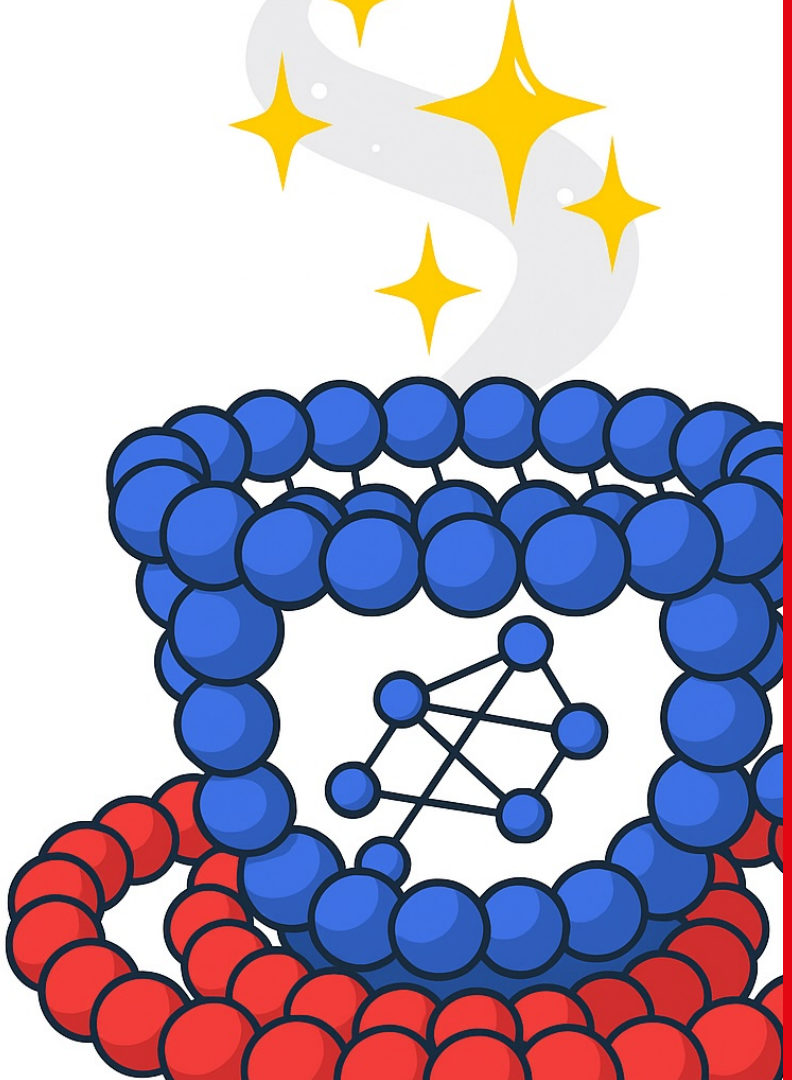


Representations for Atomistic Machine Learning

Philip Loche

Stuttgart, 2025/10/07

What is on the menu

1. Molecular descriptors (cheminformatics)
2. Atomistic descriptors (condensed matter)
3. Descriptor based neural network potentials
4. End-to-end neural network potentials

Deliver the vibe of atomistic machine learning!

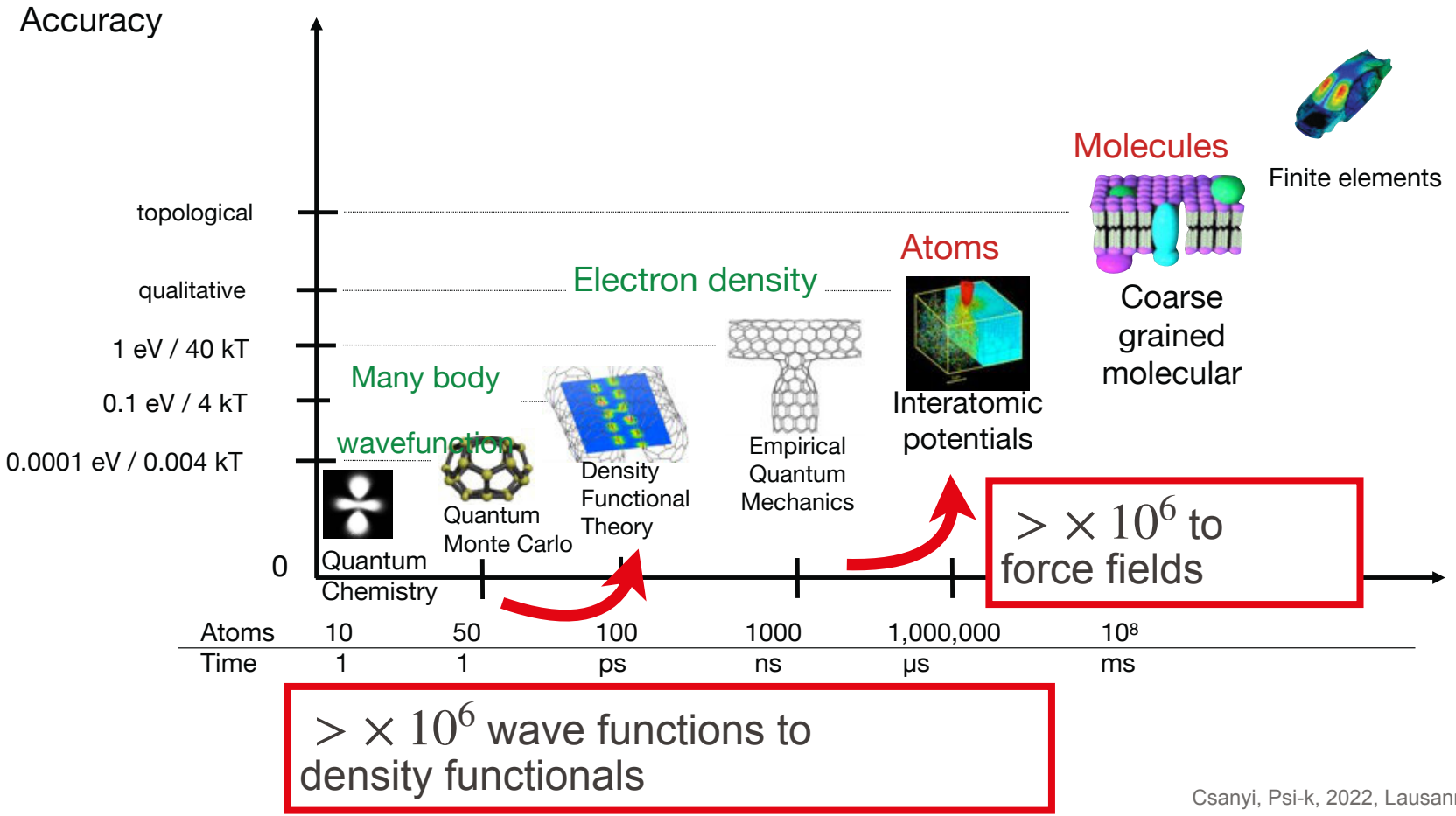
If you want to dig deeper…

- Musil et. al., “Physics-Inspired Structural Representations for Molecules and Materials”, 10.1021acs.chemrev.1c00021.
- Unke, O. T. et al. “Machine Learning Force Fields”, 10.1021/acs.chemrev.0c01111
- Andrew D. White, Deep learning for molecules & materials, https://dmol.pub/

… and try it on your own

- https://github.com/ceriottm/aml-intro
- https://github.com/HFooladi/GNNs-For-Chemists

Atomistic modeling encompasses the microscopic domain



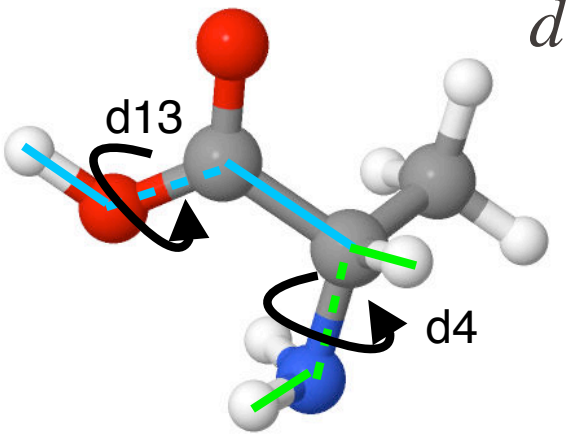
DFT: backbone of computational chemistry

- Quantum mechanics are based on the **Shrödinger's equation** $H\Psi = E\Psi$
- **Very complex equation** and concepts reduced complexity (and accuracy).
  1. Born-Oppenheimer: Total molecular wavefunction as product of an **electronic and a nuclear wavefunction**.
  2. Kohn-Scham: Base **electron energy** functional $E[\,\cdot\,]$ on **density** $\rho(\boldsymbol{r})$ instead of the wave function Ψ.

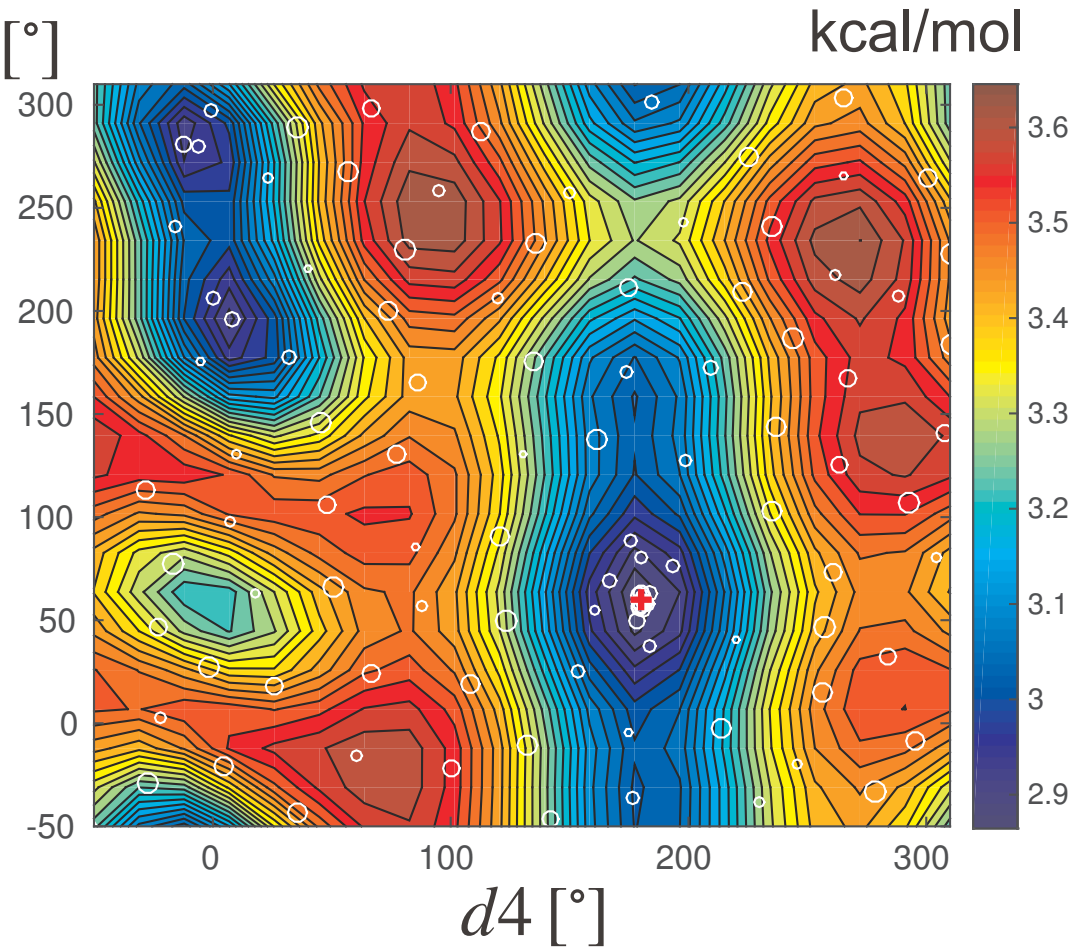
DFT is in principle an exact **ground state theory**.

The **energy functional** needs to be **approximated** in practice.

Potential Energy surface (PES)

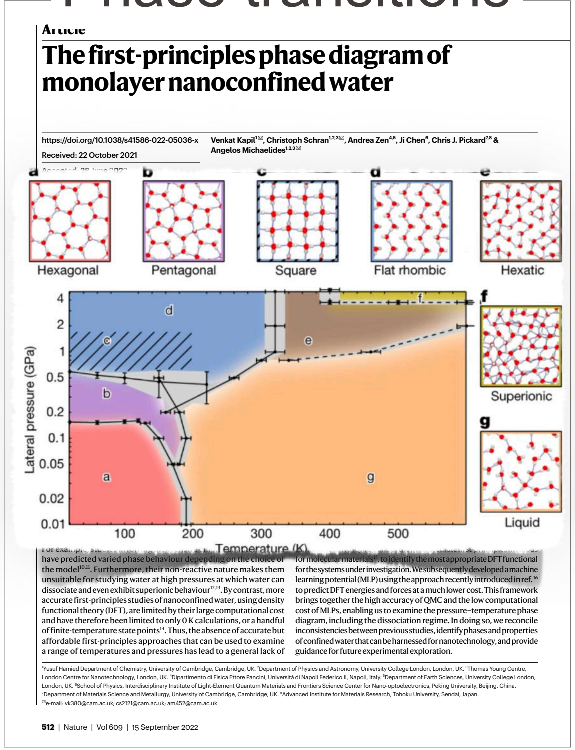


- Total energy in terms of atomic positions
- The minima are the stable isomers/ conformers



DFT: backbone of computational chemistry

Phase transitions


Article

The first-principles phase diagram of monolayer nanoconfined water

https://doi.org/10.1038/s41586-022-05036-x

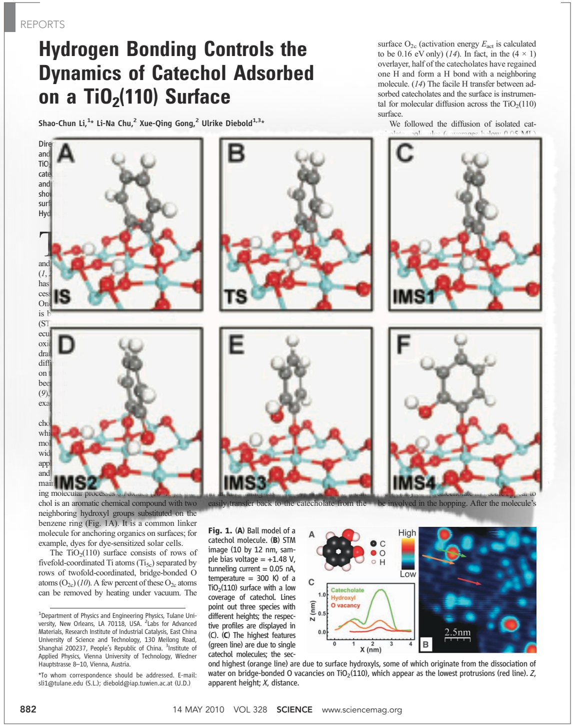
Received: 22 October 2021

Venkat Kapil[1], Christoph Schran[1,2,3], Andrea Zen[4,5], Ji Chen[6], Chris J. Pickard[7,8] & Angelos Michaelides[1,2,3]

have predicted varied phase behaviour depending on the choice of the model[10,11]. Furthermore, their non-reactive nature makes them unsuitable for studying water at high pressures at which water can dissociate and even exhibit superionic behaviour[12,13]. By contrast, more accurate first-principles studies of nanoconfined water, using density functional theory (DFT), are limited by their large computational cost and have therefore been limited to only 0 K calculations, or a handful of finite-temperature state points[14]. Thus, the absence of accurate but affordable first-principles approaches that can be used to examine a range of temperatures and pressures has lead to a general lack of

for molecular materials[15], to identify the most appropriate DFT functional for the systems under investigation. We subsequently developed a machine learning potential (MLP) using the approach recently introduced in ref.[16] to predict DFT energies and forces at a much lower cost. This framework brings together the high accuracy of QMC and the low computational cost of MLPs, enabling us to examine the pressure–temperature phase diagram, including the dissociation regime. In doing so, we reconcile inconsistencies between previous studies, identify phases and properties of confined water that can be harnessed for nanotechnology, and provide guidance for future experimental exploration.

Chemical reactions

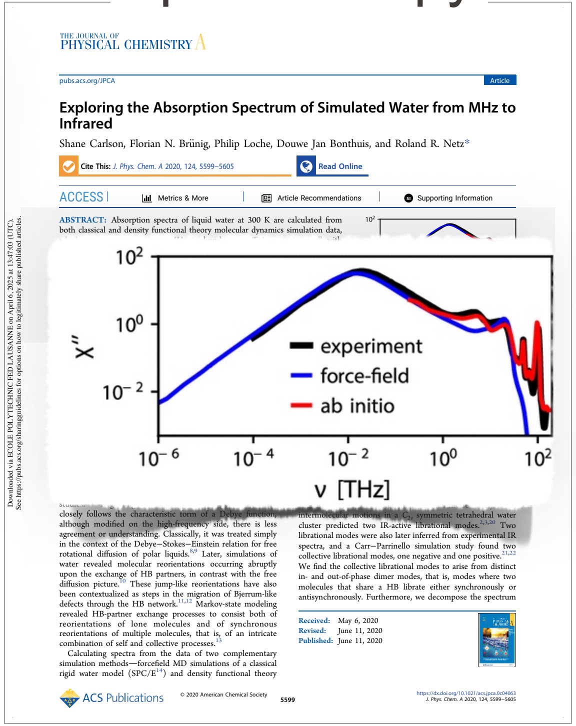

REPORTS

Hydrogen Bonding Controls the Dynamics of Catechol Adsorbed on a $TiO_2(110)$ Surface

Shao-Chun Li,[1]* Li-Na Chu,[2] Xue-Qing Gong,[2] Ulrike Diebold[1,3]*

Fig. 1. (A) Ball model of a catechol molecule. (B) STM image (10 by 12 nm, sample bias voltage = +1.48 V, tunneling current = 0.05 nA, temperature = 300 K) of a low coverage of catechol. Lines point out three species with different heights; the respective profiles are displayed in (C). (C) The highest features (green line) are due to single catechol molecules; the second highest (orange line) are due to surface hydroxyls, some of which originate from the dissociation of water on bridge-bonded O vacancies on $TiO_2(110)$, which appear as the lowest protrusions (red line). Z, apparent height; X, distance.

