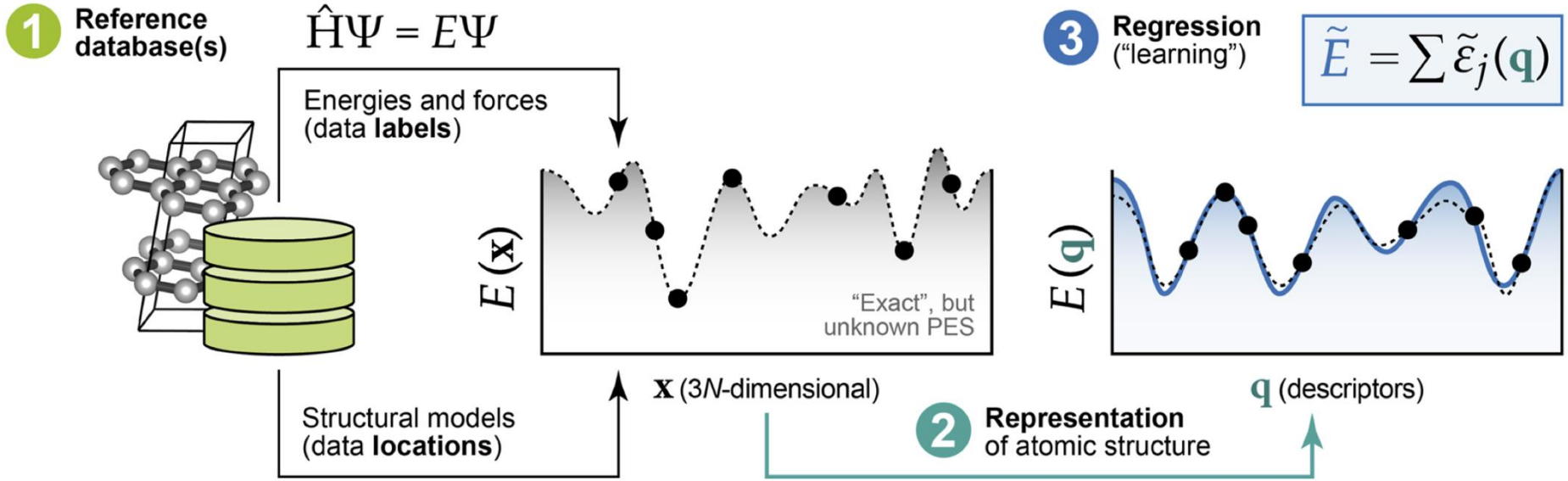
Property $= f(\boldsymbol{B}, \boldsymbol{\Theta})$ weights

ACE is used in linear and deep learning models for materials

R. Drautz, Phys. Rev. B. 99, 014104 (2019); arXiv:2311.16326 (2023)

ML Powered Molecular Dynamics

Classical models are being complemented by machine learning force fields (MLFF)

