ME 5374-ST

ML4MSD

Machine Learning for Materials Science and Discovery

Fall 2025

Asst. Prof. Peter Schindler

Lecture 10 – Featurization of Materials

- Requirements for an Ideal Descriptor
- Hierarchy of Descriptors
- Compositional Descriptors
- Local and Global Atomistic Descriptors
- Global Atomistic Descriptors
- Coarse-Grained Descriptors

Overview and Terminology: Features / Descriptors / Fingerprints

Goal: Describe a material with a numeric representation

- This numeric representation has to be rich in information for the ML algorithm to learn from the input data
- Features should be correlated with the target label

Featurization:

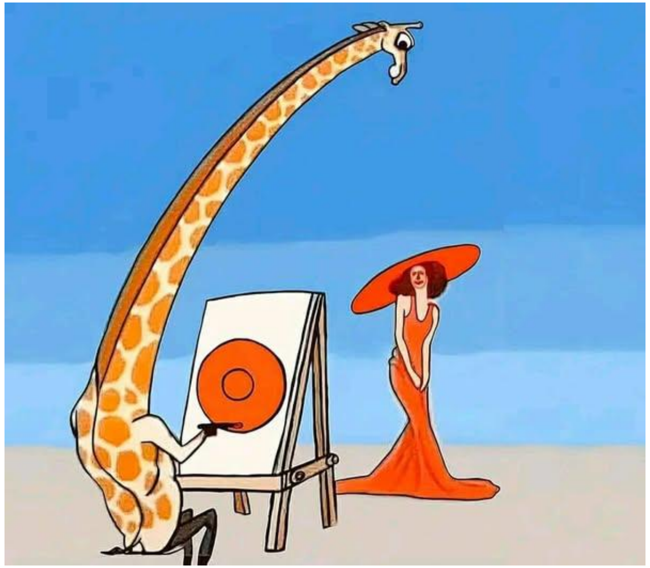
Transforming data input (material, text, image, etc.) into numerical "feature vector"

Feature Engineering:

Take feature vectors (already numeric) and improve them

Not all descriptors are equally useful…

DOES THAT SATISFY YOUR COMMITMENT TO THE TRADE DESCRIPTION ACT?
TEPID BROWN MUDDY STUFF



Upper: ChatGPT, Lower: Couldn't find original source…

Requirements for an Ideal Descriptor

i. **Meaningful** (and compact)
Relationship between descriptor and response not overly complex

ii. **Universal**
Can be applied to any existing and hypothetical material
ii.a Fixed in number
Same number of descriptors *regardless* of input material

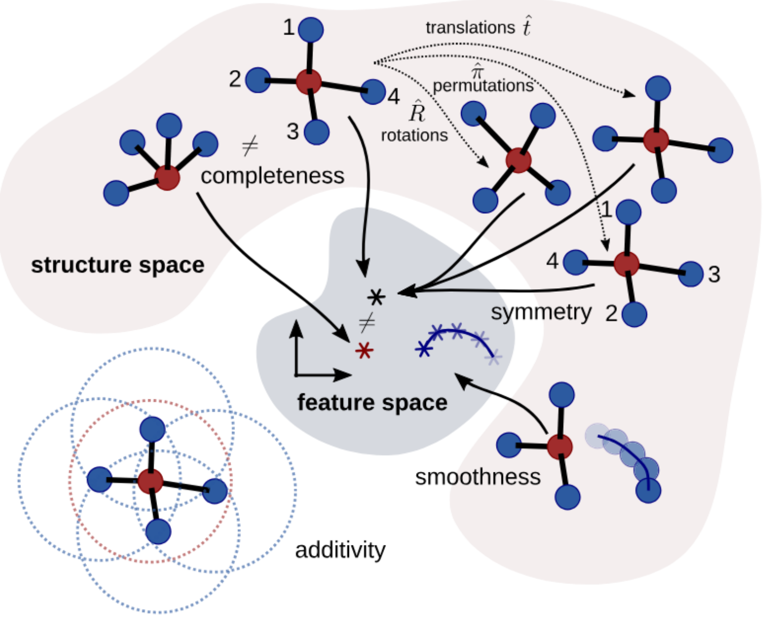
iii. **Invariant under crystal symmetries** (and permutations)

iv. **Reversible** (i.e. be a unique description)
List of descriptors can (in principal) be reversed back into a description of a material (enables *inverse design*!)

v. **Continuous**
Small change in atomic structure = small change in descriptor

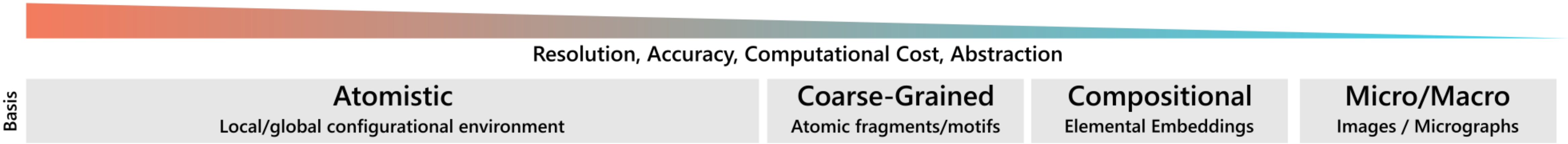
vi. **Computationally cheap(er)**
Should be easier to obtain than target property itself

vii. **Uncorrelated** (ideally)
Can be fixed with feature selection/regularization



[15]

Hierarchy of Descriptors



Resolution, Accuracy, Computational Cost, Abstraction

Basis

Atomistic
Local/global configurational environment

Coarse-Grained
Atomic fragments/motifs

Compositional
Elemental Embeddings

Micro/Macro
Images / Micrographs

Compositional Descriptors

Based on chemical formula of structure: $\boxed{A_aB_bC_c....}$

1. One-Hot Encoding Vectors
2. Stoichiometric attributes
   Number of elements, *p*-norm of fraction vector $\left\| \left[\frac{N}{a}, \frac{N}{b}, \frac{N}{c}, \ldots \right] \right\|_p$ $(N=a+b+c+\ldots)$
3. Elemental property (*P*) statistics
   $g\left[a \cdot [P(A)], b \cdot [P(B)], c \cdot [P(C)], \ldots\right]$
   g = Min, Max, Mean,
   Range, StDev,…
4. Ionic compound properties

P= a) *General properties*
Position periodic table, Mendeleev number, $N_{\text{valence electrons}}$

b) *Electronic structure*
Fraction of filled/unfilled electrons in s, p, d, and f shells

c) *Measured properties (molecular)*
Atomic mass, electron affinity, atomic radius

d) *Derived Properties (molecular)*
Covalent radius, electronegativity, polarizability

e) *Elemental crystal properties (measured/calculated)*
BCC bandgap, lattice constant, DFT volume/atom, E_{cohesive}