Spectroscopy


THE JOURNAL OF PHYSICAL CHEMISTRY A
pubs.acs.org/JPCA Article

Exploring the Absorption Spectrum of Simulated Water from MHz to Infrared

Shane Carlson, Florian N. Brünig, Philip Loche, Douwe Jan Bonthuis, and Roland R. Netz*

Cite This: J. Phys. Chem. A 2020, 124, 5599−5605

Read Online

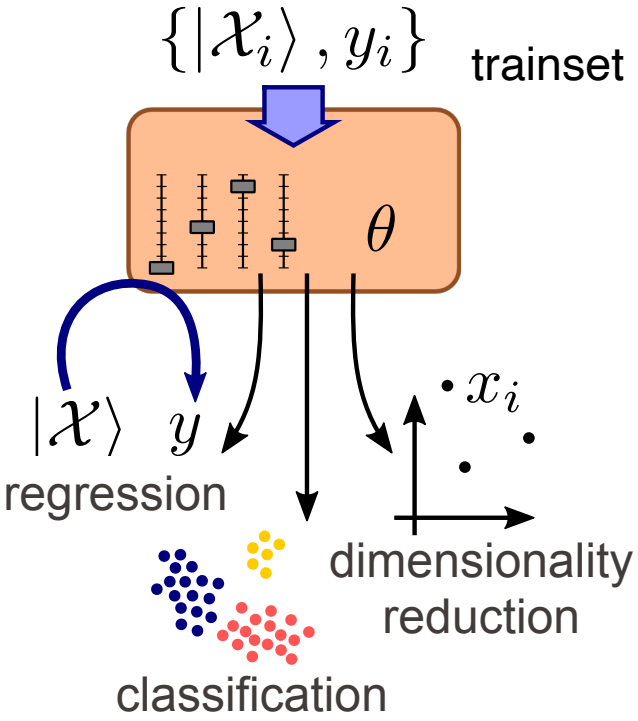
ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: Absorption spectra of liquid water at 300 K are calculated from both classical and density functional theory molecular dynamics simulation data,

Received: May 6, 2020
Revised: June 11, 2020
Published: June 11, 2020

Machine learning at the atomic scale

Chemical structures (inputs) and their properties (labels) are fed to a learning scheme, tuned by hyper parameters θ, that can then be used to perform different tasks on new data.



What is y_i (target)?

- per-structure property: energy
- per molecule property (dipole moment)
- per atom property (force)

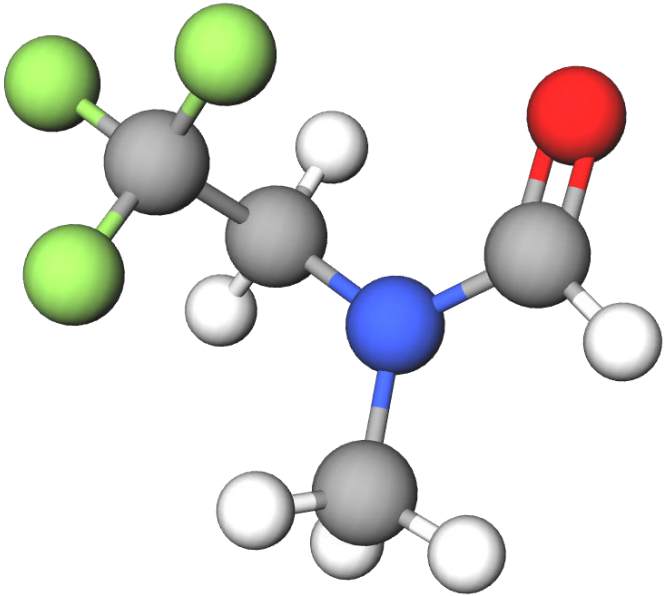
What is X_i (features)?

- also called descriptor, fingerprint or representation
- *How to do good feature engineering?*

Data structure

Data is usually of type:

molecule1, energy1
molecule2, energy2
…



Molecules are then specified in xyz coordinates:

molecule1:
atom1, x1, y1, z1
atom2, x2, y2, z2
….

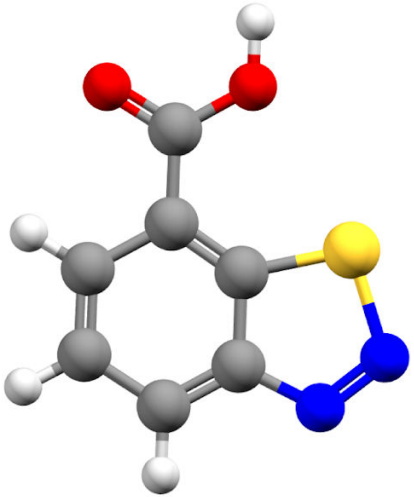
Additionally, properties (force, dipoles, etc.) might be listed.

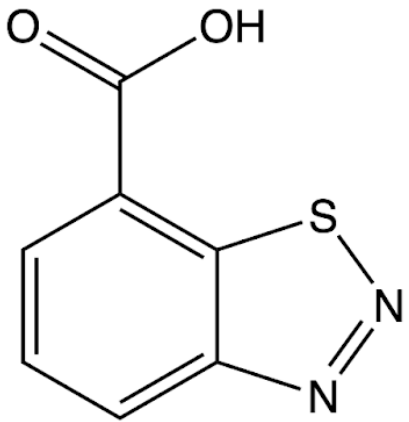
Molecular descriptors (cheminformatics)

Molecular representations

How would you describe this molecule to a machine?

Think about what molecular information could be extracted and how it could be encoded.





Acibenzolar acid
(benzo[d][1,2,3]thiadiazole-7-carboxylic acid)

Descriptors (representation)

A machine-compatible representation of a material, state, or problem.

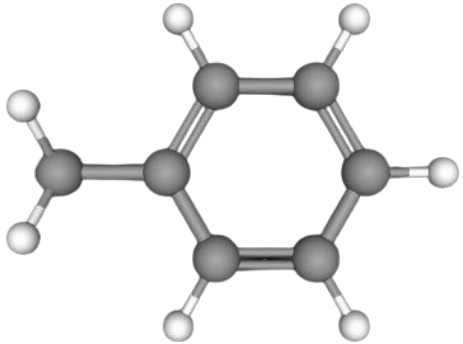
Attribute based

Encodes descriptive information about the problem.

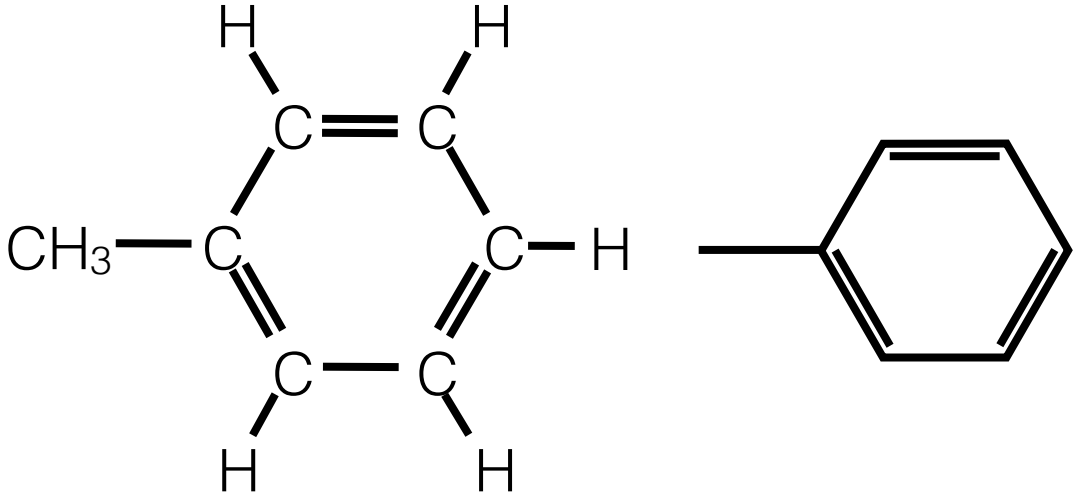
Structure based

Encodes physical or chemical structure directly.

Molecules as graphs



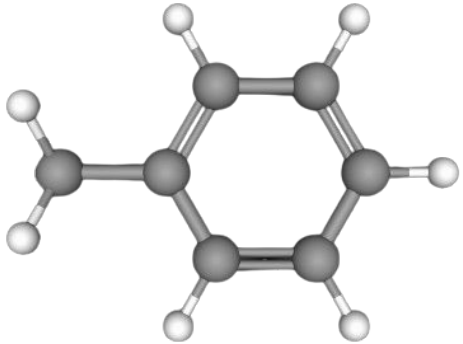
toluene

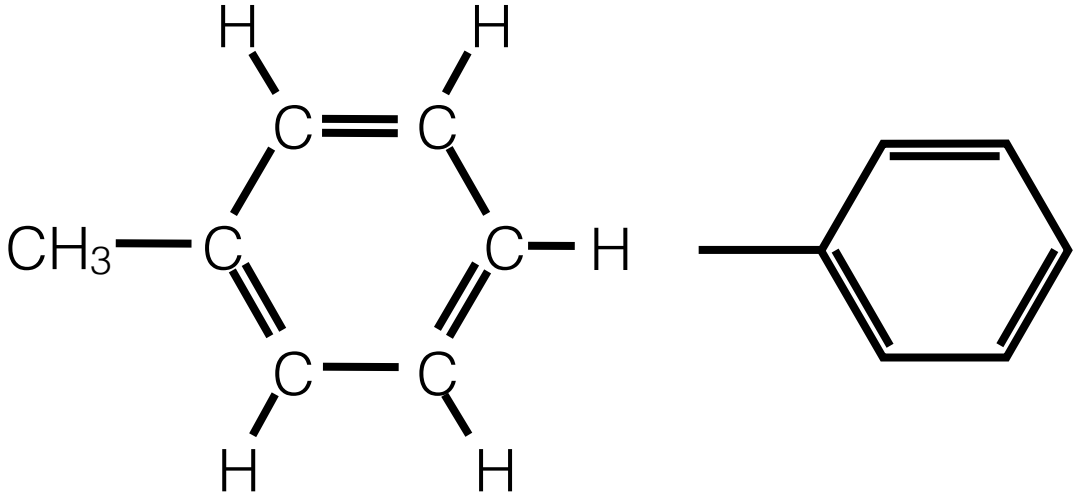


Graph representation of molecules (also called 2D)
Chemists are used to it

But we need something easier for the machines!

Molecules as graphs





toluene

C1=C(C)C=CC=C1

Simplified molecular-input line-entry system (*SMILES*)
text-based molecular encoding

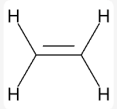
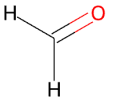
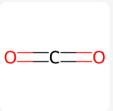
Some SMILES rules

Atoms:

- as atomic symbol: [C]
- implicit fulfilled valency: C -> [CH4]

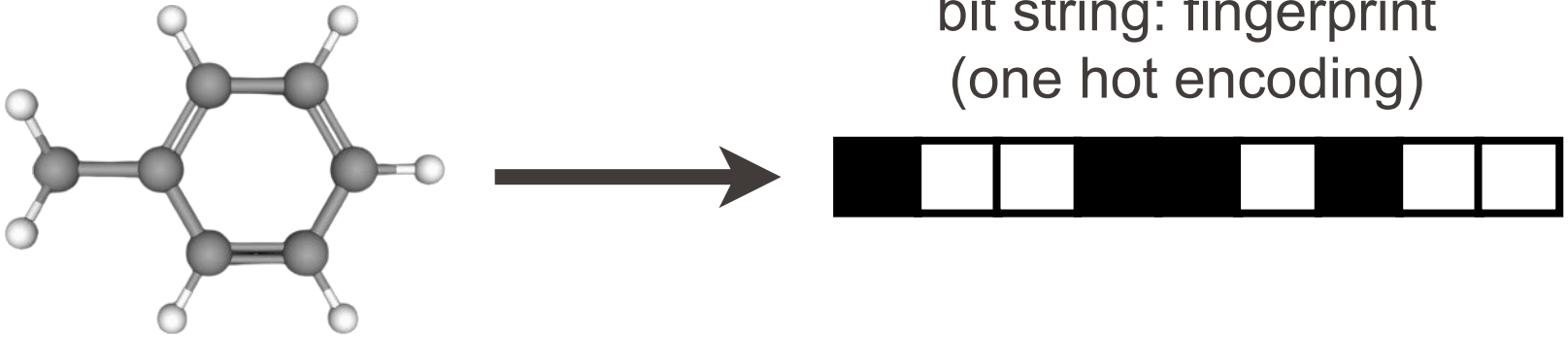
Bonds:

- single: -
- double: =
- triple: #
- Many more rules for branching, rings, charges…

Molecule	SMILES	Alternate SMILES	Structure
Ethene	C=C	[CH2][CH2]	
Formaldehyde	C=O	[CH2]=O	
Carbon Dioxide	O=C=O	C(=O)=O	

- SMILES tools and generation: CANGEN algorithm, RDkit

Fingerprints: molecules as bit strings



Fingerprinting strategy:

- identify and extract molecular features
- encode them as 1 (present) or 0 (not present)
- or, hash the features
  (= convert them to a number in a unique way)

Fingerprints: molecules as bit strings

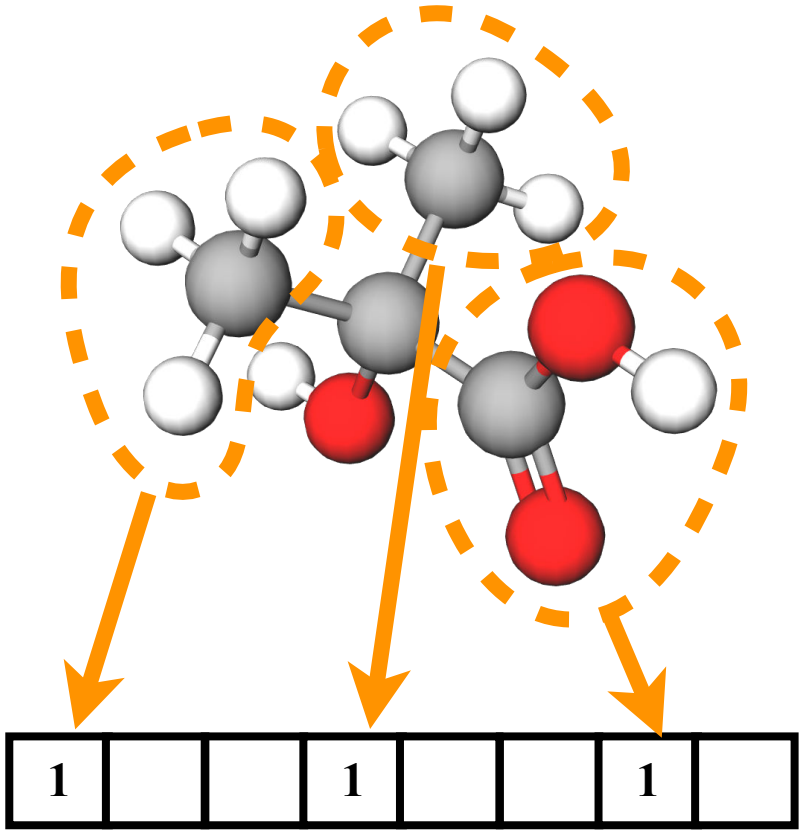
Also called dictionary fingerprints

Explicitly define features
(typically with SMILES Arbitrary Target Specification (SMARTS))

Check if features are present

Examples:

MACCS:	166 features
PubChem:	881 features
Klekota-Roth:	4860 features



Substructure fingerprints - advantages

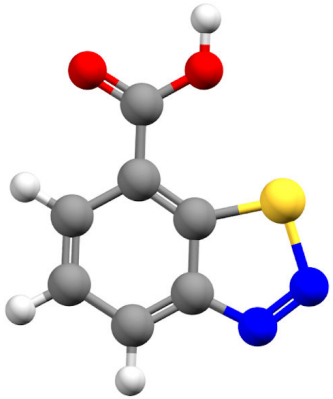
1. **Simple:**
   easy to implement and use
2. **Compact:**
   very small fingerprints, typically 2 to 8k entries
3. **Sasy similarity metrics:**
   great for similarity searches
4. **Interpretability**
   features encode chemical information directly

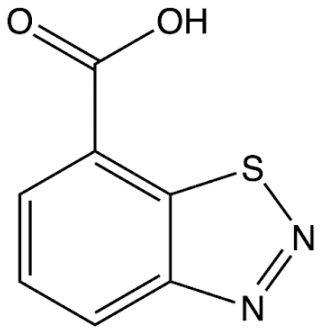
Substructure fingerprints - disadvantages

1. **Poor regression accuracy:**
   bond lengths and bond angles are not encoded
2. **Lack of transferability:**
   fixed feature selection might not translate to new domains
3. **Lack of resolution:**
   certain molecules might have the same fingerprint due to limited feature selection
4. **Not conformer sensitive:**
   since bond lengths and angle information is absent

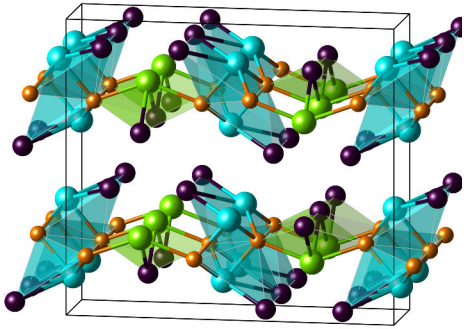
Atomistic descriptors (condensed matter)

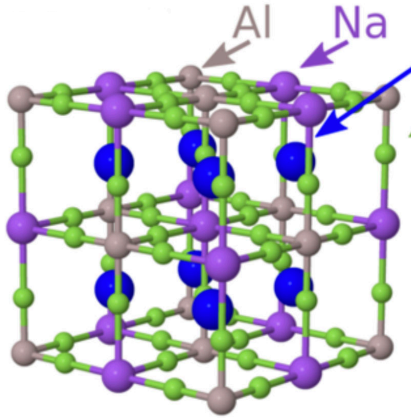
Representation for solids

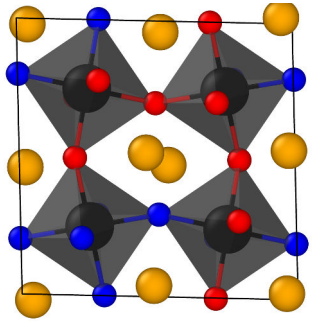




1. Speculate why molecular descriptors are not used to for solids.
2. What could a descriptor for solids look like;
   what should it include?