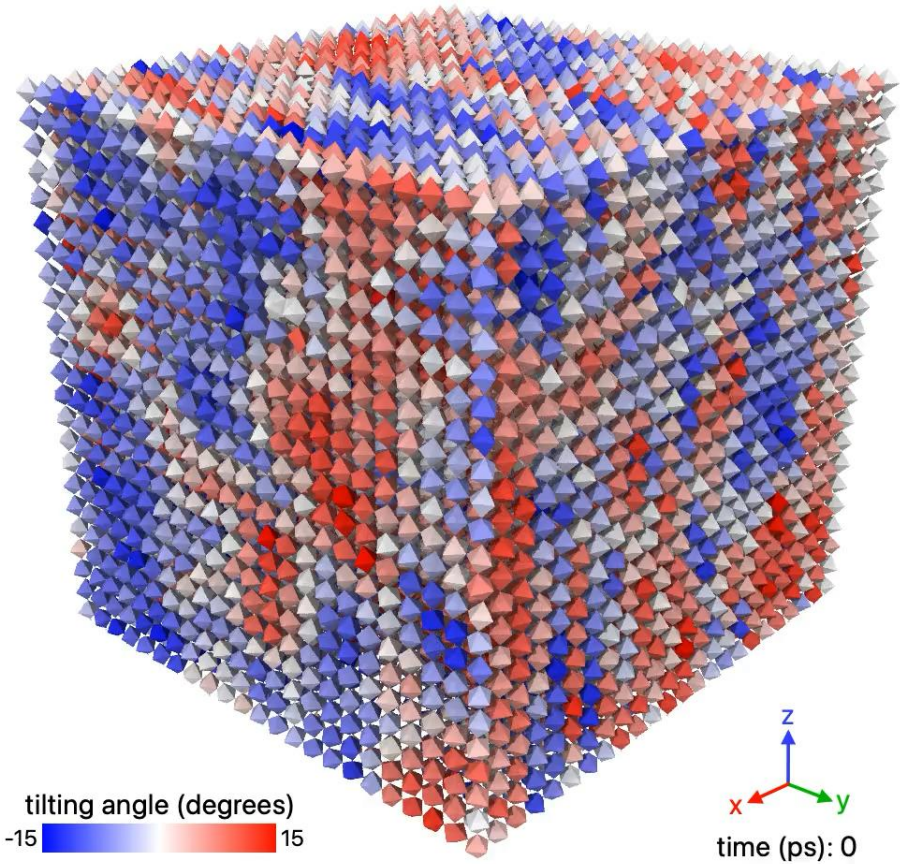


Three start-of-the-art implementations based on equivariant neural network regression are MACE, Allegro, and SevenNet

J. D. Morrow, J. L. A. Gardner and V. Deringer, J. Chem. Phys. 158, 121501 (2023)

ML Powered Molecular Dynamics

Enable large-scale simulations of complex materials such as organic-inorganic solids



69,120 atom simulation of $CsPbI_3$ perovskite based on the atomic cluster expansion (ACE)

Animation by Will Baldwin

(Small 20, 2303565, 2024)

Xia Liang et al, J. Phys. Chem. C 127, 12941 (2023)