[14]

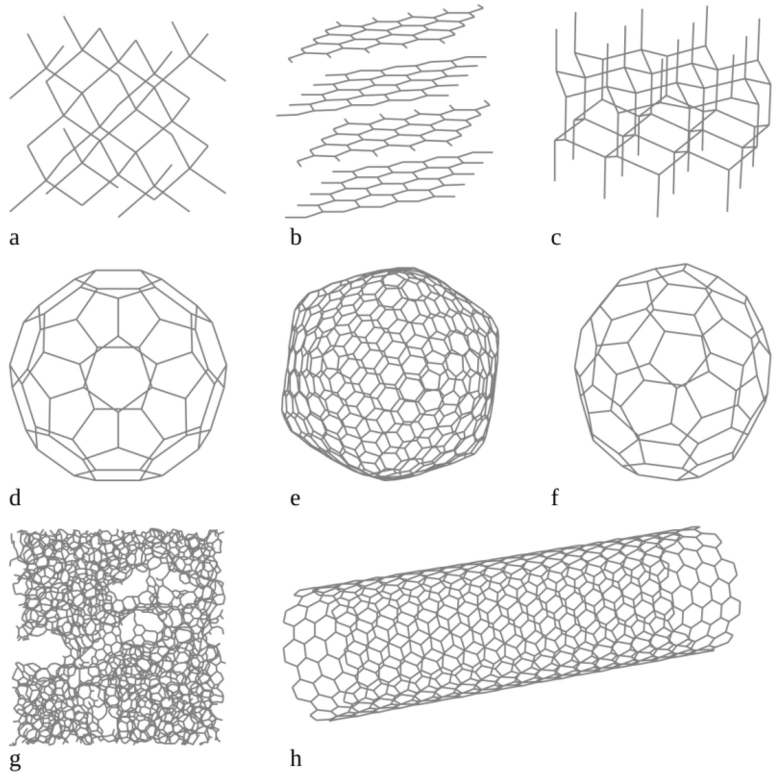
Compositional Descriptors

These descriptors are unique for any given chemical formula

However, *different phases* with same chemical formula are described by the same set of descriptors

Potential extension

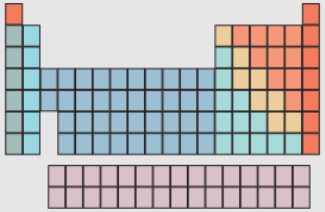
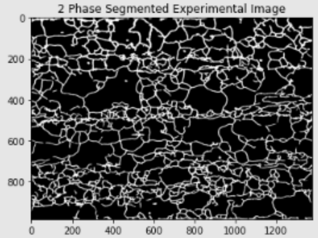
- Add information about spacegroup/crystal system or structure prototypes
- Packing fraction
- Meso-scale descriptors
- Add experimental/processing conditions



wikipedia.org

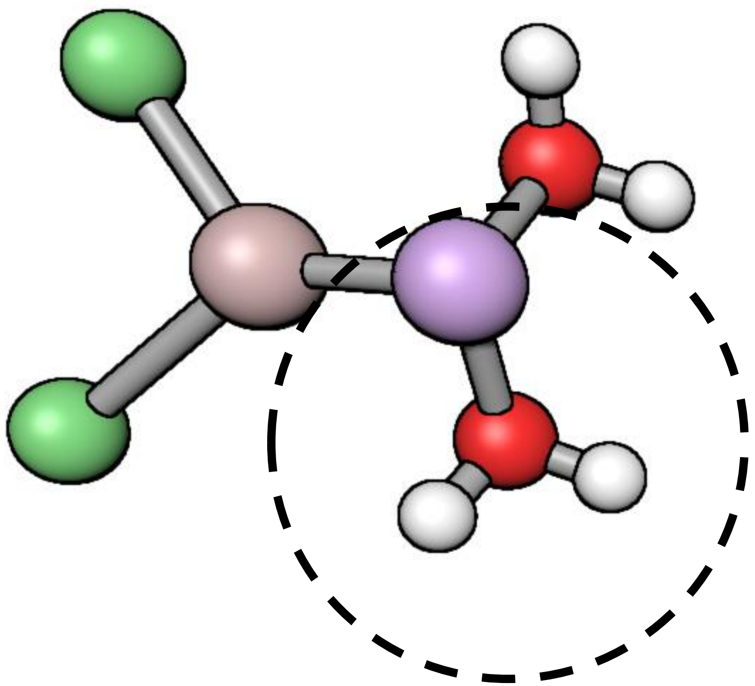
Hierarchy of Descriptors

Resolution, Accuracy, Computational Cost, Abstraction

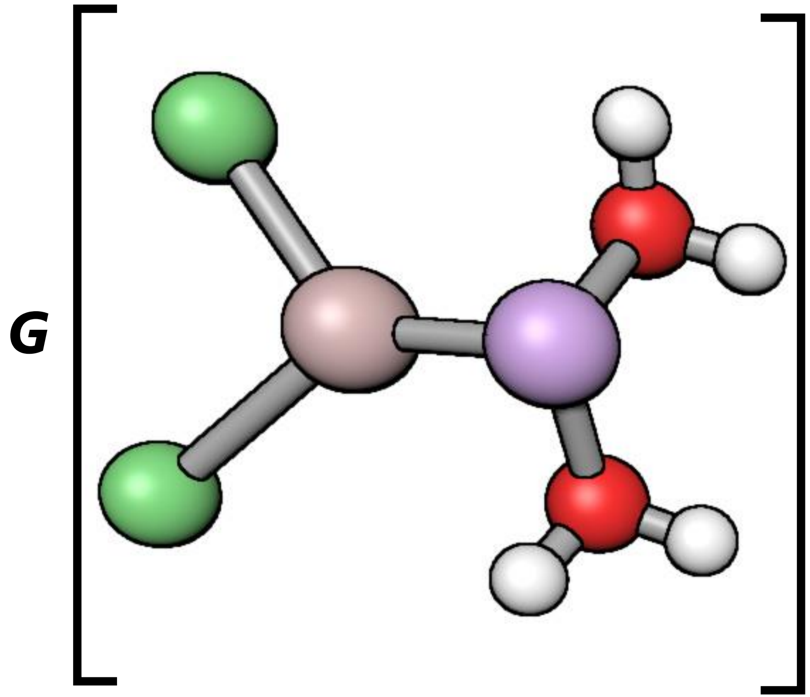
	Atomistic	Coarse-Grained	Compositional	Micro/Macro
Basis	Local/global configurational environment	Atomic fragments/motifs	Elemental Embeddings	Images / Micrographs
			 $A_aB_bC_c$...	
Input			Elemental/bulk properties	Micro/macro Properties
Implementations			One-Hot, Magpie Mat2Vec, Atom2Vec Matscholar Megnet (16-dim) Oliynyk CrystalLLM SkipAtom	Segmentation Tools pyMKS motif-learn AtomAI