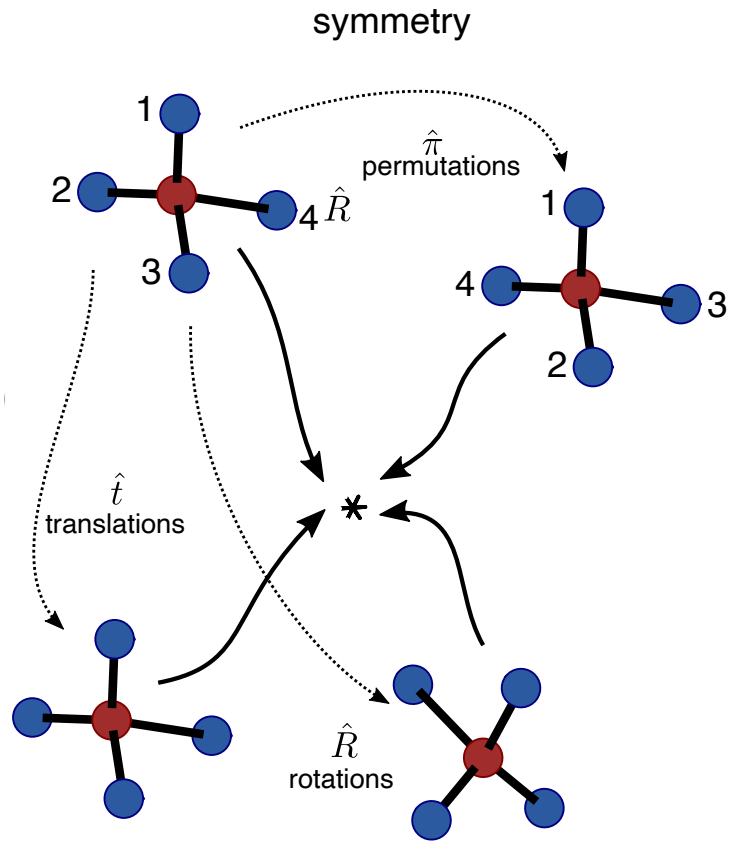


Descriptor requirements

1. Invariant with respect to spatial **translations** of the coordinate system: isometry of space

2. Invariant with respect to **rotations** of the coordinate system: isotropy of space.

3. Invariant with respect to **permutations** (e.g., elements in chemical formula, atomic indices)

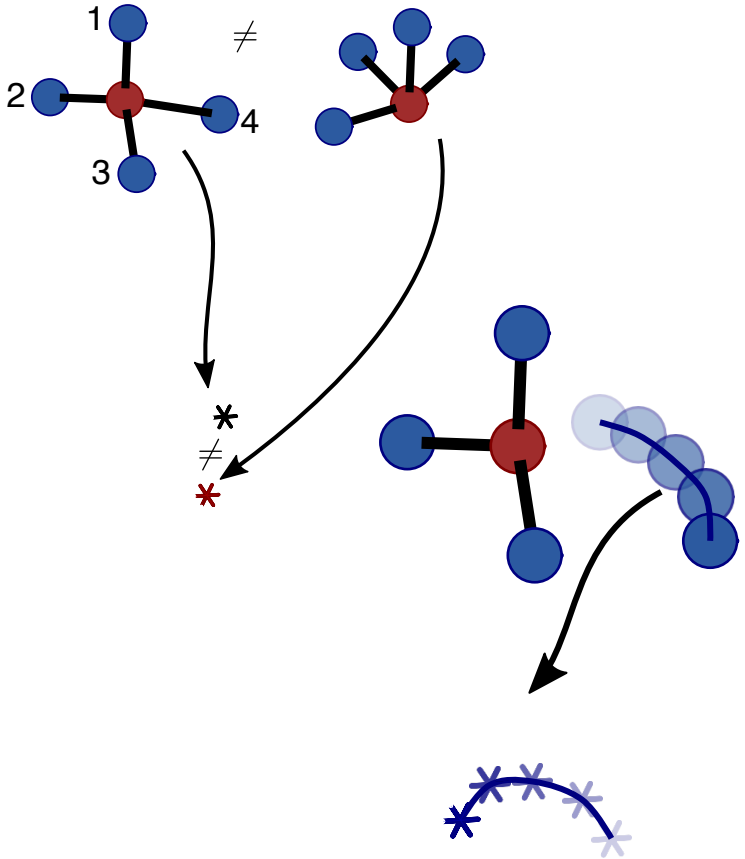


Descriptor requirements

4. Completeness/**Unique**: there is a single way to construct a descriptor for the material, problem or state

5. **Continuous**: small changes in the state should translate to small changes in the descriptor

6. Compact and **computationally efficient**

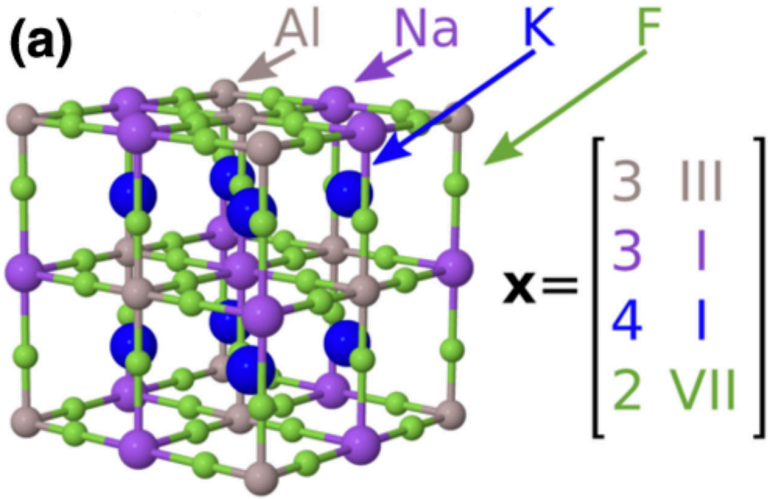


Attribute based descriptors

- Easy example: fixed crystal structure
- Elpasolite crystal structure (ABC_2D_6) is very common
- $2 \cdot 10^6$ possible combinations with only main group elements
- Cohesive energy for ~10'000 materials calculated with DFT

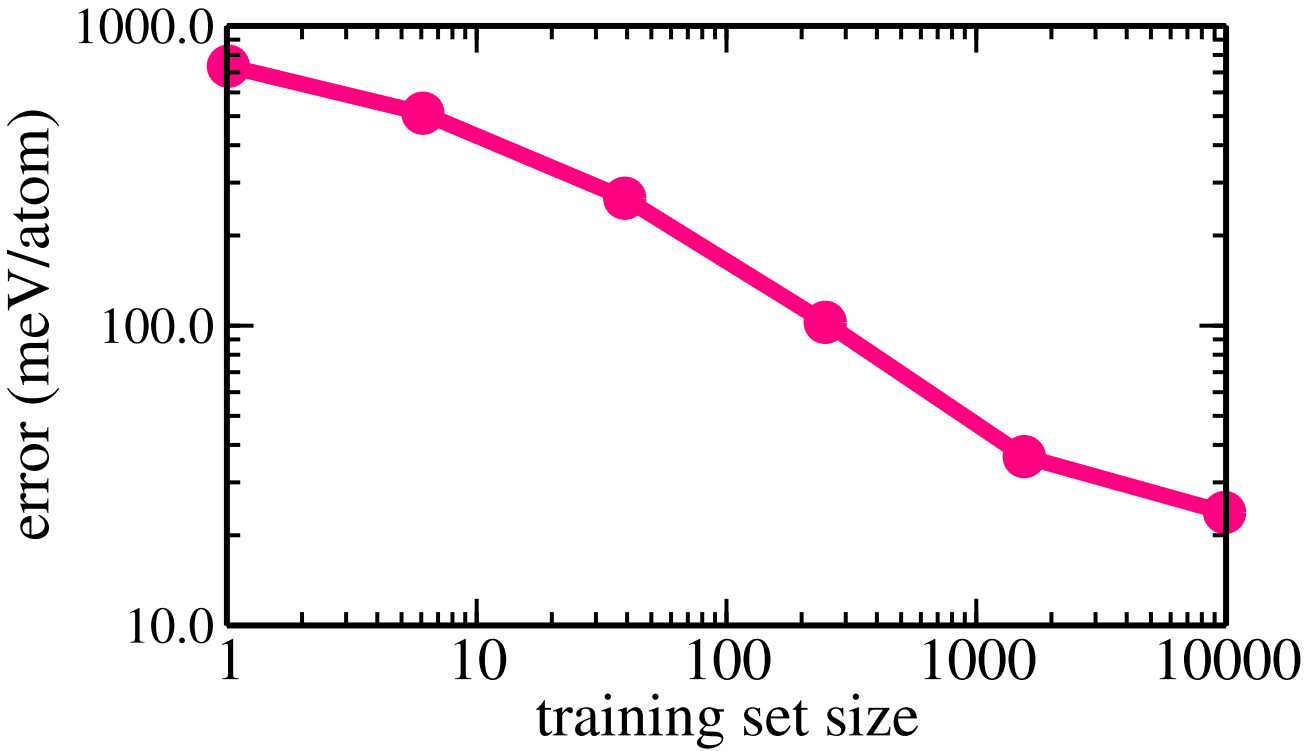
Descriptor

For each lattice position in the structure (here 4), the atom type is encoded by its column and row number in the periodic table



F. A. Faber, A. Lindmaa, O. A. von Lilienfeld, R. Armiento Phys. Rev. Lett. 117, 135502 (2016)

Application: ABC2D6 elpasolite crystals

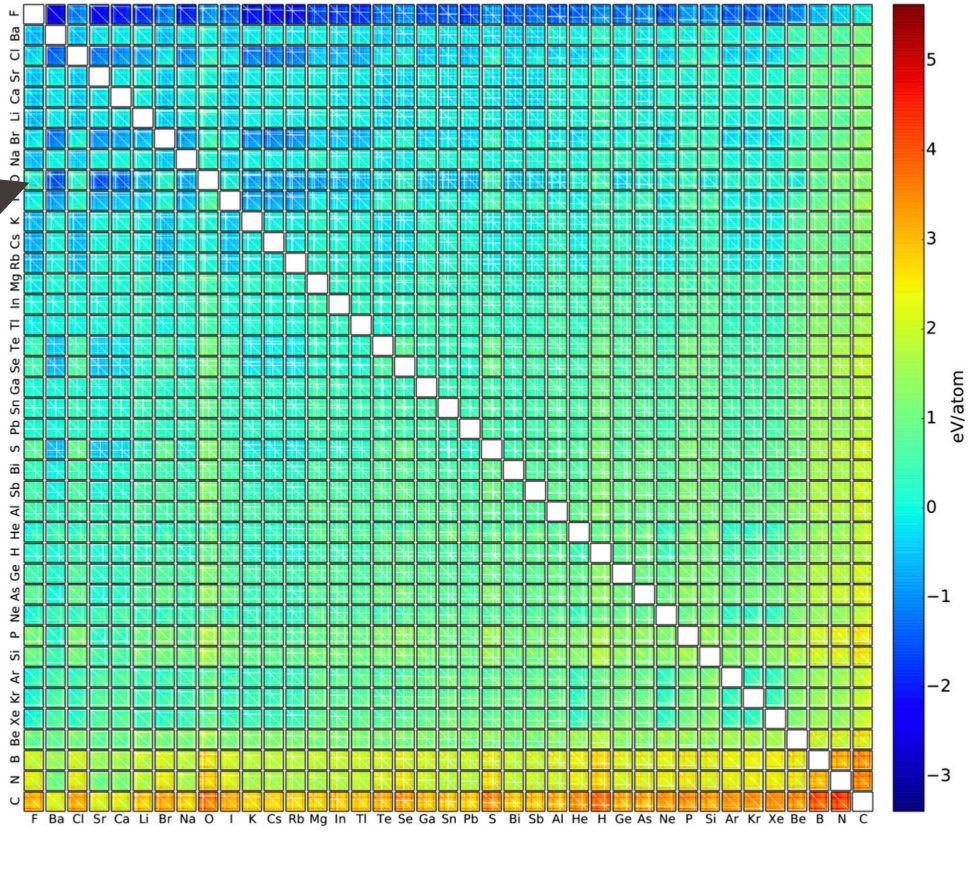


F. A. Faber, A. Lindmaa, O. A. von Lilienfeld, R. Armiento Phys. Rev. Lett. 117, 135502 (2016)

Inference: Predictions for 2M crystals

Each square is a 39 x 39 block of the 3rd and 4th element

First 2 elements



F. A. Faber, A. Lindmaa, O. A. von Lilienfeld, R. Armiento Phys. Rev. Lett. 117, 135502 (2016)

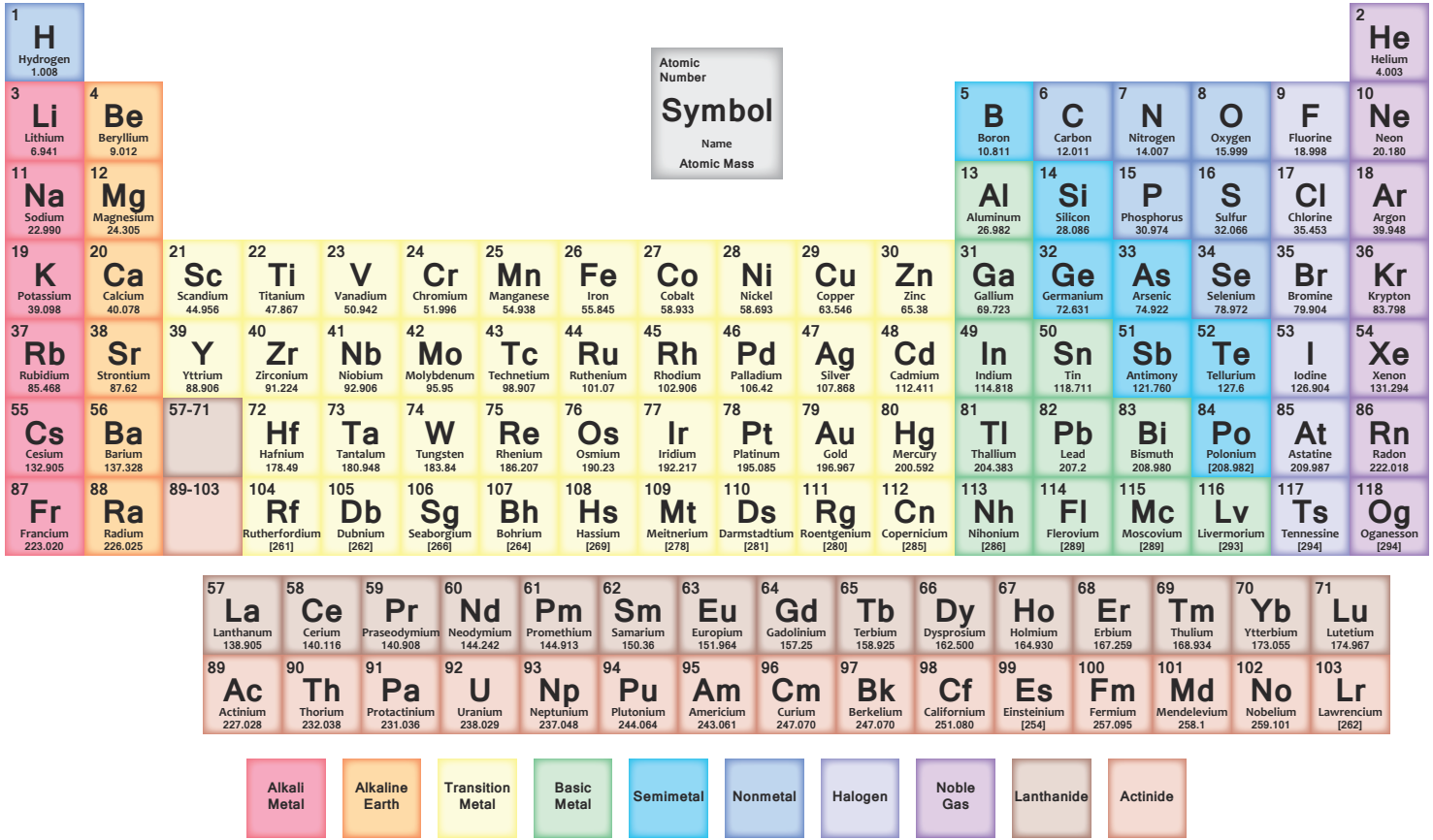
Attribute based descriptors

Problem: How do we present a list of general materials to a machine?