Class Outline

Crystal Representations

A. Compositional

B. Structural

C. Graphs

Graphs

Graphs are a representation common to many domains and problems

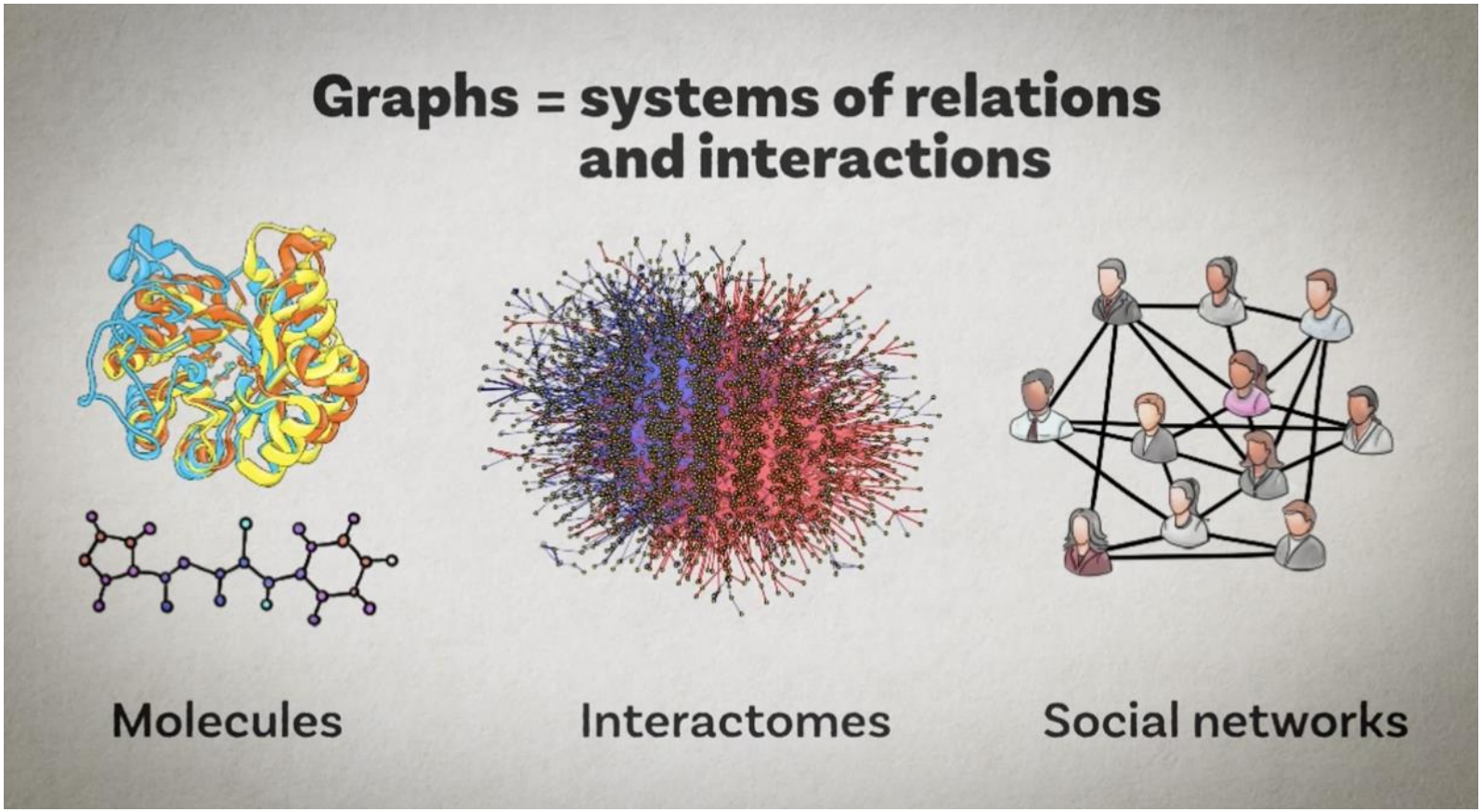
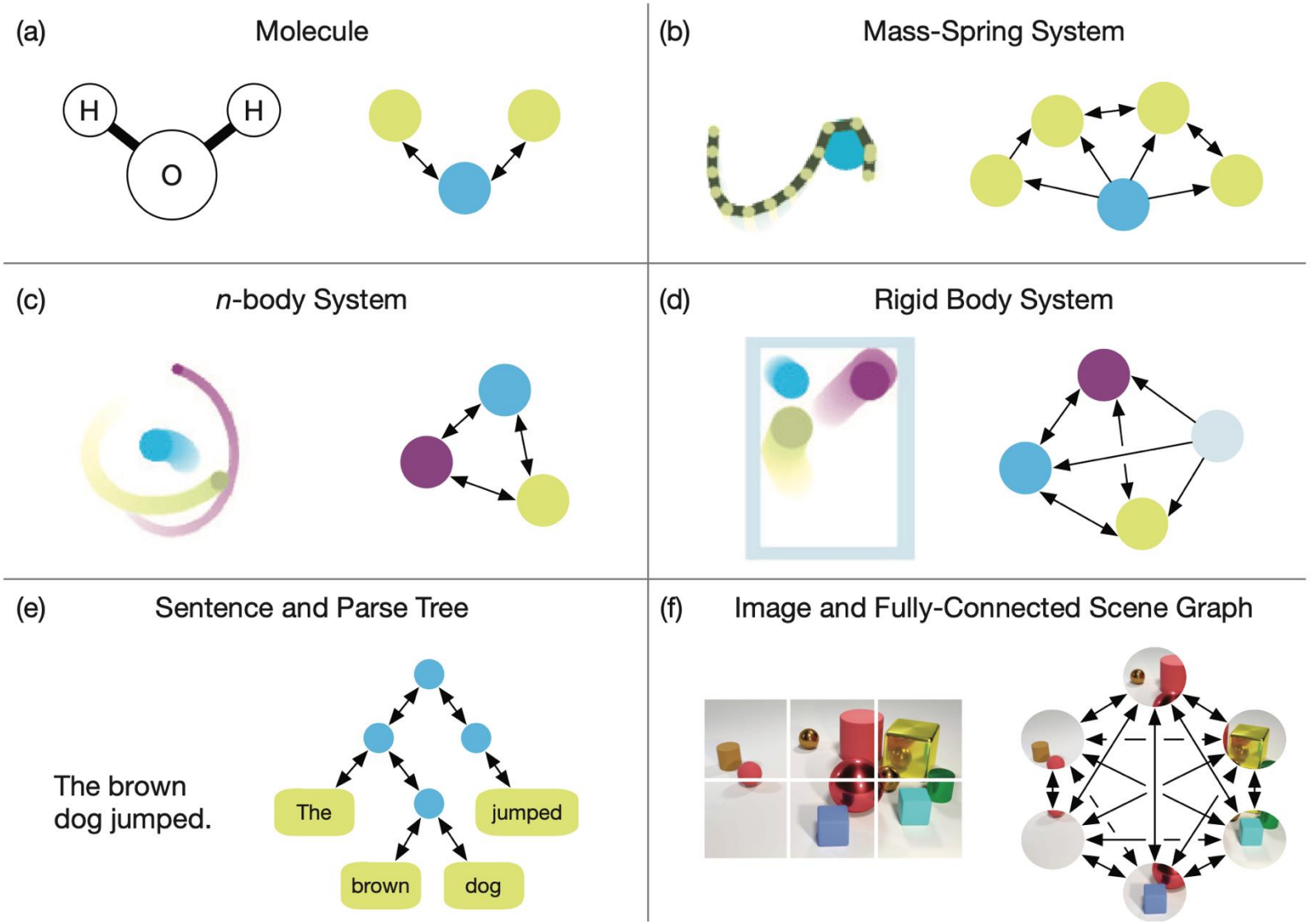