Atomistic Descriptors

Local



Global



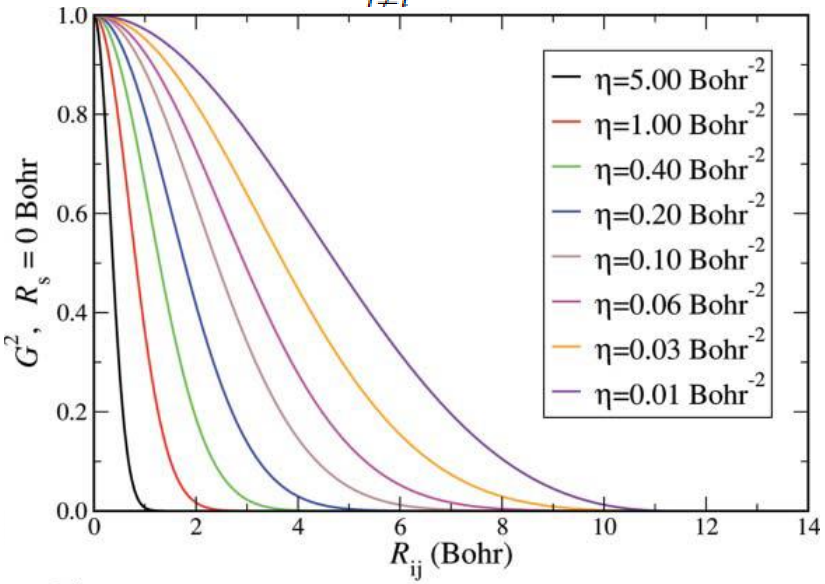
$$E_{\text{total}} = E_1(\mathbf{G}[A_1]) + E_2(\mathbf{G}[A_2]) + E_3(\mathbf{G}[A_3]) + E_4(\mathbf{G}[A_4]) + \ldots$$

$$E_{\text{total}} = E_{\text{total}}\,(\mathbf{G}[A_1, A_2, A_3, A_4, \ldots])$$

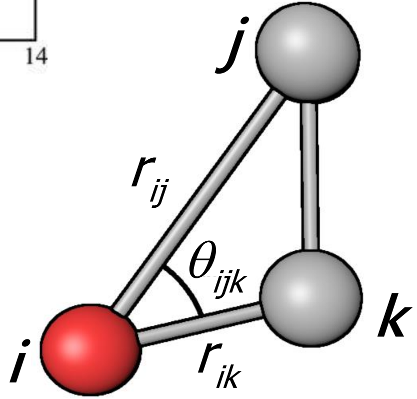
Atomistic Descriptors (Local)

Symmetry functions (Behler and Parinello)

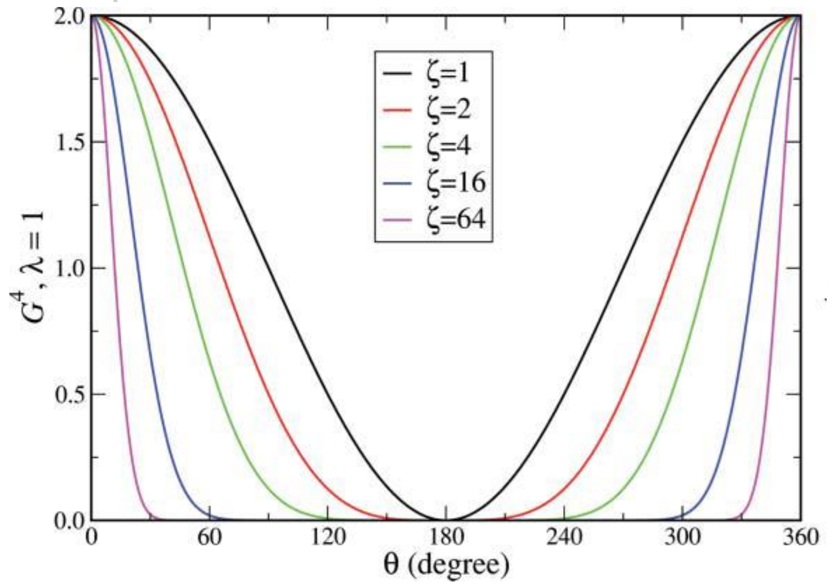
$$G_i^{\text{radial}} = \sum_{i\neq i} e^{-\eta(r_{ij}-r_s)^2} f_c(r_{ij})$$



$$E^{\text{pot}} = \sum_{j=1}^{N^a} E_j(\mathbf{G}_j)$$



$$G_i^{\text{ang.n.}} = 2^{1-\zeta} \sum_{j,k\neq i} (1+\lambda\cos\theta_{ijk})^{\zeta} e^{-\eta(r_{ij}^2+r_{ik}^2+r_{jk}^2)} f_c(r_{ij}) f_c(r_{ik}) f_c(r_{jk})$$

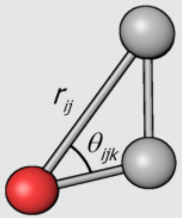
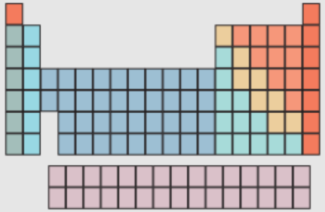
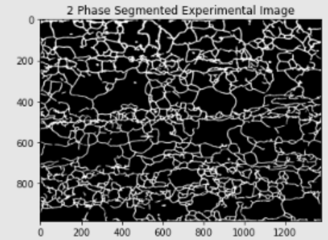


$$\vec{F}_i = -\vec{\nabla}_i E^{\text{pot}} = -\sum_{j=1}^{N^a} \vec{\nabla}_i E_j = -\sum_{j=1}^{N^a} \sum_{k=1}^{N_j^s} \frac{\partial E_j}{\partial G_{j,k}} \vec{\nabla}_i G_{j,k}$$

[1,2]