ZnO, Cu_2O_3, $YZrO_2$, $FeCrMnNiCo$, $CsPbCl_3$, $BaTiO_2$, H_2O etc.

More general solution:
Use the periodic table to expand chemical formulas into descriptor and supplement with data from **material handbooks** and electronic structure theory calculations.

Attribute based descriptors



Attribute-based descriptors: MAGPIE

Elemental Property Stats (115):

6 statistics: mean, variance, max, min, range, mode; of 22 elemental properties: Z, row, column, radius, electronegativity, s, p, d… valence electrons, unfilled states, magnetic moment….

Stoichiometry:

no. of components (fractions of elements)

Electronic-structure based:

fractions of s, p, d… electrons in the material

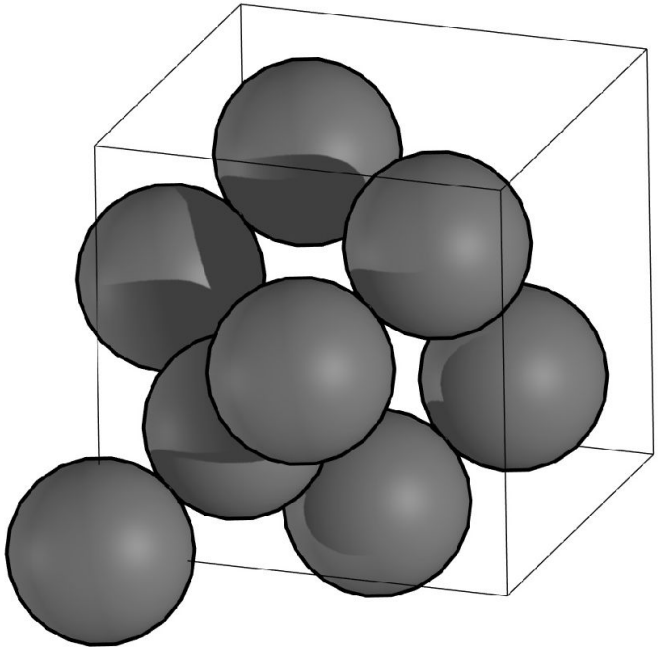
Attribute-based descriptors advantages

1. **Simple:**
   easy to implement and use
2. **Compact:**
   very, very small fingerprints
3. **Invariant:**
   invariances to rotation and translation built in
4. **Interpretability**
   features encode information on chemical elements and certain properties directly

Attribute-based descriptors disadvantages

1. **poor regression accuracy:**
   bond lengths and bond angles are not encoded
2. **no structural information:**
   polymorphs of the same material cannot be resolved
3. **lack of transferability:**
   fixed feature selection might not translate to new domains

Structure based descriptors: Coulomb Matrix



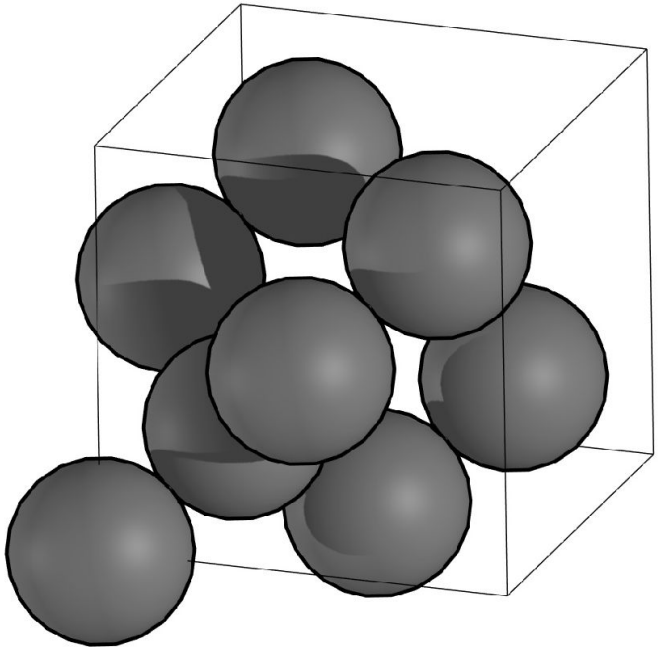
diamond

$$\mathbf{M}_{IJ} = \begin{cases} 0.5 Z_I^{2.4} & \text{for } I = J \\ \dfrac{Z_I Z_J}{|\mathbf{R}_I - \mathbf{R}_J|} & \text{for } I \neq J \end{cases}$$

atomic charges

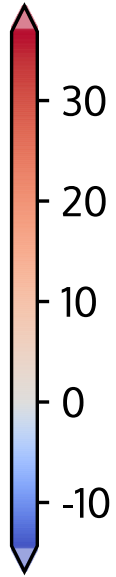
atomic positions

Coulomb Matrix



diamond

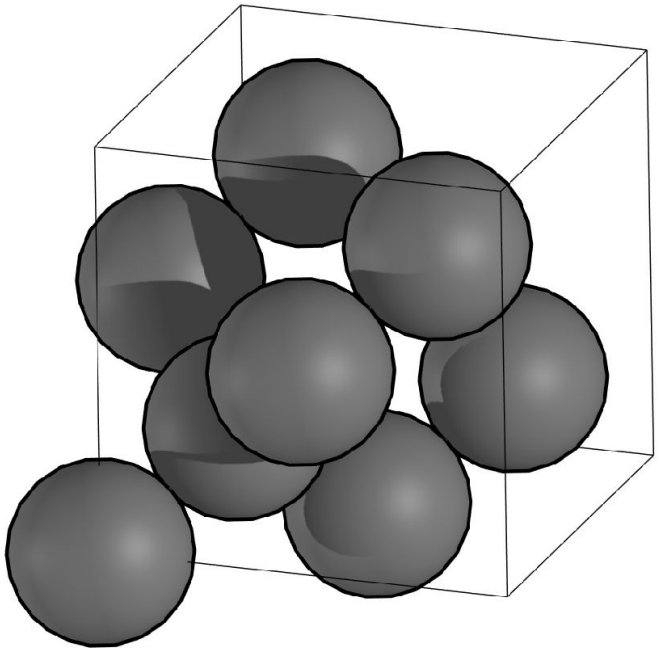
36.9	23.3	14.3	9.3	14.3	9.3	14.3	9.3
23.3	36.9	23.3	14.3	23.3	14.3	23.3	14.3
14.3	23.3	36.9	23.3	14.3	12.2	14.3	12.2
9.3	14.3	23.3	36.9	12.2	14.3	12.2	14.3
14.3	23.3	14.3	12.2	36.9	23.3	14.3	12.2
9.3	14.3	12.2	14.3	23.3	36.9	12.2	14.3
14.3	23.3	14.3	12.2	14.3	12.2	36.9	23.3
9.3	14.3	12.2	14.3	12.2	14.3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



not periodic

M. Rupp, Int. J. Quantum Chem. 115, 1058 (2015)
M. Rupp, et al, Phys. Rev. Lett. 108, 058301 (2012)

Ewald sum matrix



diamond

Electrostatic energy between atoms *i* and *j*:

$$\phi_{ij} = \sum_{\mathbf{n}} \frac{Z_i Z_j}{|\mathbf{R}_i - \mathbf{R}_j| + \mathbf{n}}$$

Infinite lattice sum:

$$\mathbf{n} = h\mathbf{a} + k\mathbf{b} + l\mathbf{c}$$

You will learn tomorrow how to evaluate!

M. Rupp, Int. J. Quantum Chem. 115, 1058 (2015)
M. Rupp, et al, Phys. Rev. Lett. 108, 058301 (2012)

Simplification: Sine matrix

$$M_{ij}^{\text{sine}} = \begin{cases} 0.5 Z_i^{2.4} & \forall\ i = j \\ \phi_{ij} & \forall\ i \neq j \end{cases}$$

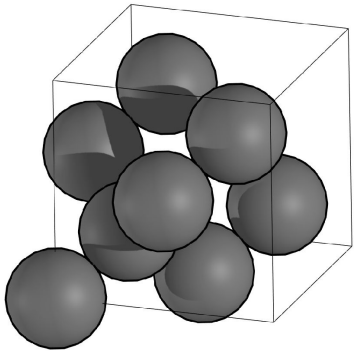
$$\phi_{ij} = Z_i Z_j |\mathbf{B} \cdot \sum_{k=\{x,y,z\}} \hat{\mathbf{e}}_k \sin^2\left(\pi \mathbf{B}^{-1} \cdot (\mathbf{R}_i - \mathbf{R}_j)\right)|^{-1}$$

matrix formed by lattice vectors

Cartesian unit vectors

F. Faber, A. Lindmaa, O. A. v. Lilienfeld, R. Armiento, Int. J. QuantumChem. 115, 1094 (2015)

Comparison of the three matrices



Coulomb matrix

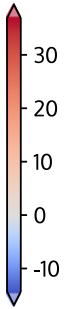
36.9	23.3	14.3	9.3	14.3	9.3	14.3	9.3
23.3	36.9	23.3	14.3	23.3	14.3	23.3	14.3
14.3	23.3	36.9	23.3	14.3	12.2	14.3	12.2
9.3	14.3	23.3	36.9	12.2	14.3	12.2	14.3
14.3	23.3	14.3	12.2	36.9	23.3	14.3	12.2
9.3	14.3	12.2	14.3	23.3	36.9	12.2	14.3
14.3	23.3	14.3	12.2	14.3	12.2	36.9	23.3
9.3	14.3	12.2	14.3	12.2	14.3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



not periodic

periodic

L. Himanen, et. al Comp. Phys. Commun. 247, 106949 (2020)

Ewald and sine matrices - advantages

1. **Simple:**
   easy to implement and use
2. **Compact:**
   small descriptor
3. **Invariant:**
   invariances to rotation and translation built in
4. **Quantitative structure information**
   bond lengths now encoded

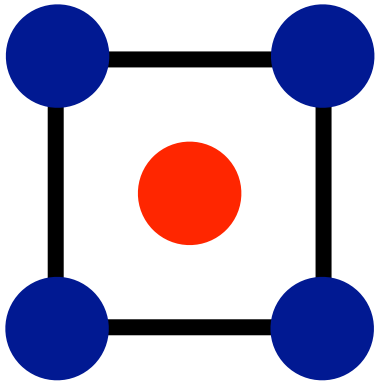
L. Himanen, et. al Comp. Phys. Commun. 247, 106949 (2020)

Ewald and sine matrices - disadvantages

1. **Moderate regression accuracy:**
   structural information still limited
2. **Not continuous:**
   small changes in the structure can switch the rows and columns
3. **Lack of interpretability:**
   features correspond to inverse distances, but are becoming hard to interpret
4. **Not transferable**
   Descriptor size changes with number of atoms

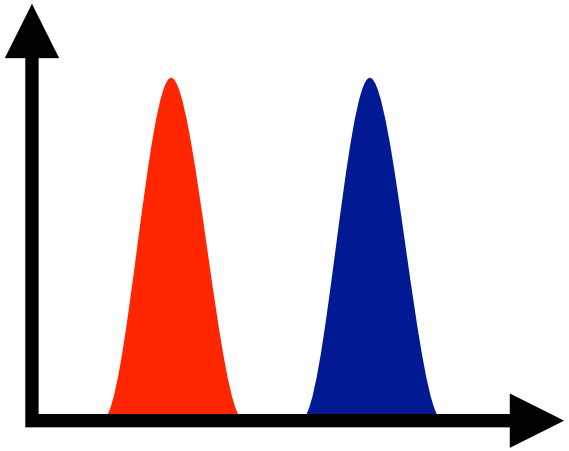
L. Himanen, et. al Comp. Phys. Commun. 247, 106949 (2020)

From global to local

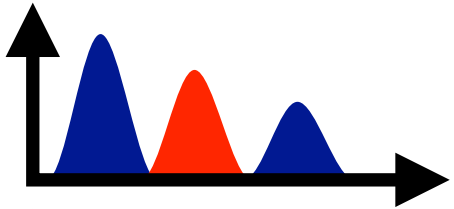


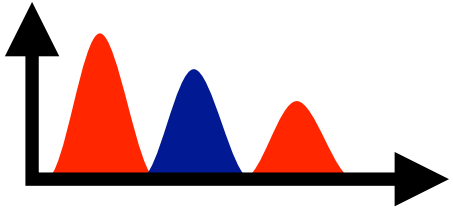
global

One descriptor for the **whole structure**



local

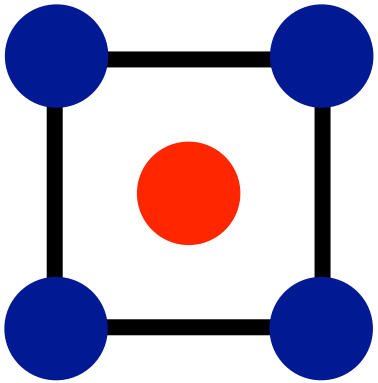




One descriptor for the **every atom**

L. Himanen, et. al Comp. Phys. Commun. 247, 106949 (2020)

From global to local



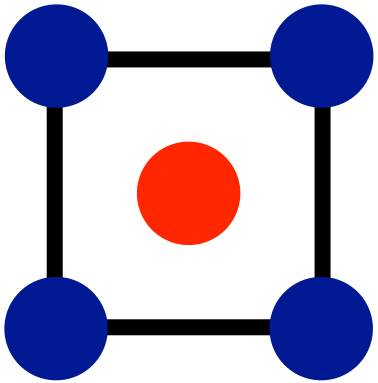
Locality principle:
many properties are local
or approximately additive

The total energy in DFT is not additive!

But we treat it as such approximately:

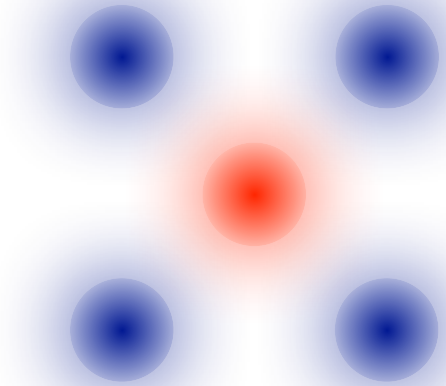
$$E = \sum_{i}^{N_{atoms}} E_i$$

Smooth overlap of atomic orbitals (SOAP)



Gaussian broadening atomic positions into a density field.

“Density trick”

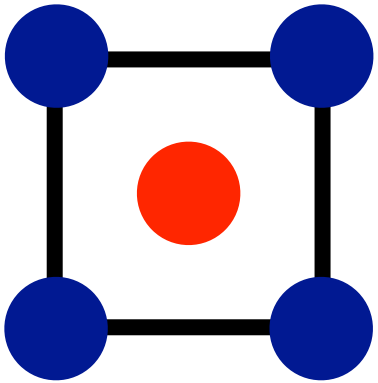


$$\rho_i(\boldsymbol{r}) = \sum_j e^{-\frac{1}{2\sigma^2}|r - r_{ij}|^2}$$

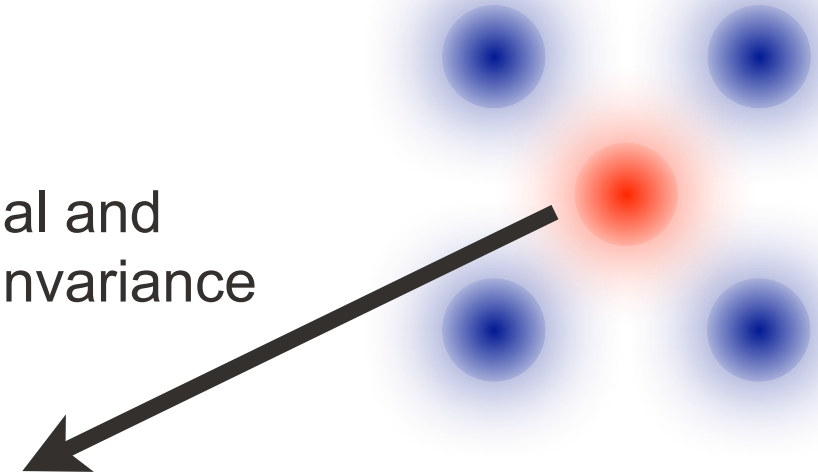
$$\boldsymbol{r}_{ij} = \boldsymbol{R}_i - \boldsymbol{R}_j$$