Image courtesy of Michael Bronstein (University of Oxford)

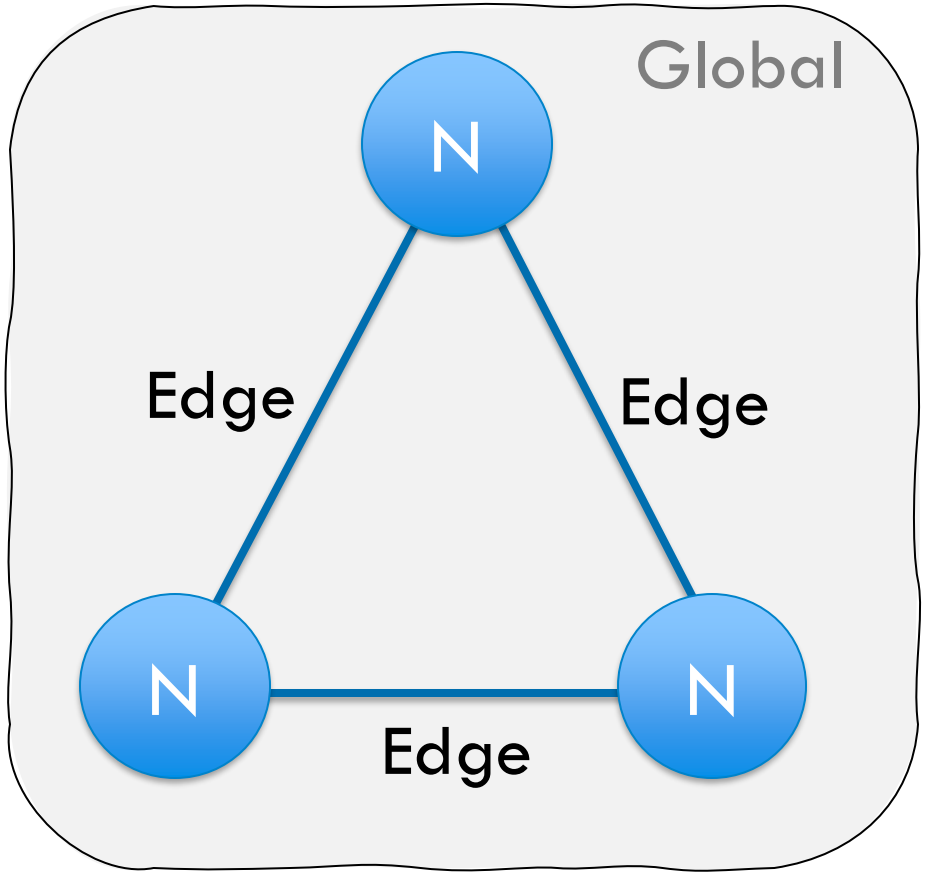
Graphs



P. W. Battaglia et al, arXiv:1806.01261 (2018)

Graph Components

Nodes (Vertices), **E**dges, **G**lobal Attributes



Crystal systems

N – atoms

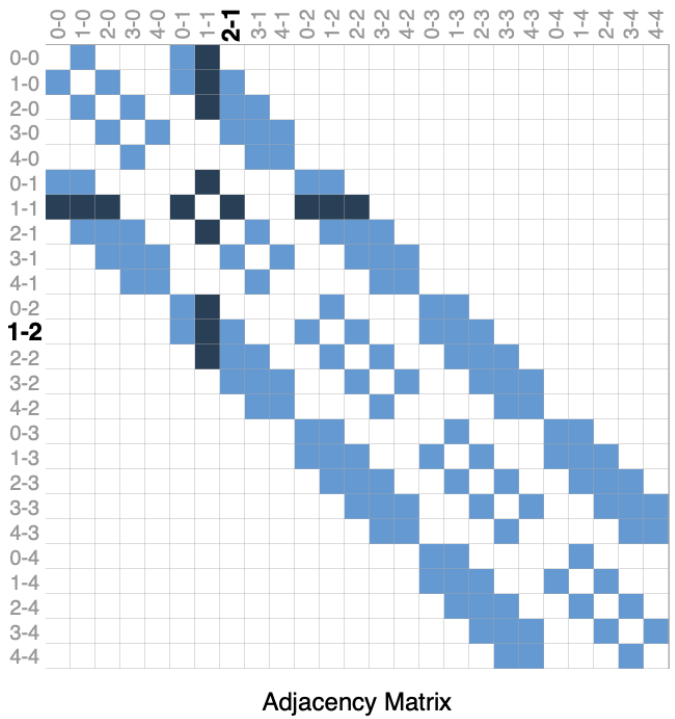
E – bonds

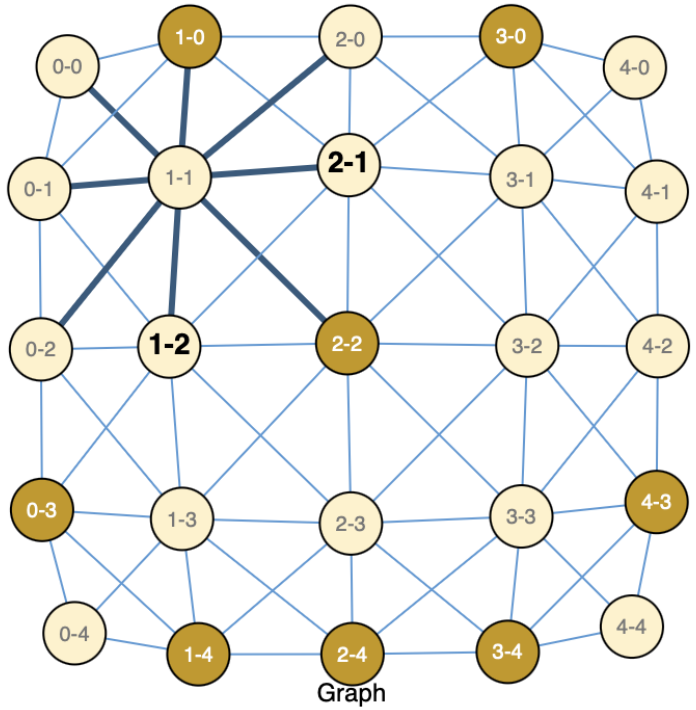
G – unit cell or materials properties

Vectors can be associated with each component to encode & exchange information

Graph Components

Nodes (Vertices), **E**dges, **G**lobal Attributes



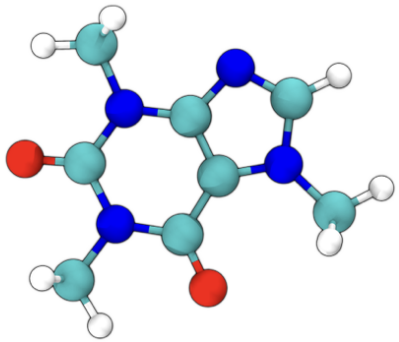


Graphs can be fully connected (every node connected to every other node), but sparse connections are often used