Hierarchy of Descriptors

Resolution, Accuracy, Computational Cost, Abstraction

Basis	**Atomistic** Local/global configurational environment		**Coarse-Grained** Atomic fragments/motifs	**Compositional** Elemental Embeddings	**Micro/Macro** Images / Micrographs
	Local	Global			
				 $A_aB_bC_c$...	
Input	Inter-atomic/structural properties 🖥			Elemental/bulk properties 🖥 🧪	Micro/macro Properties 🧪
Implementations	Atom-centered symmetry functions			One-Hot, Magpie Mat2Vec, Atom2Vec Matscholar Megnet (16-dim) Oliynyk CrystalLLM SkipAtom	Segmentation Tools pyMKS motif-learn AtomAI

Atomistic Descriptors (Global)

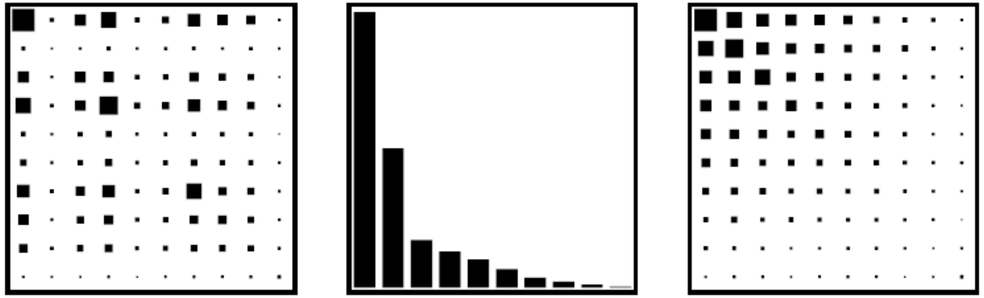
Coulomb matrix

$$M_{ij}^{\text{Coulomb}} = \begin{cases} 0.5 Z_i^{2.4} & \forall\, i = j \\ \frac{Z_i Z_j}{|\boldsymbol{R}_i - \boldsymbol{R}_j|} & \forall\, i \neq j \end{cases}$$

Issues 1: Number of atoms changes size of matrix

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

Issues 2: Order of indexing atoms (N! permutations)



[3-5]

Atomistic Descriptors (Global)

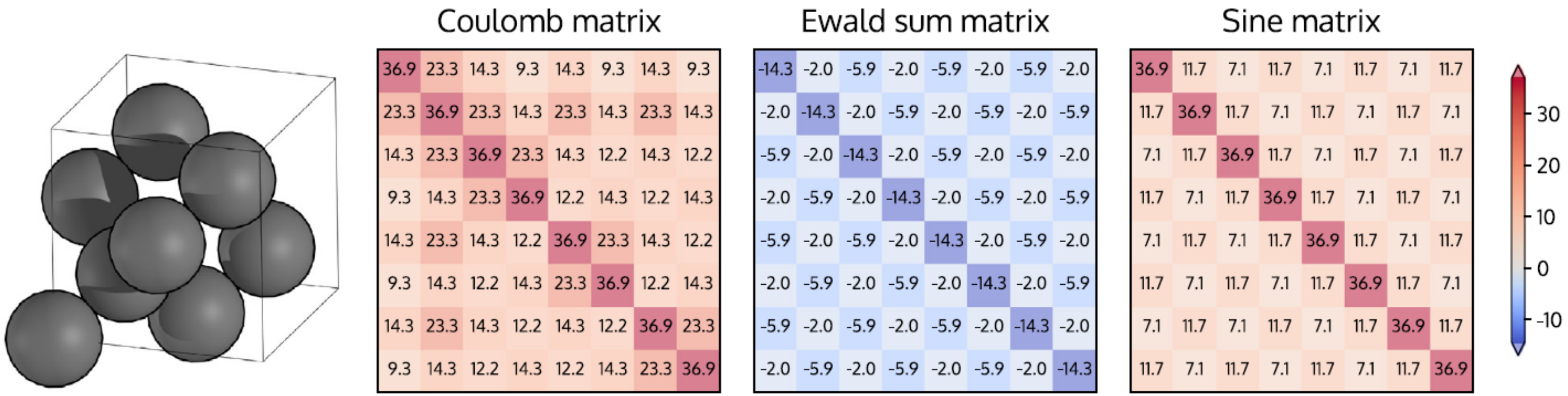
Ewald sum matrix

$$\phi_{ij} = \sum_{\mathbf{n}} \frac{Z_i Z_j}{\left|\boldsymbol{R}_i - \boldsymbol{R}_j\right| + \mathbf{n}} \qquad \mathbf{n} = h\mathbf{a} + k\mathbf{b} + l\mathbf{c}.$$

$$M_{ij}^{\text{Ewald}} = \begin{cases} \phi_{ij}^{\text{real}} + \phi_{ij}^{\text{recip}} + \phi_{ij}^{\text{self}} + \phi_{ij}^{\text{bg}} & \forall\ i = j \\ 2\left(\phi_{ij}^{\text{real}} + \phi_{ij}^{\text{recip}} + \phi_{ij}^{\text{bg}}\right) & \forall\ i \neq j \end{cases}$$

Sine matrix

$$\phi_{ij} = Z_i Z_j \left|\mathbf{B} \cdot \sum_{k=\{x,y,z\}} \hat{\mathbf{e}}_k \sin^2\left(\pi \mathbf{B}^{-1} \cdot \left(\boldsymbol{R}_i - \boldsymbol{R}_j\right)\right)\right|^{-1}$$



Coulomb matrix

36.9	23.3	14.3	9.3	14.3	9.3	14.3	9.3
23.3	36.9	23.3	14.3	23.3	14.3	23.3	14.3
14.3	23.3	36.9	23.3	14.3	12.2	14.3	12.2
9.3	14.3	23.3	36.9	12.2	14.3	12.2	14.3
14.3	23.3	14.3	12.2	36.9	23.3	14.3	12.2
9.3	14.3	12.2	14.3	23.3	36.9	12.2	14.3
14.3	23.3	14.3	12.2	14.3	12.2	36.9	23.3
9.3	14.3	12.2	14.3	12.2	14.3	23.3	36.9

Ewald sum matrix

-14.3	-2.0	-5.9	-2.0	-5.9	-2.0	-5.9	-2.0
-2.0	-14.3	-2.0	-5.9	-2.0	-5.9	-2.0	-5.9
-5.9	-2.0	-14.3	-2.0	-5.9	-2.0	-5.9	-2.0
-2.0	-5.9	-2.0	-14.3	-2.0	-5.9	-2.0	-5.9
-5.9	-2.0	-5.9	-2.0	-14.3	-2.0	-5.9	-2.0
-2.0	-5.9	-2.0	-5.9	-2.0	-14.3	-2.0	-5.9
-5.9	-2.0	-5.9	-2.0	-5.9	-2.0	-14.3	-2.0
-2.0	-5.9	-2.0	-5.9	-2.0	-5.9	-2.0	-14.3