A. P. Bartók, R. Kondor, G. Csányi, Phys. Rev. B 87 184115 (2013)

Smooth overlap of atomic orbitals (SOAP)



translational and
rotational invariance

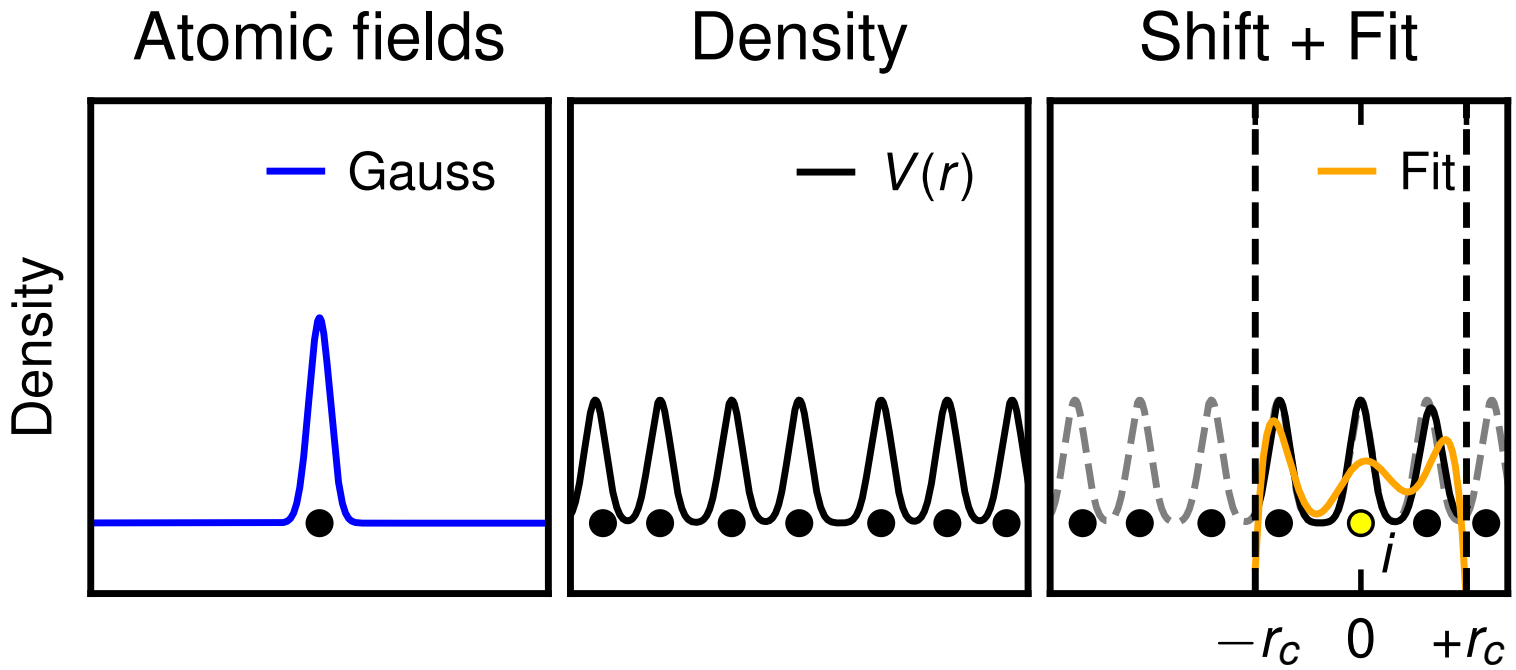


Expand in on a basis spherical harmonics:

$$\rho_i(\boldsymbol{r}) = \sum_{nlm} c_{i,nlm} R_n(r) Y_{lm}(\hat{\boldsymbol{r}})$$

A. P. Bartók, R. Kondor, G. Csányi, Phys. Rev. B 87 184115 (2013)

Smooth overlap of atomic orbitals (SOAP)

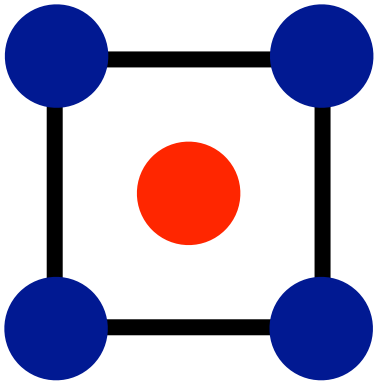


Project/fit the density on a basis built from radial & angular functions.

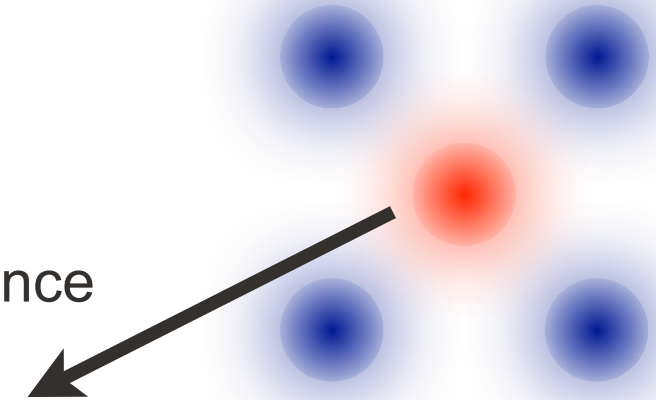
Take the coefficients of that projection as pairwise representation also called spherical expansion $\rho(r)$.

A. P. Bartók, R. Kondor, G. Csányi, Phys. Rev. B 87 184115 (2013)

Smooth overlap of atomic orbitals (SOAP)



translational and
rotational invariance



Combine spherical expansions and take zeroth order term of for power spectrum:

$$p_{nn'l}^{Z_1,Z_2} = \pi \sqrt{\frac{8}{2l+1}} \sum_m \left(c_{nlm}^{Z_1}\right)^* c_{n'lm}^{Z_2}$$

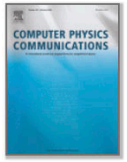
final SOAP descriptor

A. P. Bartók, R. Kondor, G. Csányi, Phys. Rev. B 87 184115 (2013)

Descriptor libraries

In practice, implementation of atomic descriptors is a bit painful… 😖

Computer Physics Communications
Volume 247, February 2020, 106949



DScribe: Library of descriptors for machine learning in materials science ☆

Lauri Himanen [a], Marc O.J. Jäger [a], Eiaki V. M
David Z. Gao [b, c], Patrick Rinke [a, f], Adam S. Foste

Show more

https://doi.org/10.1016/j.cpc.2019.106949

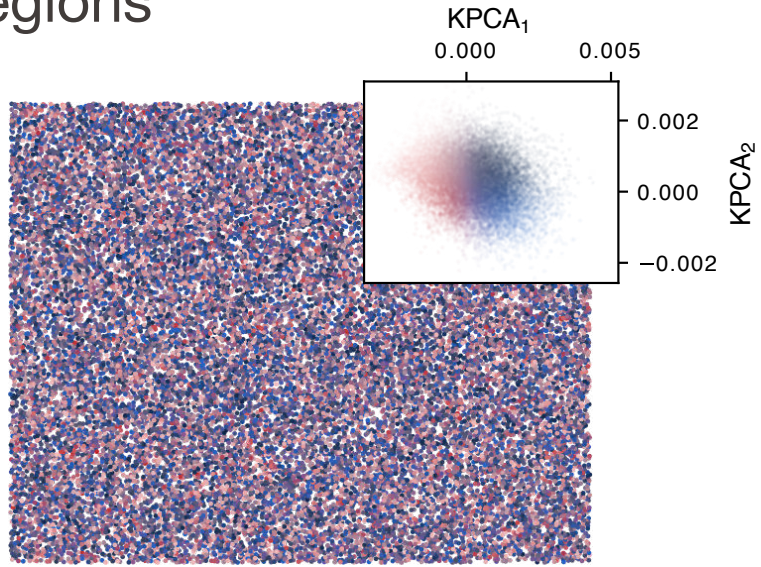
`metatensor` and `metatomic`: foundational libraries for interoperable atomistic machine learning

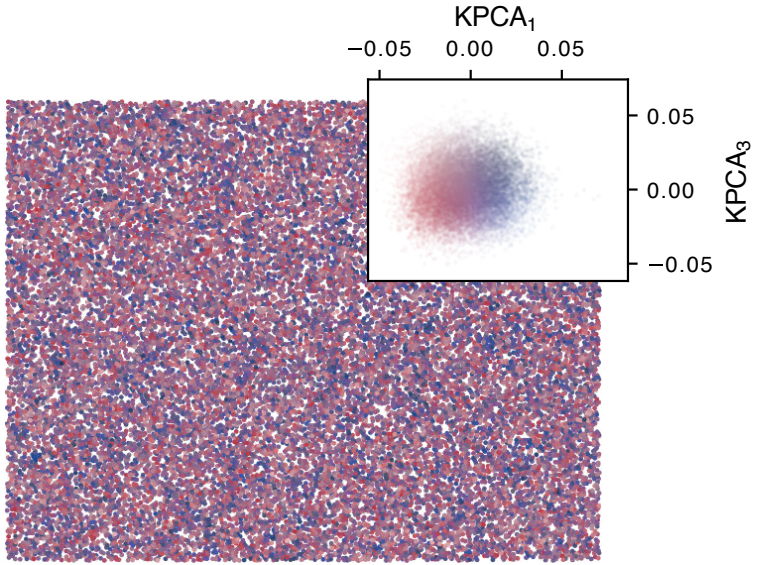
Filippo Bigi,[a] Joseph W. Abbott,[a] Philip Loche, Arslan Mazitov, Davide Tisi, Marcel F. Langer, Alexander Goscinski, Paolo Pegolo, Sanggyu Chong, Rohit Goswami, Sofiia Chorna, Matthias Kellner, Michele Ceriotti, and Guillaume Fraux
Laboratory of Computational Science and Modeling, Institute of Materials, École Polytechnique Fédérale de Lausanne, 1015 Lausanne, Switzerland

Incorporation of machine learning (ML) techniques into atomic-scale modeling has proven to be an extremely effective strategy to improve the accuracy and reduce the computational cost of simulations. It also entails conceptual and practical challenges, as it involves combining very different mathematical

Application: Supercooled iron

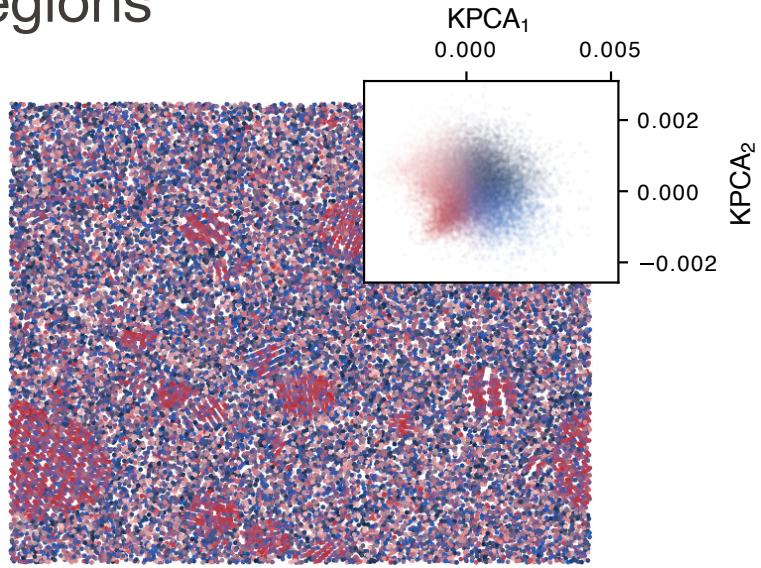
Color atoms based on whether they are in liquid or solid regions

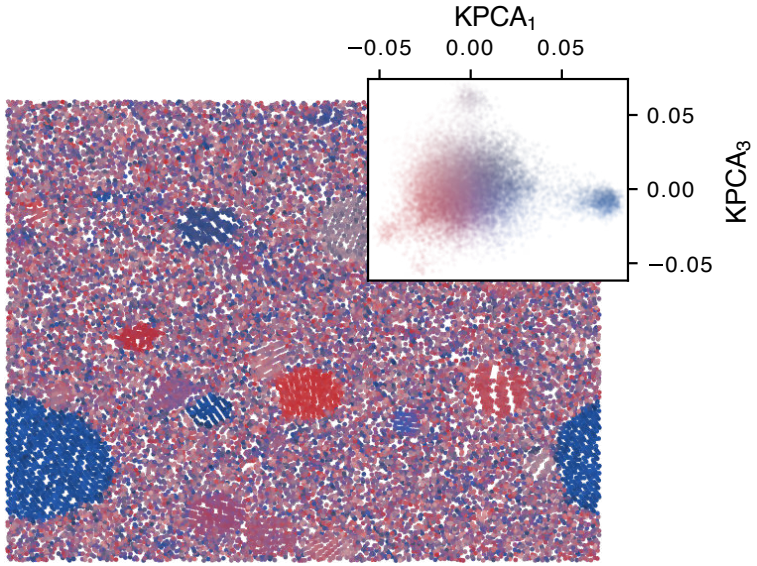




Application: Supercooled iron

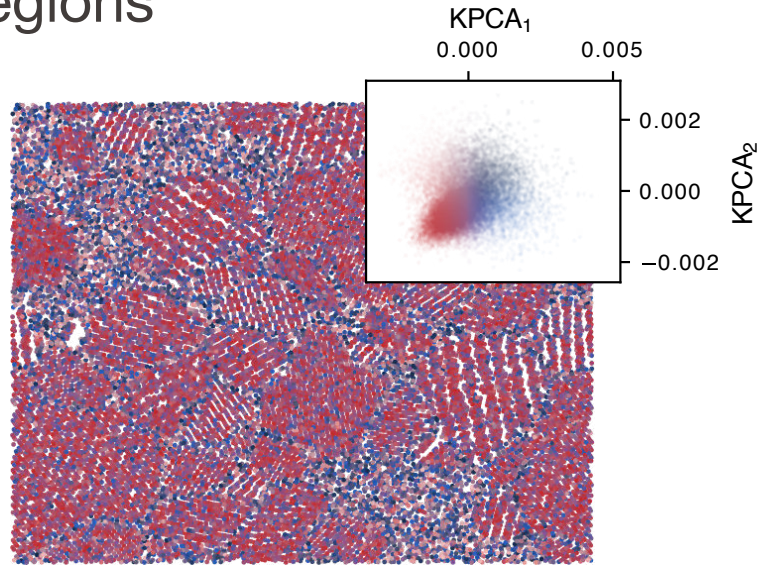
Color atoms based on whether they are in liquid or solid regions

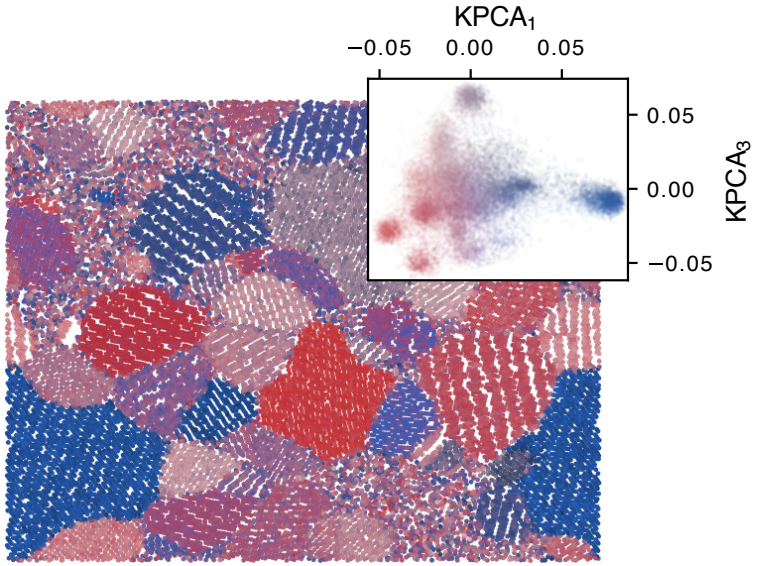




Application: Supercooled iron

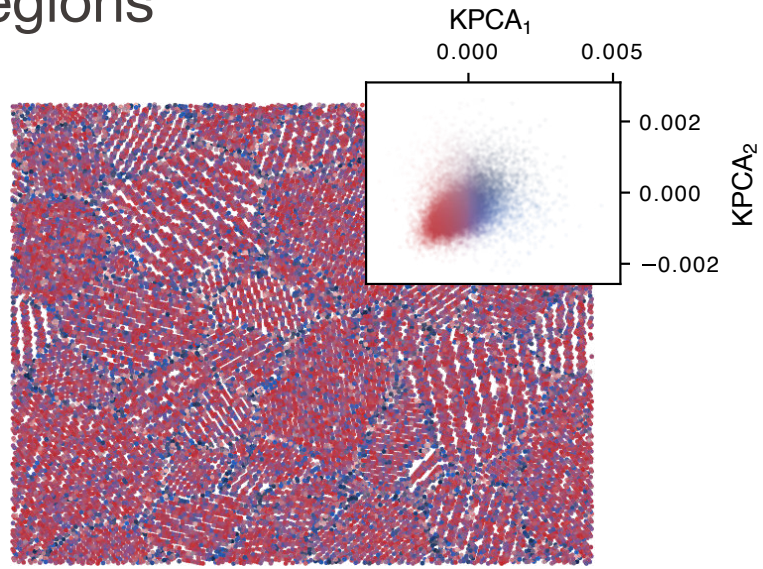
Color atoms based on whether they are in liquid or solid regions

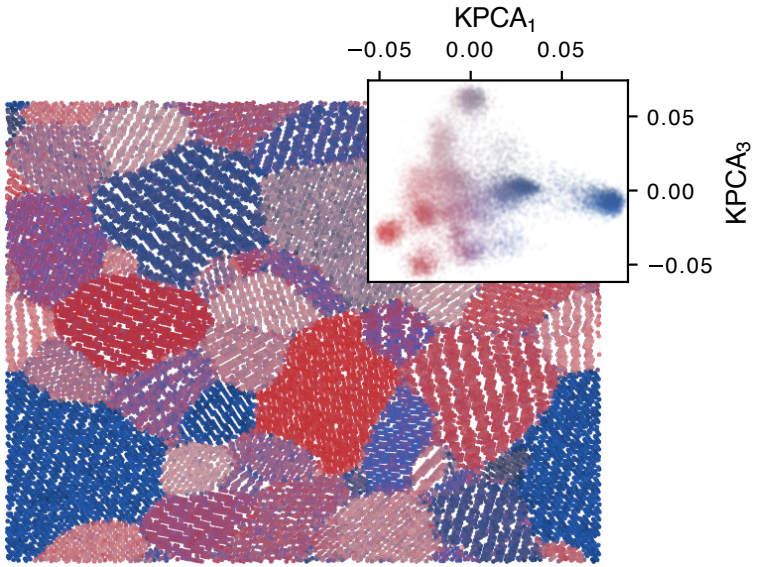




Application: Supercooled iron

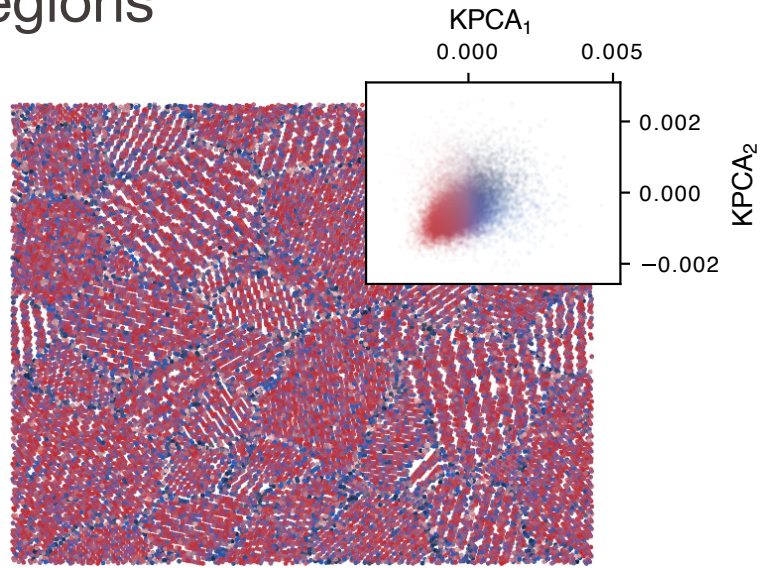
Color atoms based on whether they are in liquid or solid regions