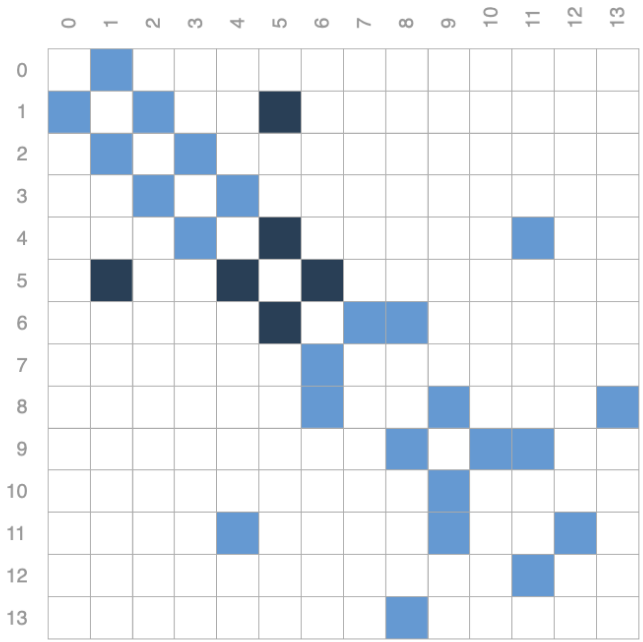
Image from https://distill.pub/2021/gnn-intro

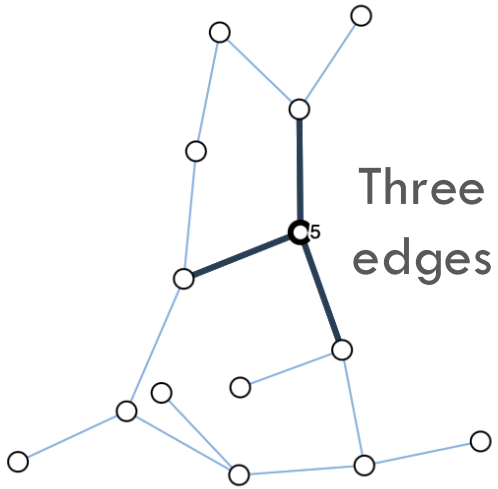
Graph Components

Nodes (or Vertices), **E**dges, **G**lobal Attributes



Molecule
(including H)





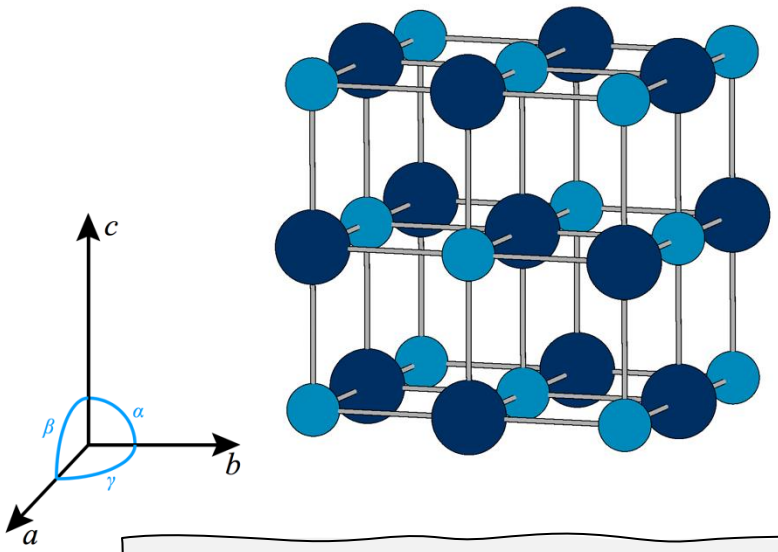
Graph
(excluding H nodes)

For chemical problems, nearest-neighbour connectivity is common, as used in “ball and stick” representations

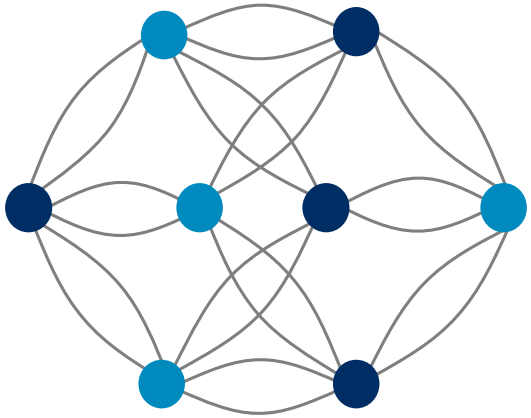
Image from https://distill.pub/2021/gnn-intro