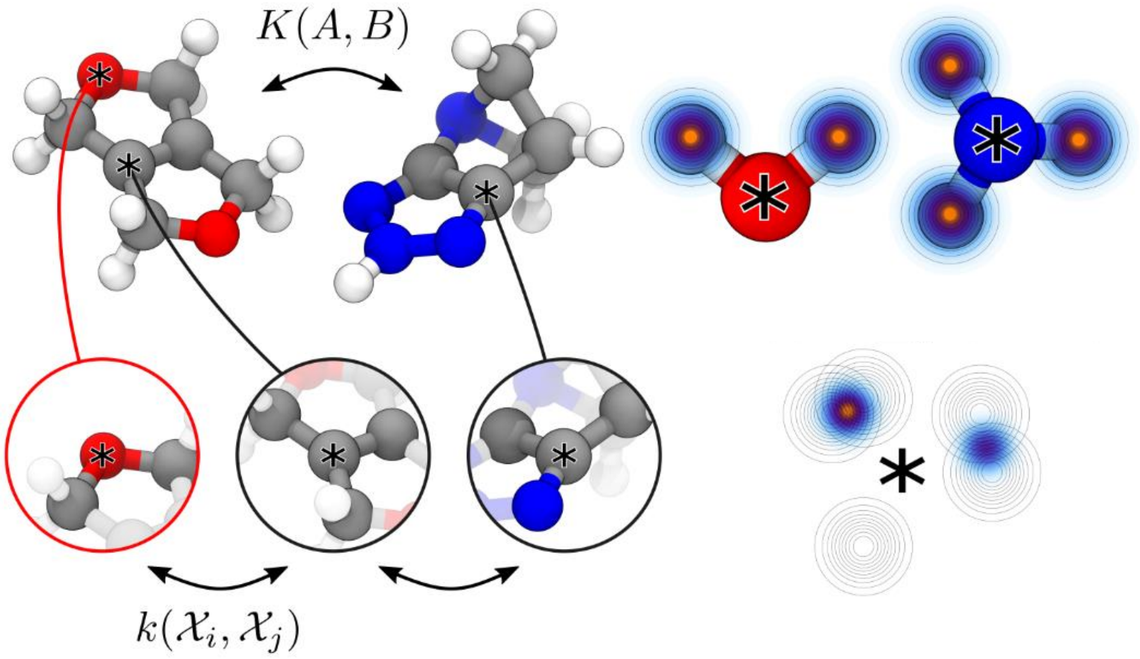
Sine matrix

36.9	11.7	7.1	11.7	7.1	11.7	7.1	11.7
11.7	36.9	11.7	7.1	11.7	7.1	11.7	7.1
7.1	11.7	36.9	11.7	7.1	11.7	7.1	11.7
11.7	7.1	11.7	36.9	11.7	7.1	11.7	7.1
7.1	11.7	7.1	11.7	36.9	11.7	7.1	11.7
11.7	7.1	11.7	7.1	11.7	36.9	11.7	7.1
7.1	11.7	7.1	11.7	7.1	11.7	36.9	11.7
11.7	7.1	11.7	7.1	11.7	7.1	11.7	36.9

[3]

Atomistic Descriptors (Global)

Smooth Overlap of Atomic Positions (SOAP) Kernel



$$\rho(\mathbf{r}) = \sum_i \mathrm{e}^{-\alpha|\mathbf{r}-\mathbf{r}_i|^2}.$$

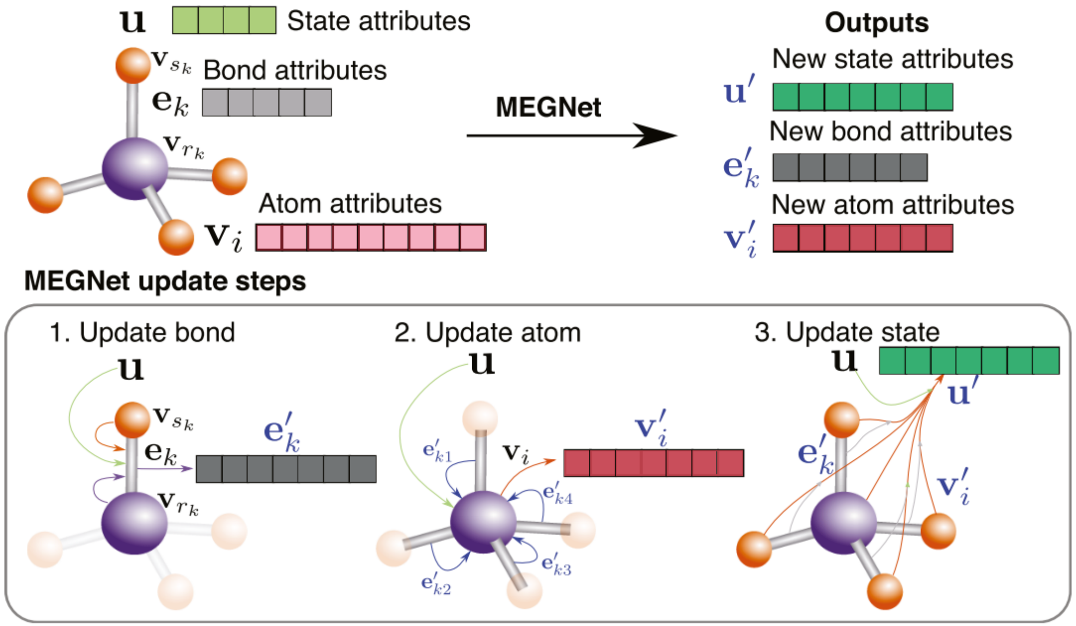
$$k(\rho, \rho\prime) = \int d\hat{R} \int d\mathbf{r}\, \rho(\mathbf{r})\rho\prime(\hat{R}\mathbf{r}).$$

Generalization of symmetry functions:
Capable of characterizing entire atomic environment at once

[6]

Atomistic Descriptors (Global)