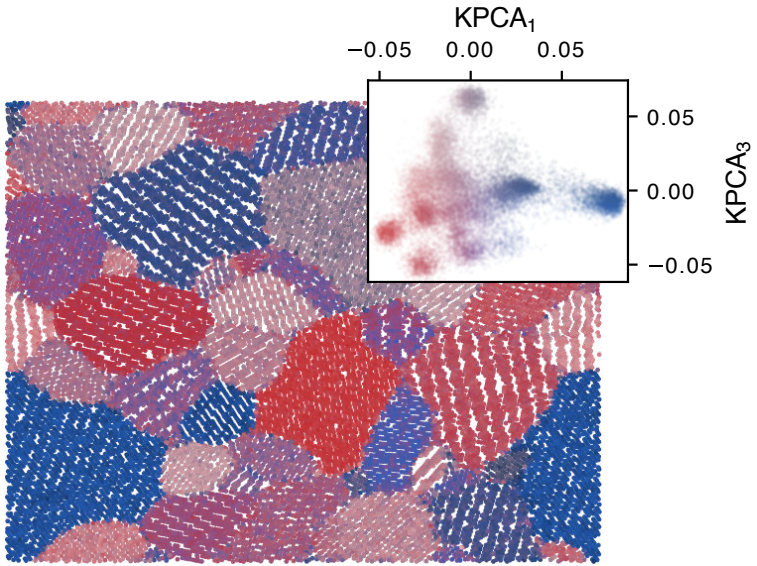


Application: Supercooled iron

Color atoms based on whether they are in liquid or solid regions



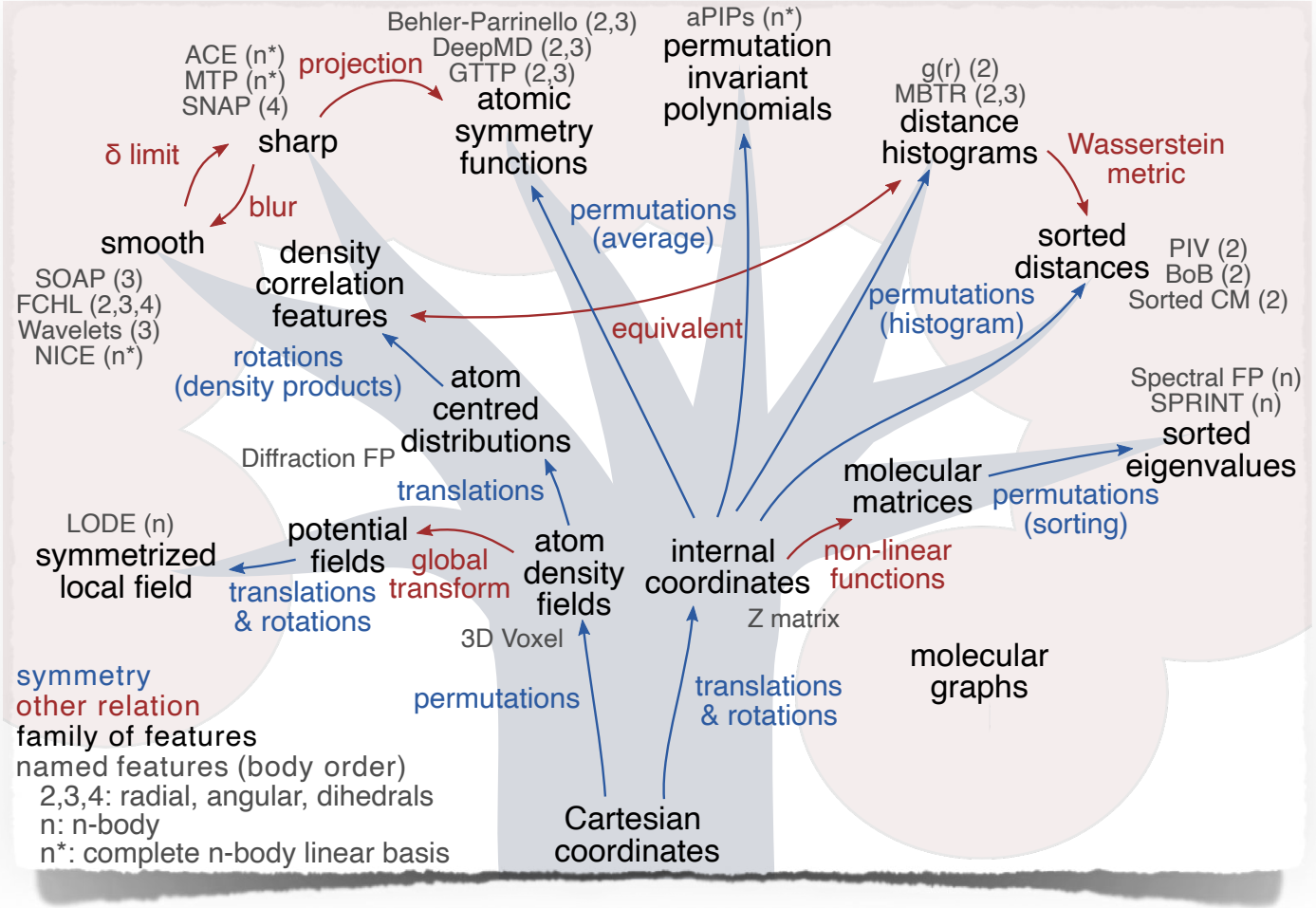
Rotational invariant



“Rotational non-invariant”

Understanding what goes into a representation is key to achieve meaningful results from automated data analytics

There are MANY related representations



Local descriptors - advantages

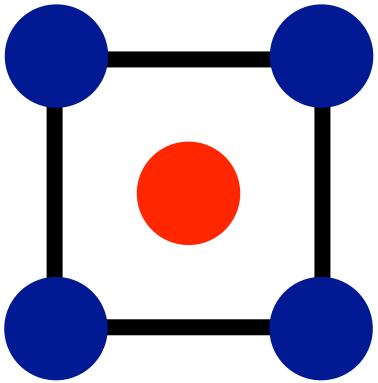
1. **Invariant:**
   invariances to rotation and translation built in
2. **Quantitative structure information**
   bond lengths, angles and higher orders encoded
3. **Continuity**
   continuous for small structural changes
4. **Locality**
   can resolve local environments (unlike global descriptors)
5. **Accuracy**
   among the most accurate descriptors for regression
6. **Versatility**
   works for molecules and periodic structures

Local descriptors - advantages

1. **Hyperparameters:**
   local descriptors often have many hyperparameters that need to be tuned, which is cumbersome and timely
2. **Size:**
   descriptors are large (~10^4 features)
3. **Lack of interpretability:**
   entries have no direct physical or chemical meaning

Descriptor based neural network potentials

From descriptors to inter-atomic potentials



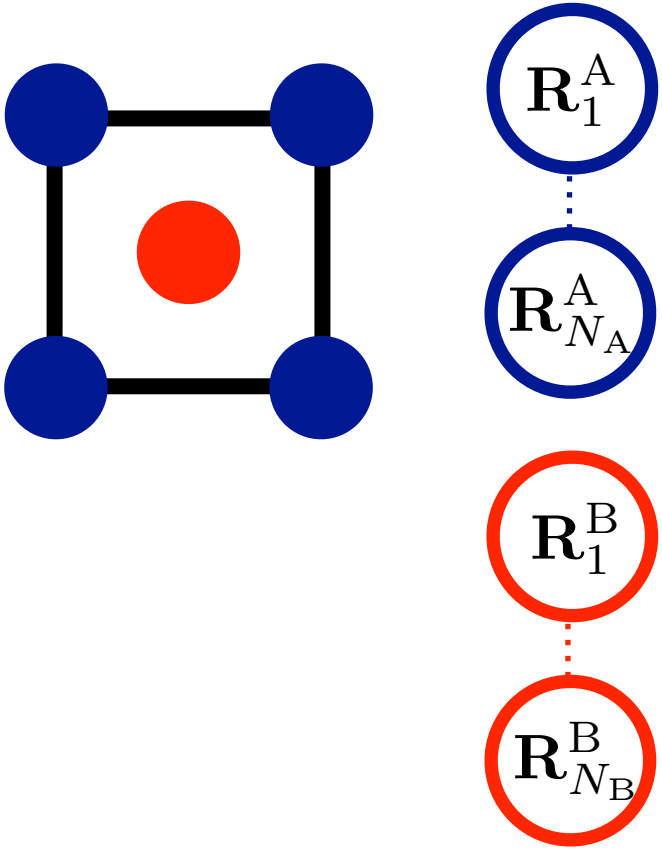
Locality as basis for inter-atomic potentials:

$$E = \sum_{i}^{N_{atoms}} E_i$$

$$E = \sum_{j}^{N_{elements}} \sum_{i}^{N_{atoms,j}} E_i^j$$

J. Behler, Int. J. Quant. Chem. 115, 1032 (2015)

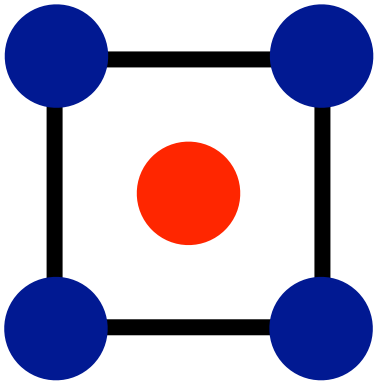
Behler-Parinello neural network potential

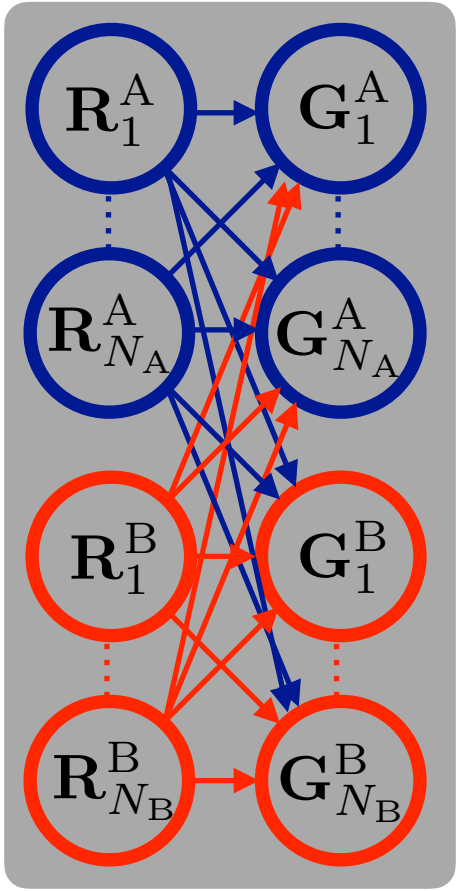


xyz coordinates of all atoms

Behler, J. & Parrinello, Phys. Rev. Lett. 98, 146401 (2007)

Behler-Parinello neural network potential





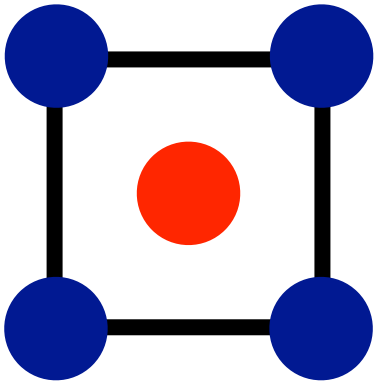
Mapping over via descriptors functions

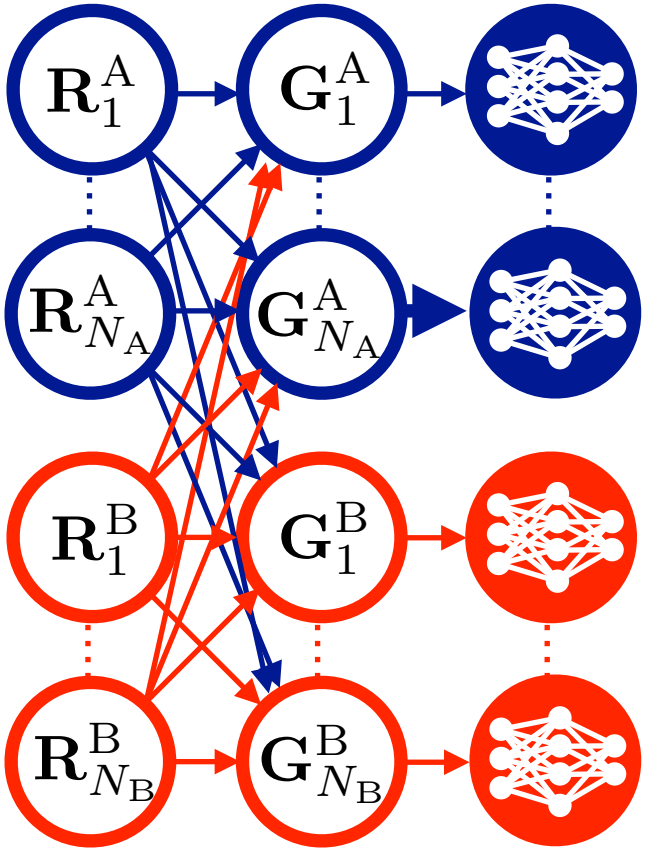
Descriptor:

- enforces invariances
- encodes local environments

Behler, J. & Parrinello, Phys. Rev. Lett. 98, 146401 (2007)

Behler-Parinello neural network potential





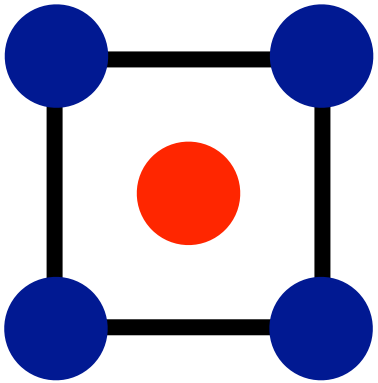
Add 1 neural network (NN) per atomic environment

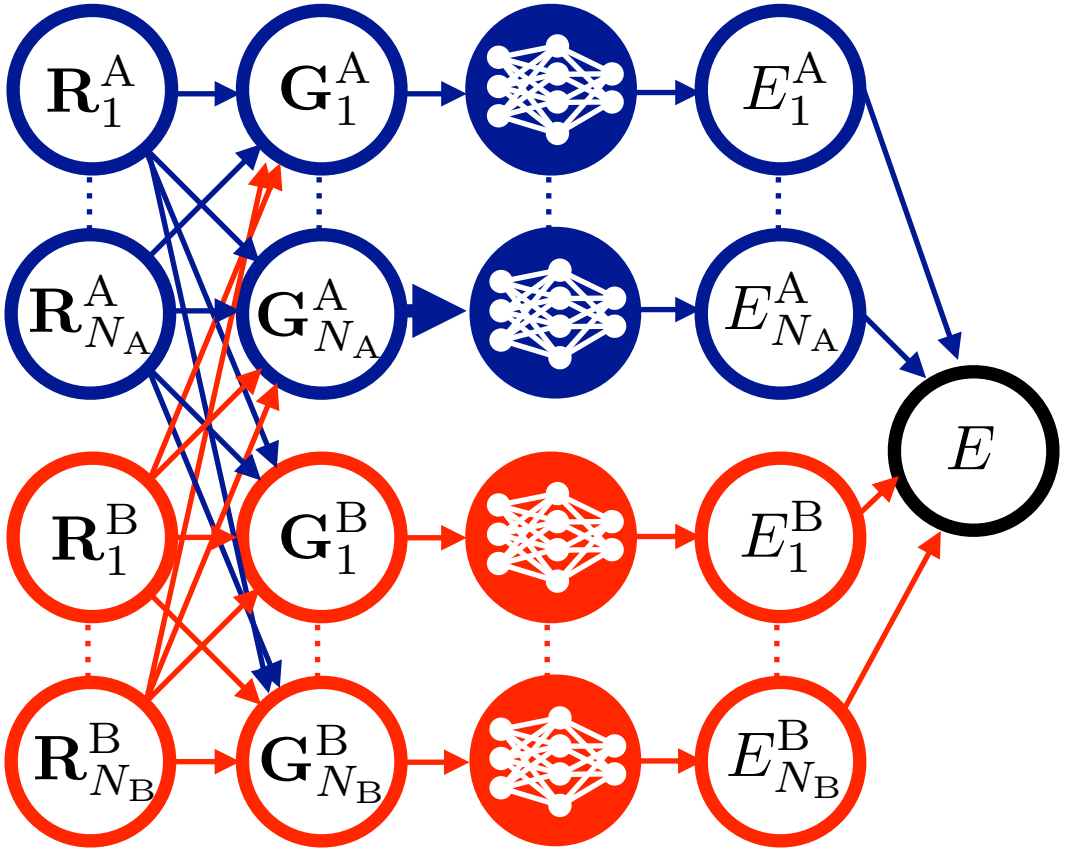
Only 1 NN per species

Difference comes in through local environments

Behler, J. & Parrinello, Phys. Rev. Lett. 98, 146401 (2007)

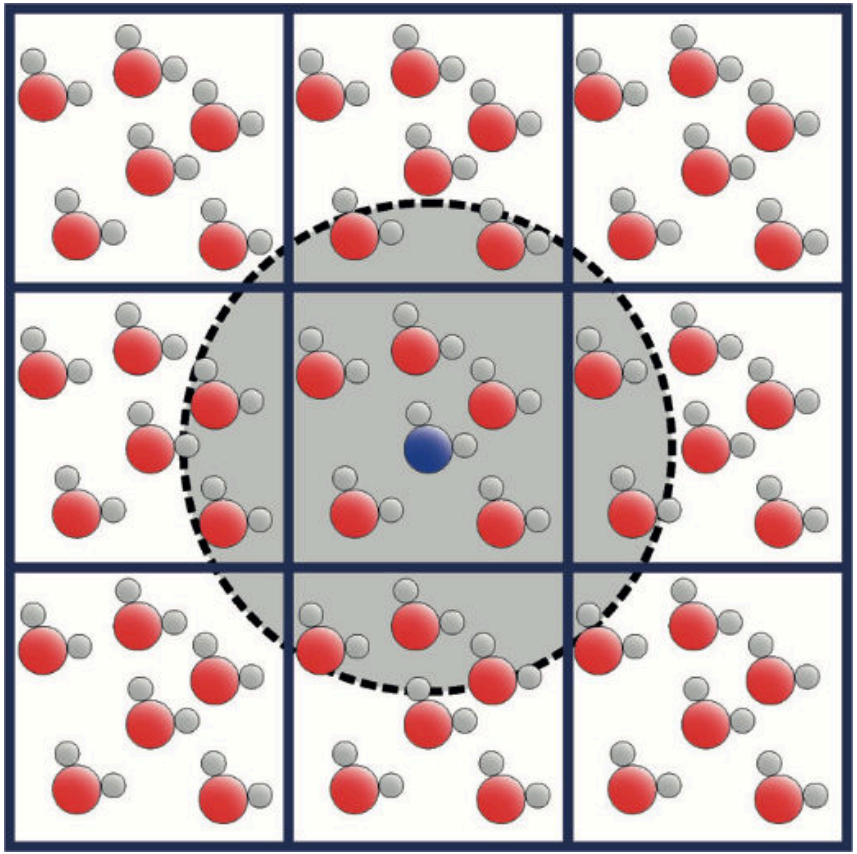
Behler-Parinello neural network potential





Behler, J. & Parrinello, Phys. Rev. Lett. 98, 146401 (2007)

Environment dependence



The **cut-off function** controls the number of atoms in the local environment.