Crystal Graphs

Standard crystallographic representation of materials



Crystal graph representation



Fractional positions xyz of atoms within a unit cell formed of lattice vectors *abc*

Effective for humans

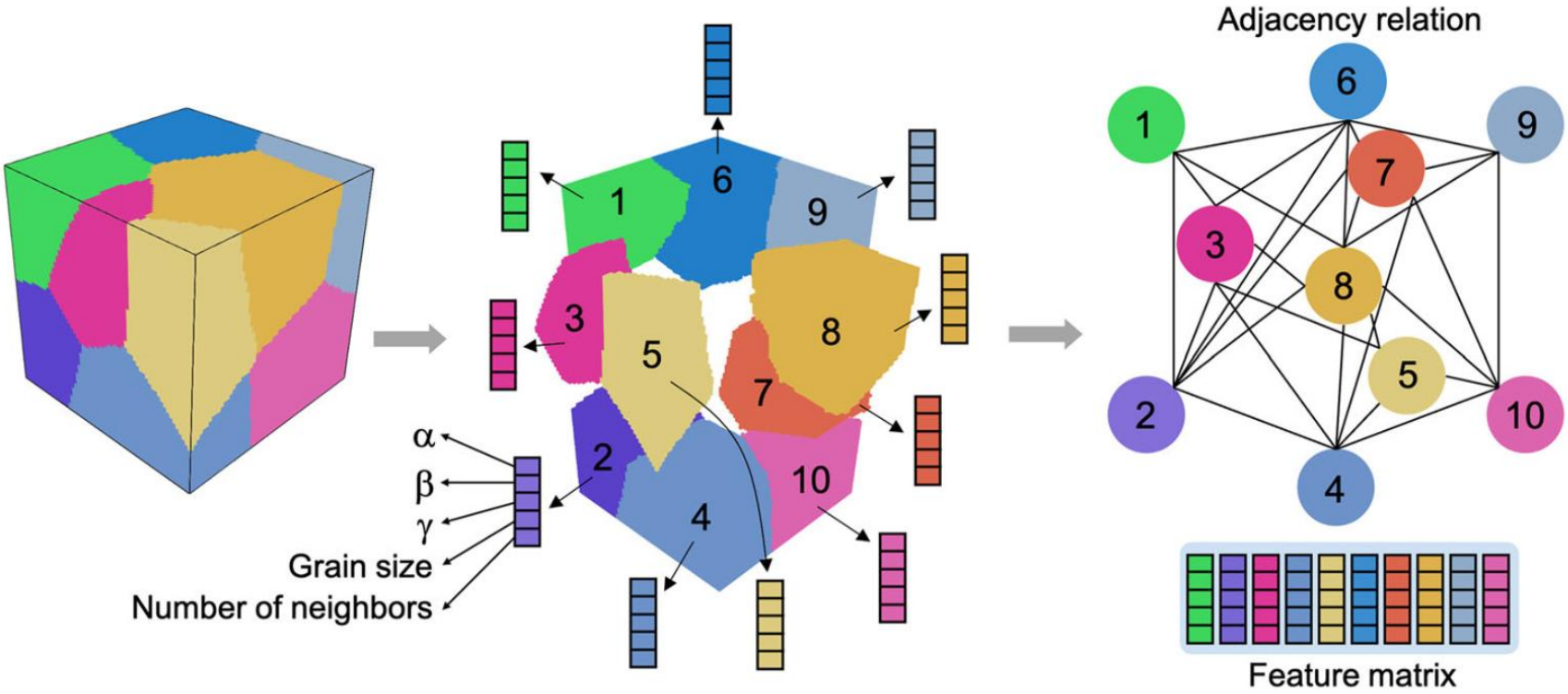
Nodes (atoms) connected by edges (bonds). Multiple edges can describe periodicity