Graph representations



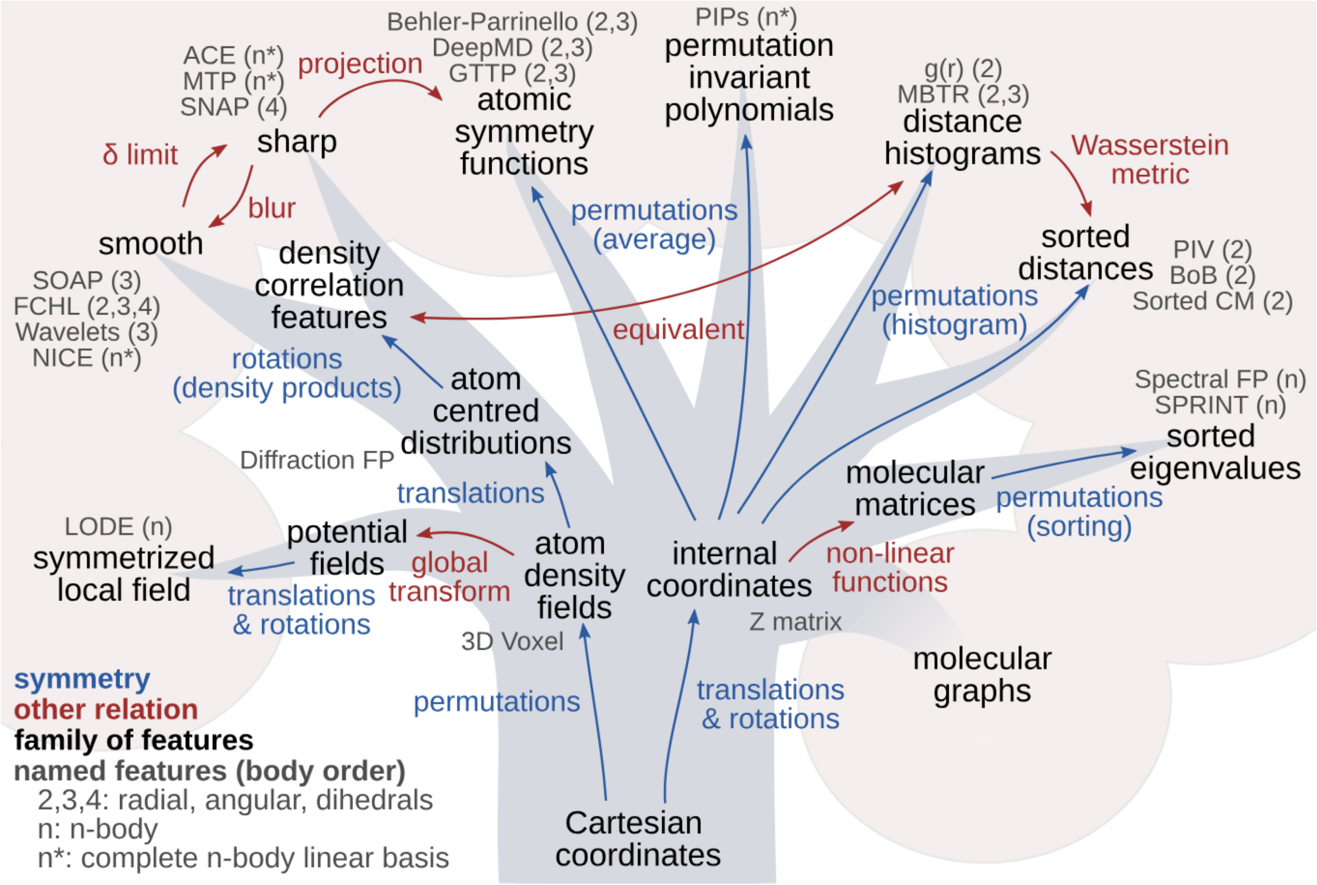
CGCNN, MEGNet, SchNet, PointNet, …

Others global descriptors

Many-Body Tensor Representation (MBTR), Voronoi Tesselation

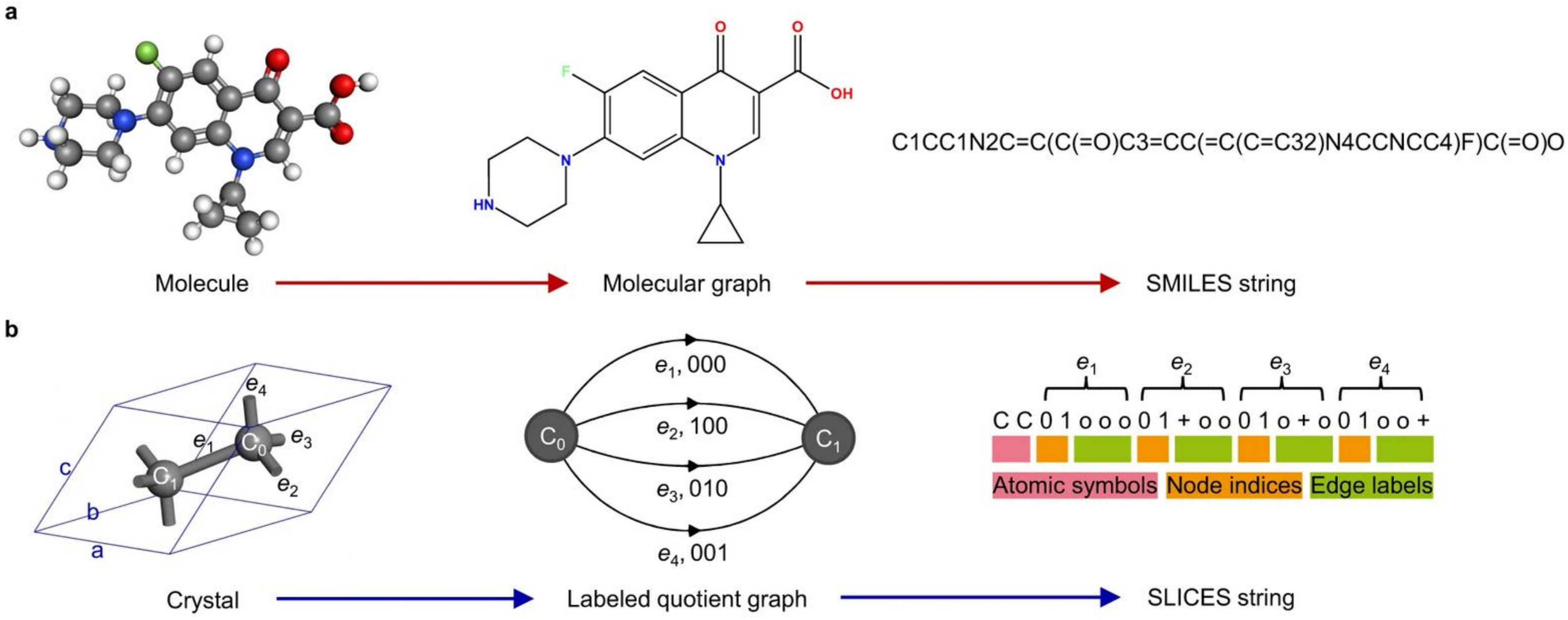
[3,7-10]