Machine learning interatomic potentials (MLIP) can be trained for smaller structures and then applied to larger.

MLIPs scale usually linear with the number of particles

J. Behler, Angew. Chem. Int. Ed. 56, 12828 (2017)

How to get the forces and the stress

Force on an atom: $\boldsymbol{F}_i = -\dfrac{\partial E}{\partial \boldsymbol{R}_i}$

without efforts
by auto grad ✨

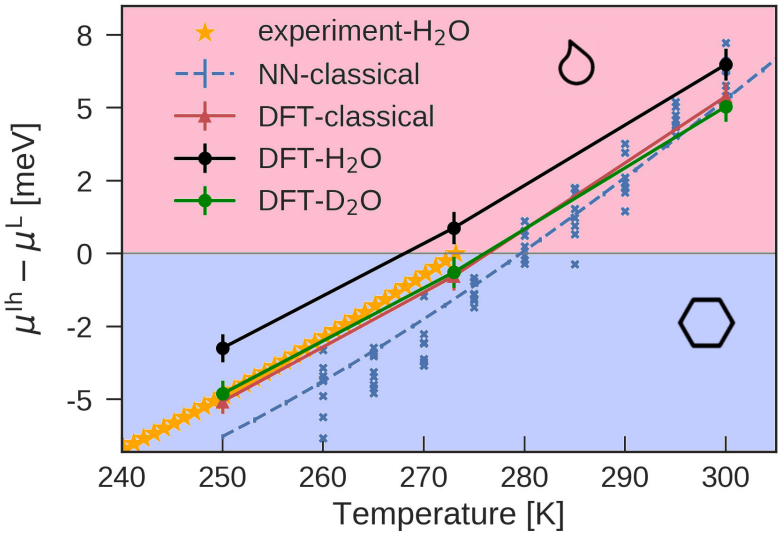
Stress: $V\sigma = \dfrac{\partial E}{\partial \epsilon}\bigg|_{\epsilon=0} = \sum_i^{N_{\text{atoms}}} \boldsymbol{R}_i \otimes \dfrac{\partial E}{\partial \boldsymbol{R}_i}$

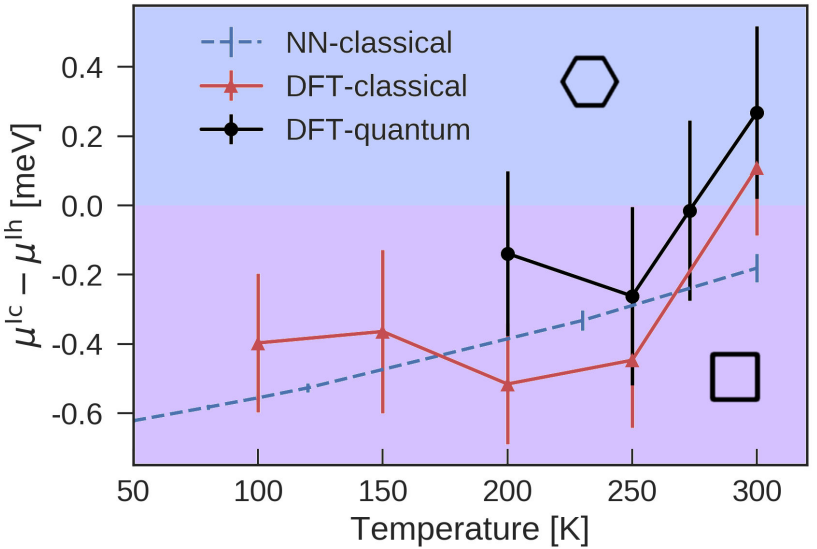
ϵ is the strain

Forces and stress are are usually part of the loss function in training.

Ab initio (thermo)dynamics made easy

Simulating matter at finite temperature, including quantum nuclear effects and dynamics is now much more affordable!

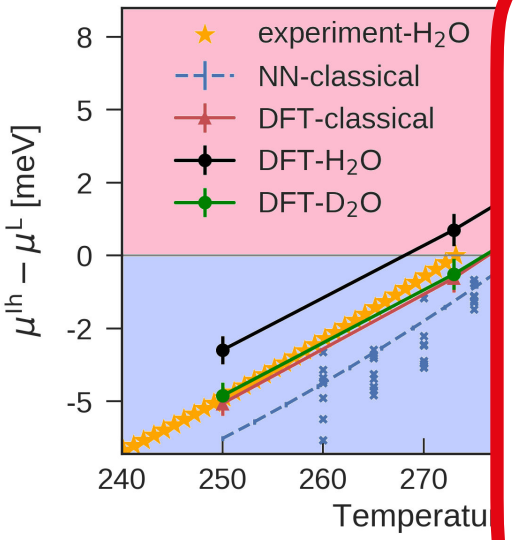




Cheng et. al. PNAS (2019)

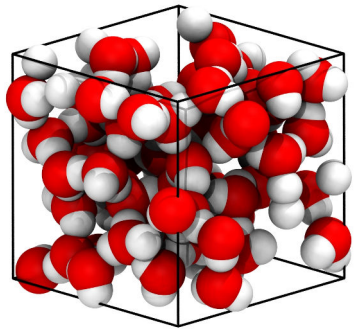
Ab initio (thermo)dynamics made easy

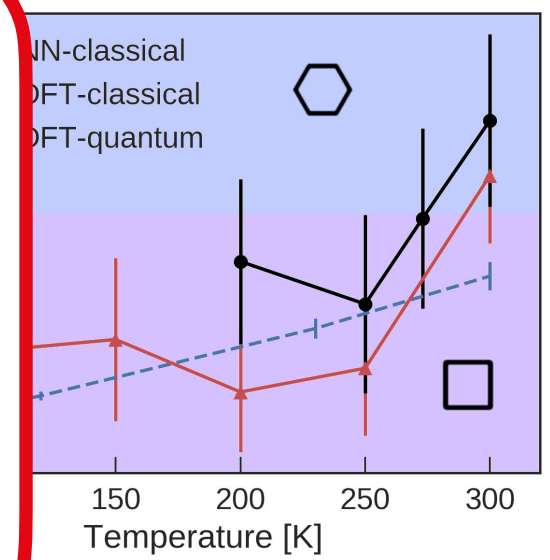
Simulating matter at finite temperature, including quantum nuclear effects and dynamics is now much more affordable!



Dataset

1,593 diverse reference structures of 64 molecules of liquid water at revPBE0 level.





Cheng et. al. PNAS (2019)

Descriptor based NN MLIPS - advantages

1. **invariant:**
   invariance of the energy to rotation and translation built in
2. **flexible functional form**
   species NNs can adjust to training data; apart from the symmetry functions for the local environments, no functional form for the potential is constraining
3. **locality**
   trained on smaller structures, but applied to larger, more complex ones
4. **accuracy**
   significantly more accurate than conventional force fields
5. **versatility**
   works for molecules and periodic structures

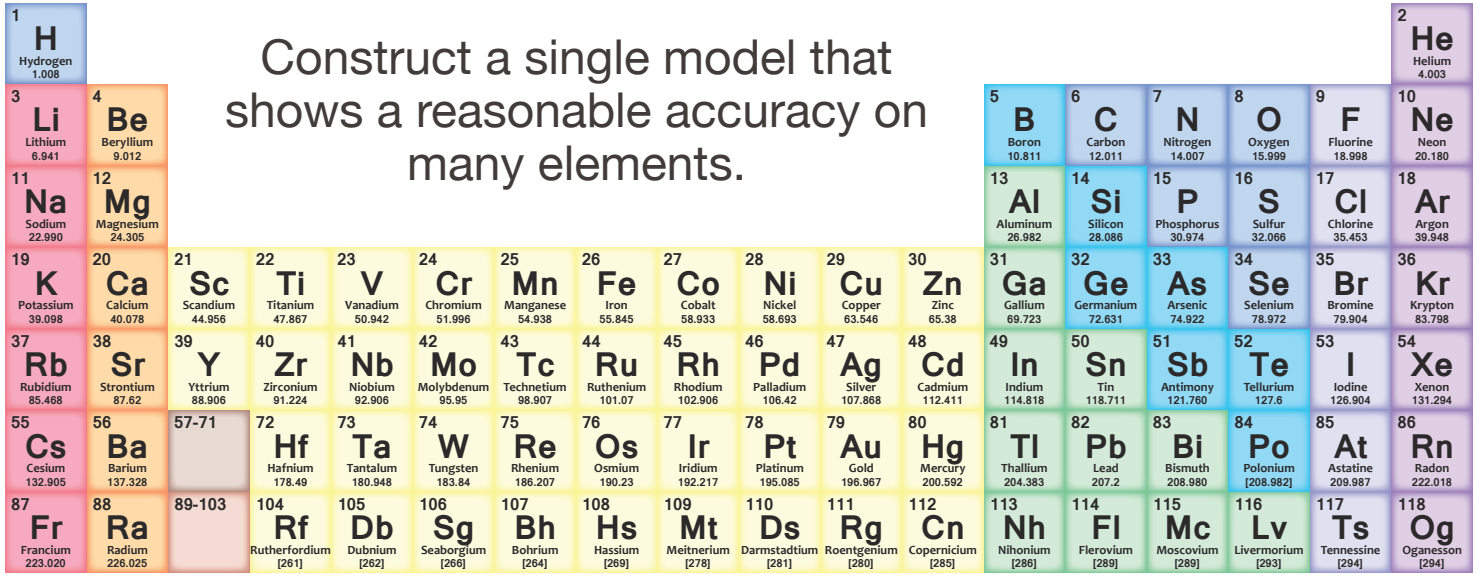
Descriptor based NN MLIPS - advantages

1. **training data:**
   require significantly more training data than conventional force fields
2. **transferability:**
   often generalise poorly to new or different environments
3. **scaling with species:**
   scale exponentially with the number of chemical species.

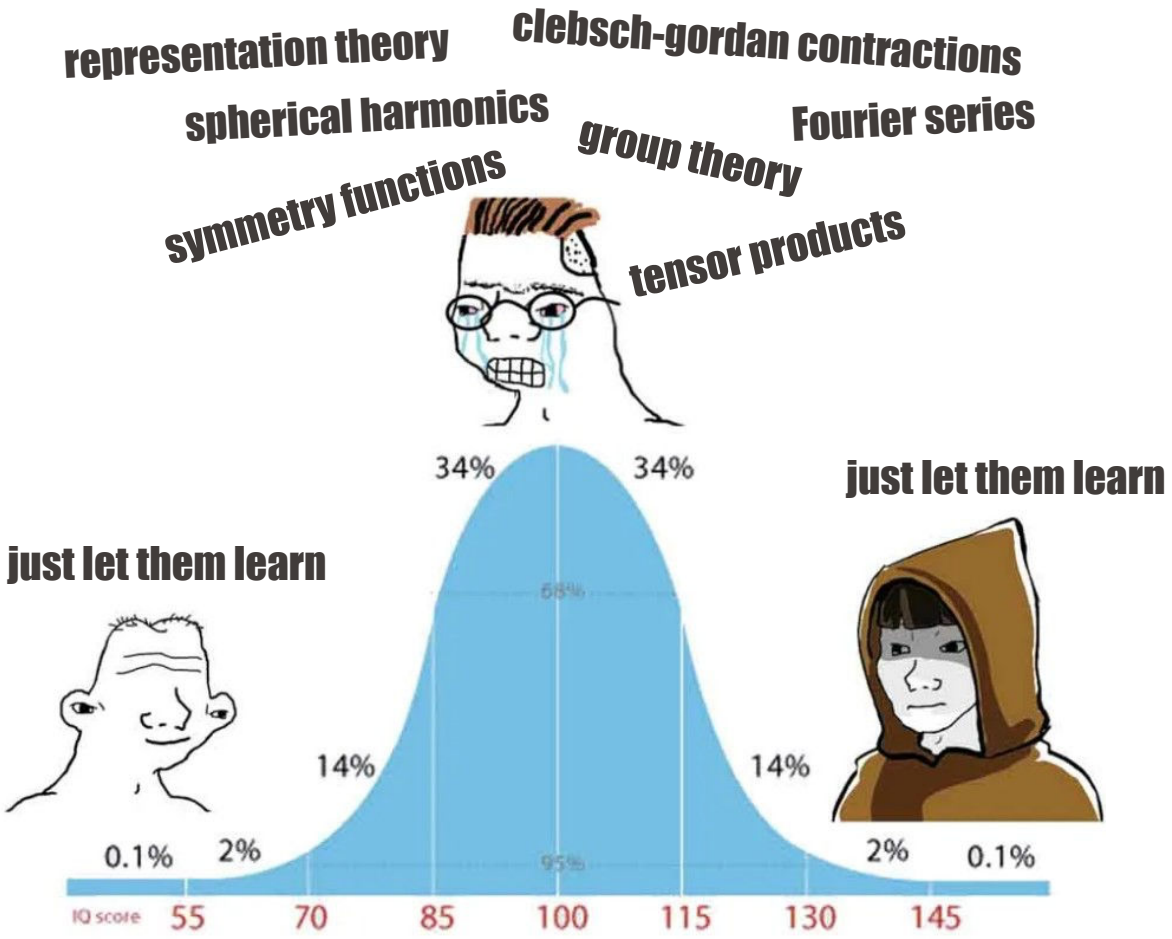
End-to-end neural network potentials

Can we do better?

Construct a single model that shows a reasonable accuracy on many elements.



Can we do better?