Effective for ML models

T. Xie and J. C. Grossman, Phys. Rev. Lett. 120, 145301 (2018)

Materials Graphs

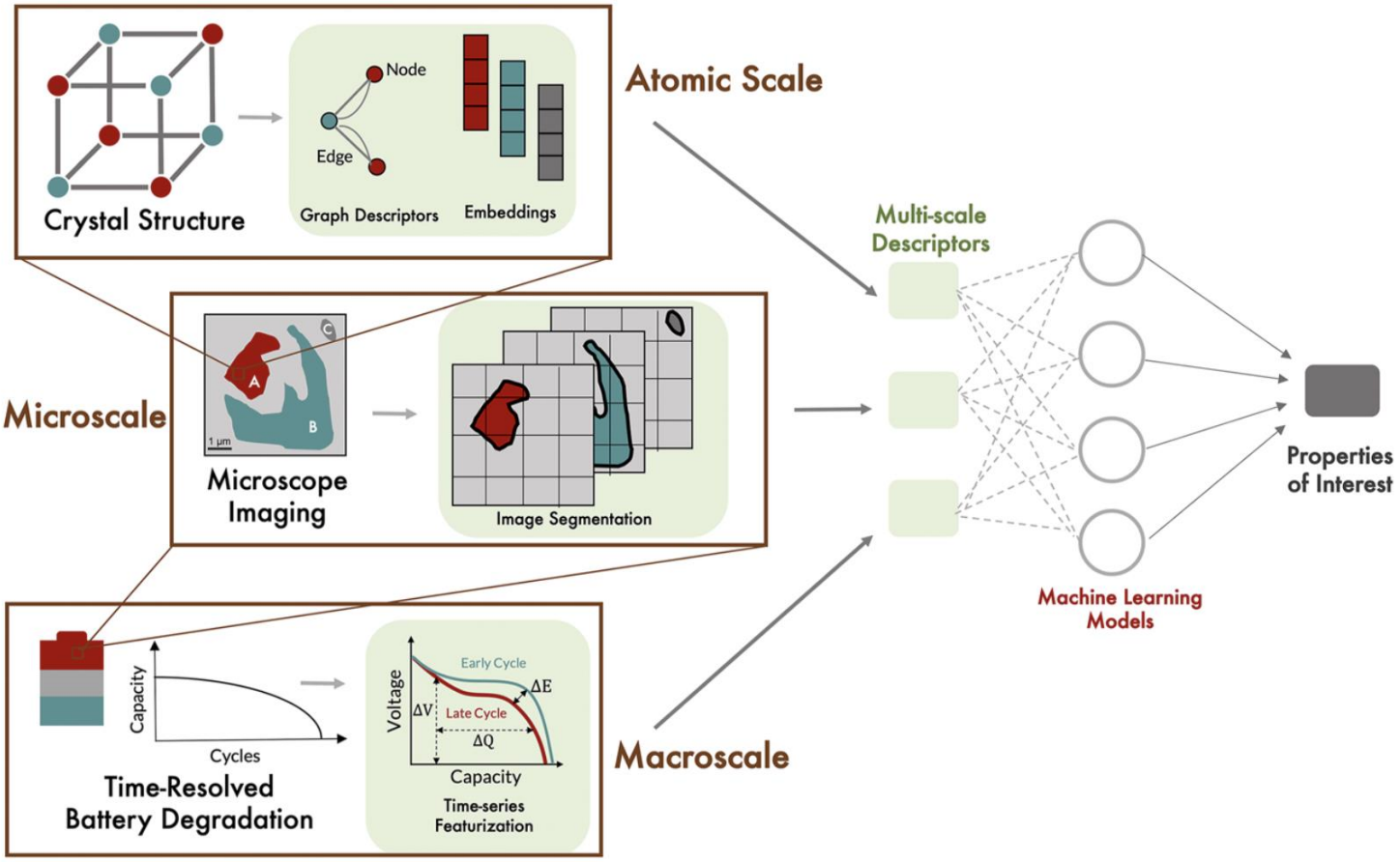
Nodes can be used to represent larger structural units of a crystal or even entire grains



M. Dai et al, npj Comp. Mater. 7, 103 (2021)

Multi-Scale Representations

Ongoing efforts to combine features that bridge from the micro to macroscale; from atoms to devices



S. B. Torrisi et al, APL Machine Learning 1, 020901 (2023)

Class Outcomes

1. Describe the ways that chemical composition can expanded into vectors
2. Explain how the structure of a material can be represented for machine learning
3. Consider the limitations of a graph-based description of a three-dimensional structure

Activity:

Chemical space