Overview: Atomistic Descriptors



Text-Based Global Descriptors

Simplified Molecular-Input Line-Entry System (SMILES)



Simplified Line-Input Crystal-Encoding System (SLICES)

[16]

Hierarchy of Descriptors

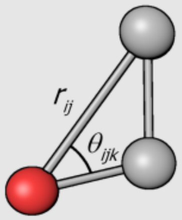
Resolution, Accuracy, Computational Cost, Abstraction

Basis

Atomistic

Local/global configurational environment

Local

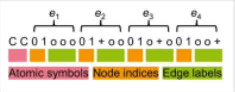


Global

Numeric

-14.3	-2.0	-5.9	-2.0	-5.9	-2.0	-5.9	-2.0
-2.0	-14.3	-2.0	-5.9	-2.0	-5.9	-2.0	-5.9
-5.9	-2.0	-14.3	-2.0	-5.9	-2.0	-5.9	-2.0
-2.0	-5.9	-2.0	-14.3	-2.0	-5.9	-2.0	-5.9
-5.9	-2.0	-5.9	-2.0	-14.3	-2.0	-5.9	-2.0
-2.0	-5.9	-2.0	-5.9	-2.0	-14.3	-2.0	-5.9
-5.9	-2.0	-5.9	-2.0	-5.9	-2.0	-14.3	-2.0
-2.0	-5.9	-2.0	-5.9	-2.0	-5.9	-2.0	-14.3

Text



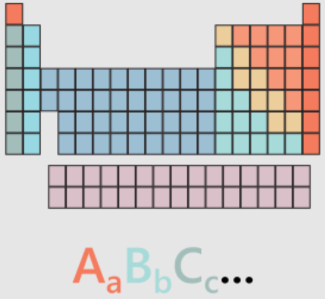
Graphs

Coarse-Grained

Atomic fragments/motifs

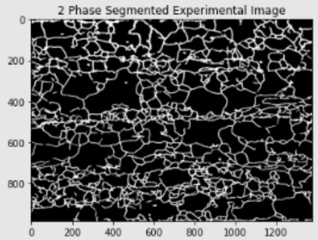
Compositional

Elemental Embeddings



Micro/Macro

Images / Micrographs



Input

- Atomistic: Inter-atomic/structural properties
- Compositional: Elemental/bulk properties
- Micro/Macro: Micro/macro Properties

Implementations

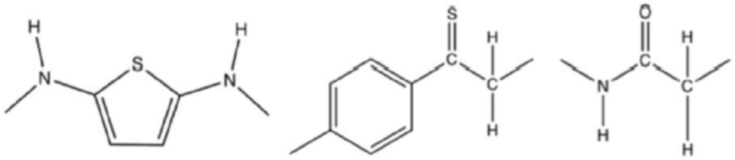
- Local: Atom-centered symmetry functions
- Numeric: SOAPs; Coulomb, Ewald, and Sine matrices; MBTR; Voronoi tesselation
- Text: SMILES (Molecule); SLICES (Crystal)
- Graphs: CGCNN, ALIGNN, MEGNet, M3GNet, SEGNN, E3NN, SchNet, PointNet, Equiformer-v2
- Compositional: One-Hot, Magpie, Mat2Vec, Atom2Vec, Matscholar, Megnet (16-dim), Oliynyk, CrystalLLM, SkipAtom
- Micro/Macro: Segmentation Tools, pyMKS, motif-learn, AtomAI

Coarse-Grained Descriptors

Fragment/Simplex/Motif fingerprints

For polymers, basic 7 units: CH_2, CO, CS, O, NH, C_6H_4, C_4H_2S

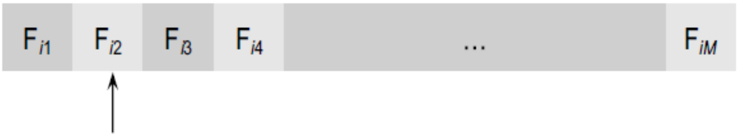
Typical organic fragment types



Pairs: 7x7

Triplets: 7x7x7

Fingerprint of polymer *i*

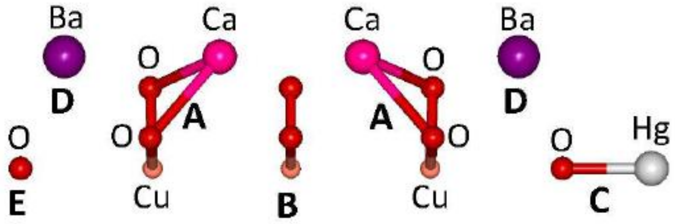


Number fraction of fragment type 2

→ KRR

Also used for crystals:

- Binning compositions by crystal structure prototypes
- Bounded/Unbounded simplexes

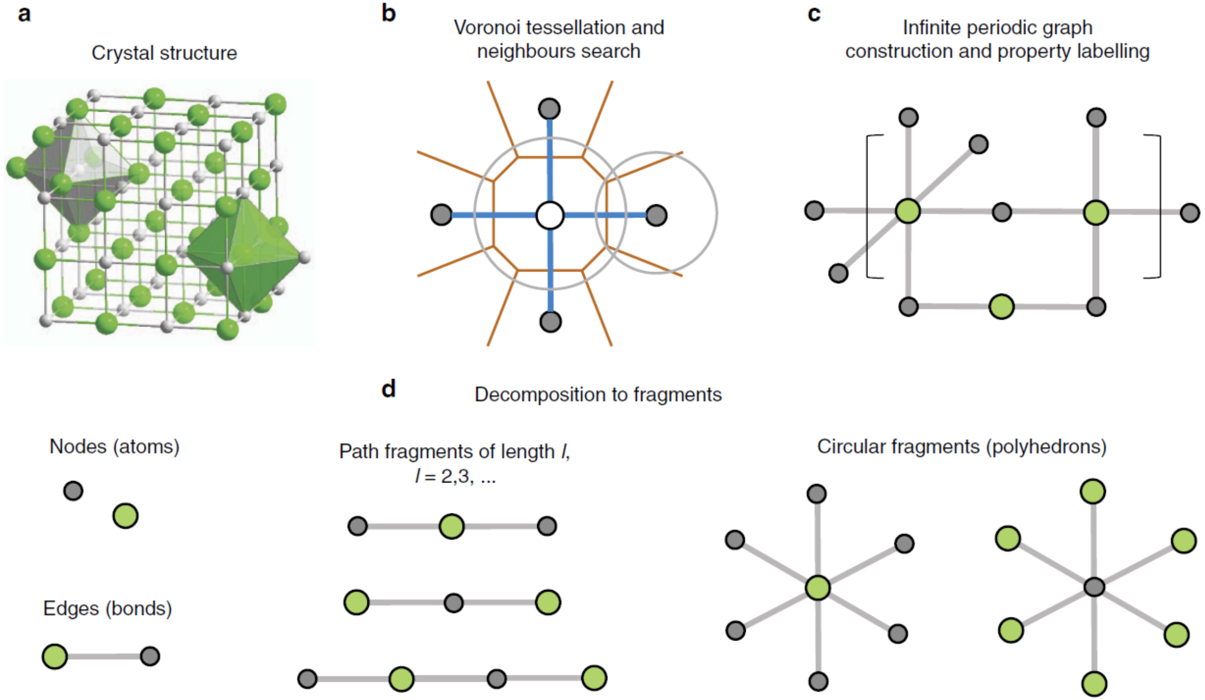


$Ba_2Ca_2Cu_3HgO_8$ = $\mathbf{A_2BCD_2E}$

[11,12]

Coarse-Grained Descriptors

Property-labeled materials fragments

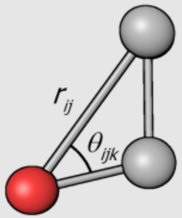
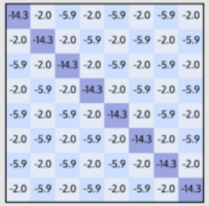
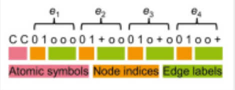
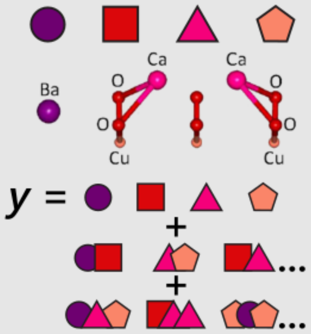
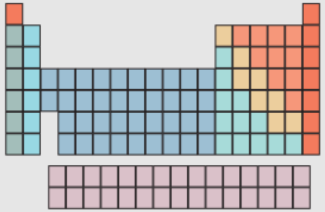
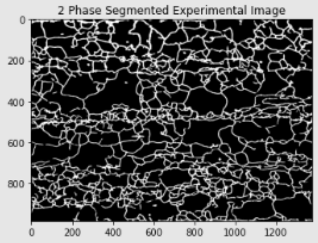

a
Crystal structure
b
Voronoi tessellation and neighbours search
c
Infinite periodic graph construction and property labelling
d
Decomposition to fragments
Nodes (atoms)
Edges (bonds)
Path fragments of length *l*, *l* = 2,3, ...
Circular fragments (polyhedrons)

$$T = \sum_{i,j} \left| q_i - q_j \right| M_{ij},$$

[13]

Hierarchy of Descriptors

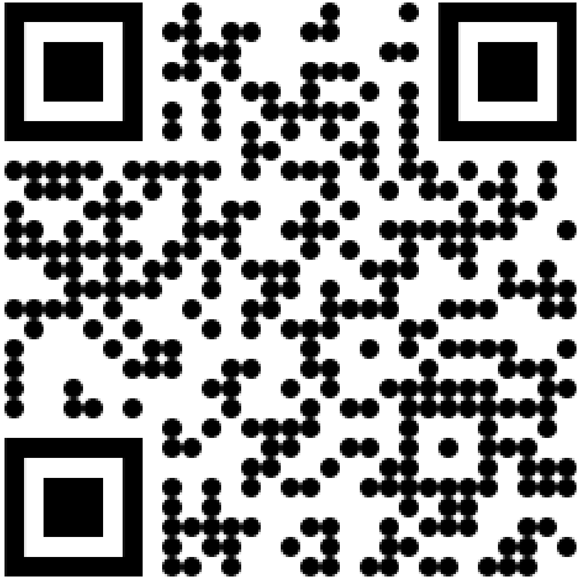
Resolution, Accuracy, Computational Cost, Abstraction

Basis	**Atomistic** Local/global configurational environment				**Coarse-Grained** Atomic fragments/motifs	**Compositional** Elemental Embeddings	**Micro/Macro** Images / Micrographs
	Local	**Global**					
		**Numeric** $\begin{pmatrix} -14.3 & -2.0 & -5.9 & -2.0 & -5.9 & -2.0 & -5.9 & -2.0 \\ -2.0 & -14.3 & -2.0 & -5.9 & -2.0 & -5.9 & -2.0 & -5.9 \\ -5.9 & -2.0 & -14.3 & -2.0 & -5.9 & -2.0 & -5.9 & -2.0 \\ -2.0 & -5.9 & -2.0 & -14.3 & -2.0 & -5.9 & -2.0 & -5.9 \\ -5.9 & -2.0 & -5.9 & -2.0 & -14.3 & -2.0 & -5.9 & -2.0 \\ -2.0 & -5.9 & -2.0 & -5.9 & -2.0 & -14.3 & -2.0 & -5.9 \\ -5.9 & -2.0 & -5.9 & -2.0 & -5.9 & -2.0 & -14.3 & -2.0 \\ -2.0 & -5.9 & -2.0 & -5.9 & -2.0 & -5.9 & -2.0 & -14.3 \end{pmatrix}$	**Text**  e_1 e_2 e_3 e_4 C C 0 1 o o o 0 1 + o o 0 1 o + o 0 1 o o + Atomic symbols Node indices Edge labels	**Graphs** 	 Ca Ca O O Ba O O Cu Cu y = + + ...	$A_aB_bC_c$...	 2 Phase Segmented Experimental Image 0 200 400 600 800 0 200 400 600 800 1000 1200
Input	Inter-atomic/structural properties 🖥				Coarse-grained properties 🖥 🧪	Elemental/bulk properties 🖥 🧪	Micro/macro Properties 🧪
Implementations	Atom-centered symmetry functions	SOAPs Coulomb, Ewald, and Sine matrices MBTR Voronoi tesselation	SMILES (Molecule) SLICES (Crystal)	CGCNN ALIGNN MEGNet M3GNet SEGNN, E3NN SchNet PointNet Equiformer-v2	Fragment/Simplex/Motif Fingerprints Property-Labeled Materials Fragments	One-Hot, Magpie Mat2Vec, Atom2Vec Matscholar Megnet (16-dim) Oliynyk CrystalLLM SkipAtom	Segmentation Tools pyMKS motif-learn AtomAI

DScribe

ElementEmbeddings

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!**