Can we do better?

representation theory

clebsch-gordan contractions

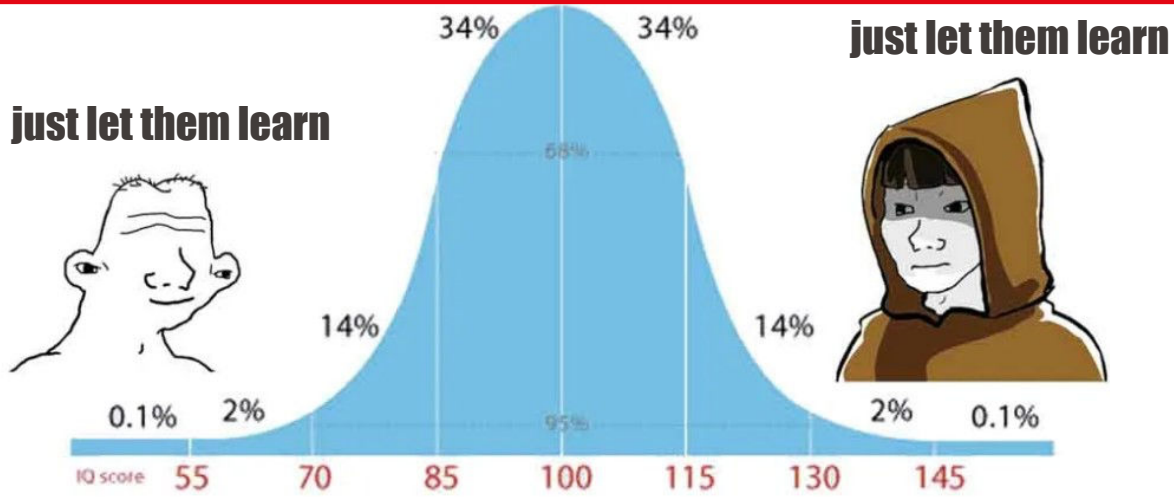
spherical harmonics

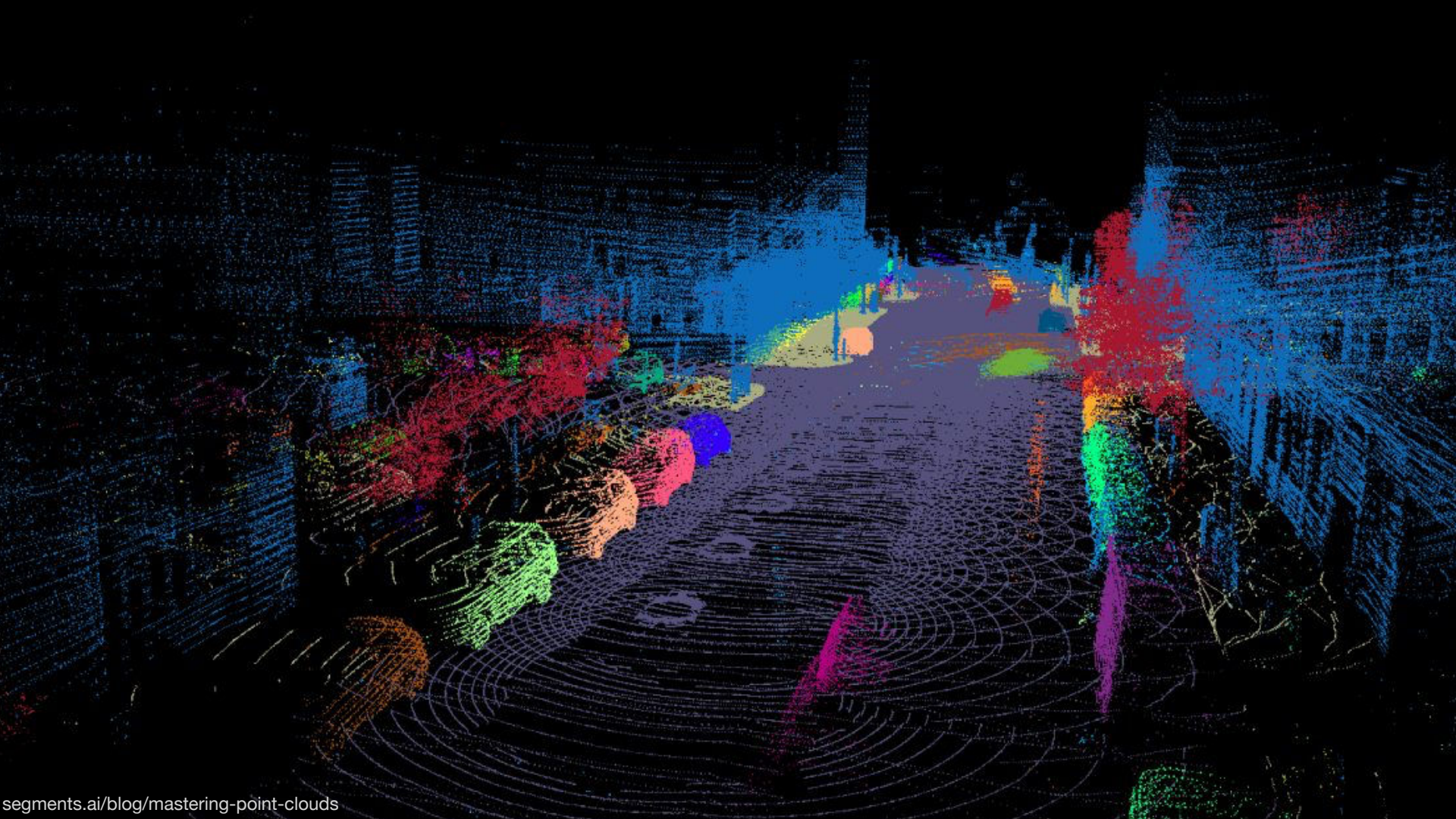
Fourier series

group the

Current state-of-the-art MLIP:

Equivariant message passing graph neural networks





segments.ai/blog/mastering-point-clouds

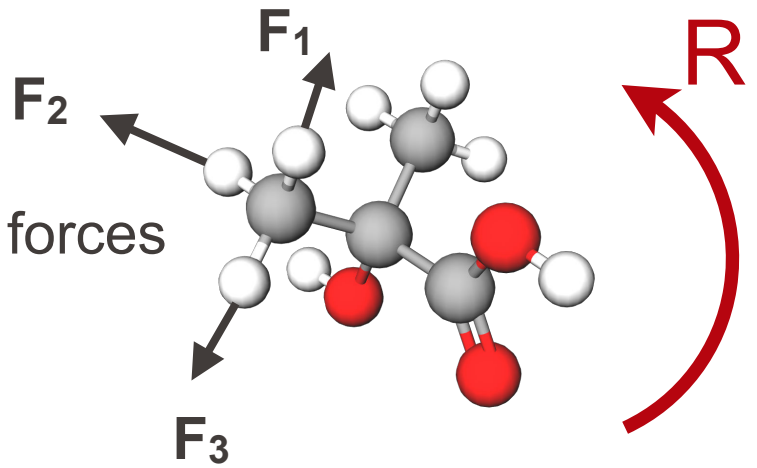
Invariance and equivariance

Equivariant message passing graph neural networks

Invariance with respect to rotation $\boldsymbol{R}$:

$E(\boldsymbol{R}x) = E(x)$ or in general $f(\boldsymbol{R}x) = f(x)$ for all $\boldsymbol{R} \in \mathbb{R}^{(3\times3)}$

Rotational invariance works for energy, but not forces!



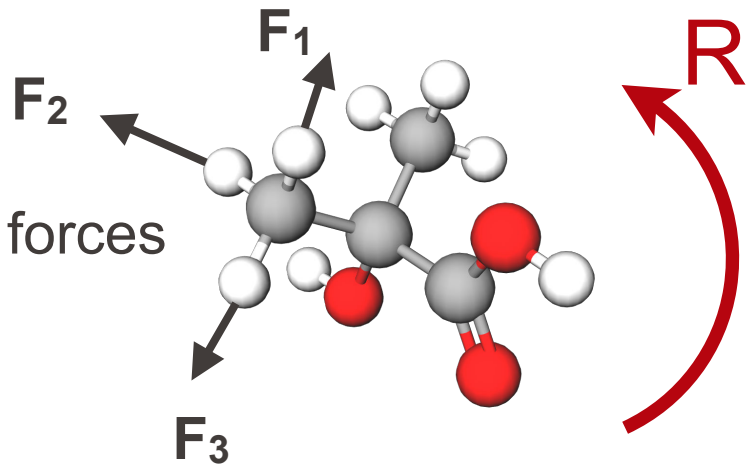
K. Schütt, O. Unke, M. Gastegger, PMLR 139, 9377 (2021)

Invariance and equivariance

Vectors (and tensors) change under rotations

Equivariance with respect to rotation $\boldsymbol{R}$:

$f(\boldsymbol{R}x) = \boldsymbol{R}f(x)$ for all $\boldsymbol{R} \in \mathbb{R}^{(3\times3)}$



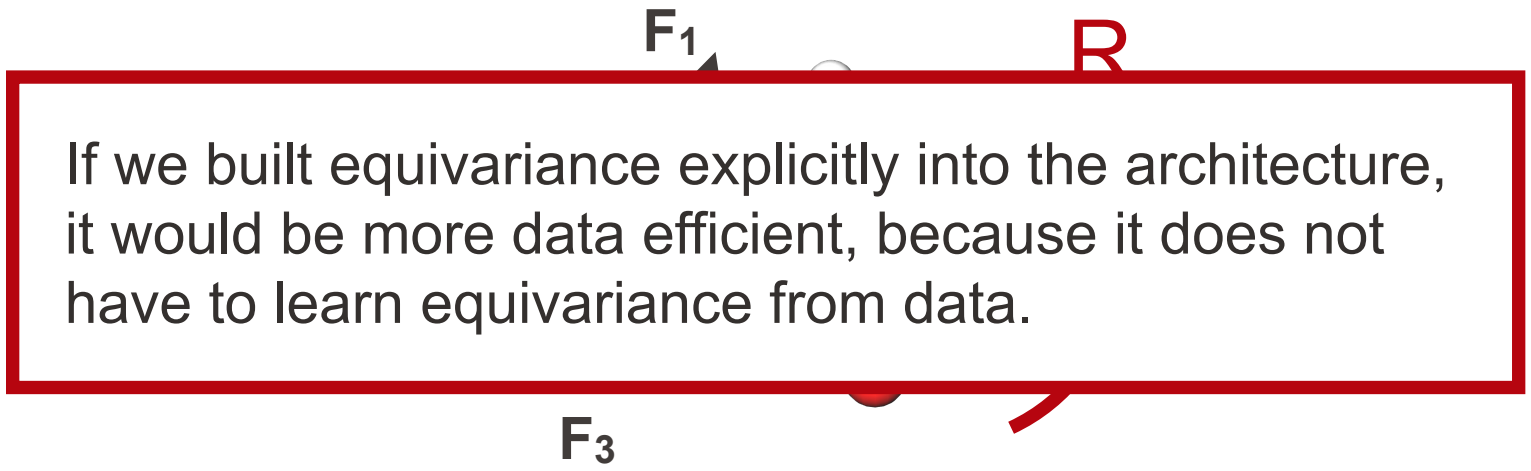
K. Schütt, O. Unke, M. Gastegger, PMLR 139, 9377 (2021)

Invariance and equivariance

Vectors (and tensors) change under rotations

Equivariance with respect to rotation $\boldsymbol{R}$:

$f(\boldsymbol{R}x) = \boldsymbol{R}f(x)$ for all $\boldsymbol{R} \in \mathbb{R}^{(3\times3)}$

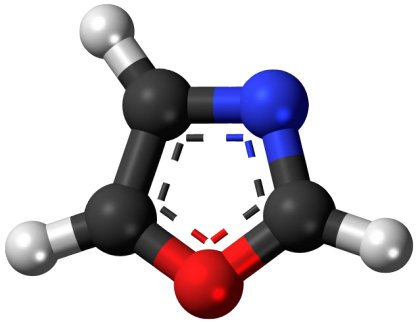


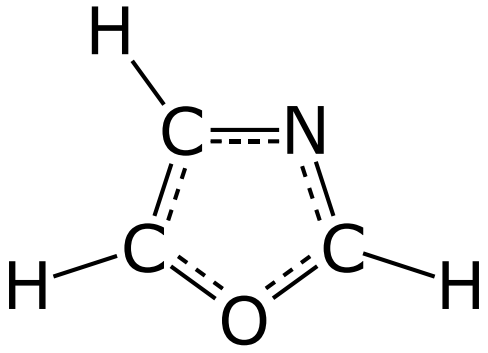
If we built equivariance explicitly into the architecture, it would be more data efficient, because it does not have to learn equivariance from data.

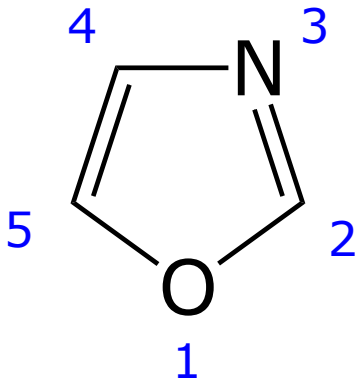
K. Schütt, O. Unke, M. Gastegger, PMLR 139, 9377 (2021)

Graph-based representation

Equivariant message passing graph neural networks



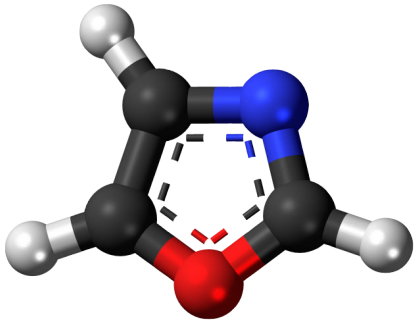




Representing molecules as graphs
(similar as molecular representations) - works also for solids

Graph-based representation

Equivariant message passing graph neural networks

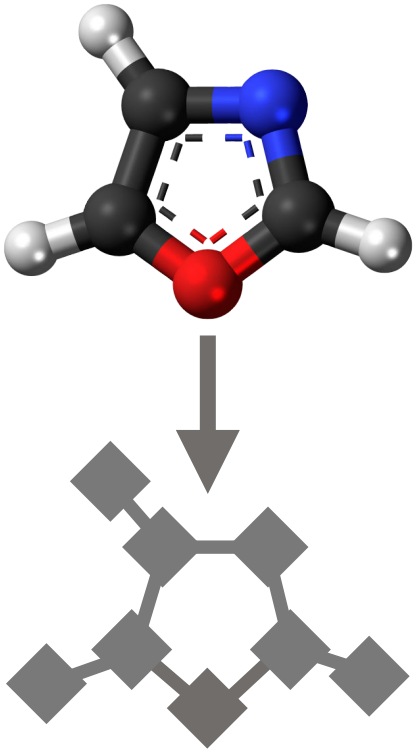


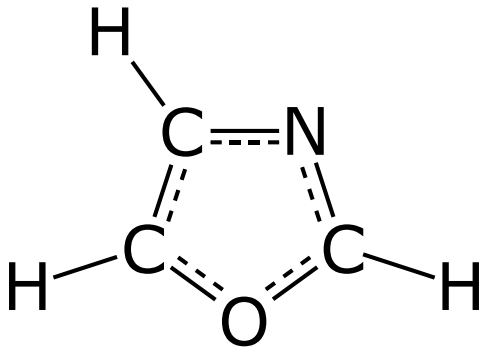
adjacency matrix encodes graph structure

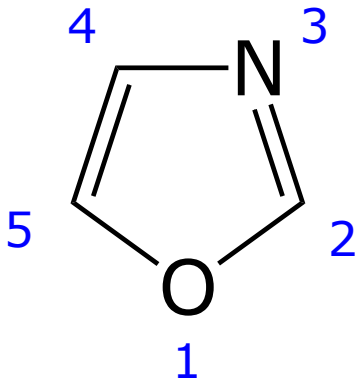
	1	2	3	4	5	6	7	8
1	0	1	0	0	1	0	0	0
2	1	0	1	0	0	0	1	0
3	0	1	0	1	0	0	0	0
4	0	0	1	0	1	0	0	1
5	1	0	0	1	0	1	0	0
6	0	0	0	0	1	0	0	0
7	0	1	0	0	0	0	0	0
8	0	0	0	1	0	0	0	0

Graph-based representation

Equivariant message passing graph neural networks





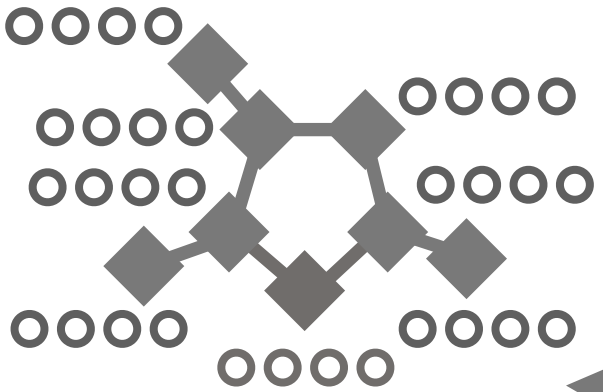


Each atom becomes a **node** in the graph

Each “bond” becomes an **edge** in the graph

K. Schütt, O. Unke, M. Gastegger, PMLR 139, 9377 (2021)

Graph-based representation



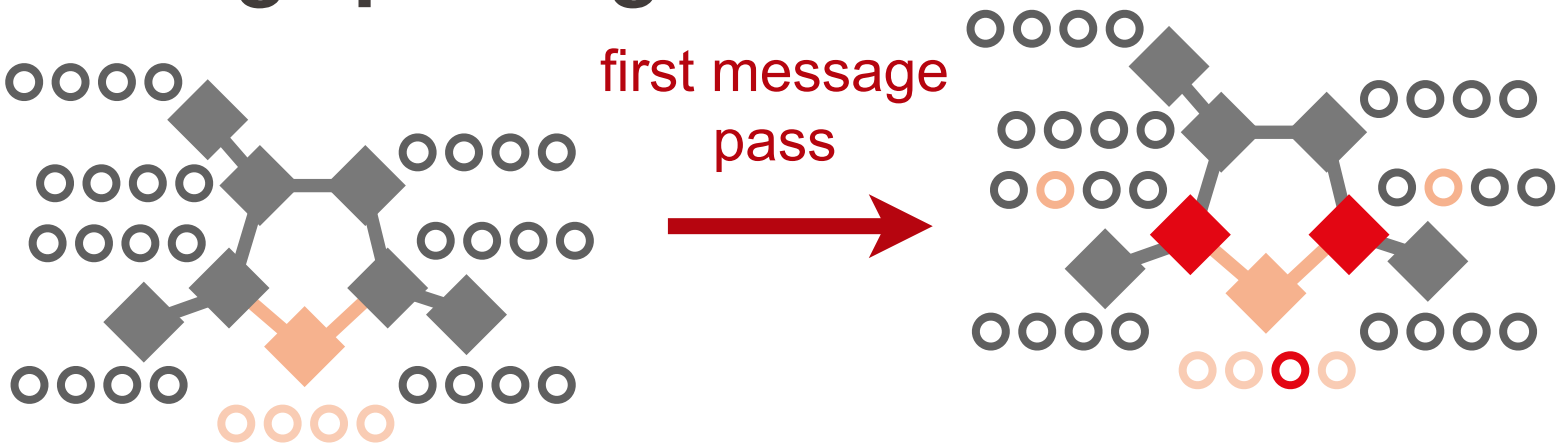
graph structure

each node carries information, called attributes:

- scalars, vectors, tensors,…
- encoded as neurons

K. Schütt, O. Unke, M. Gastegger, PMLR 139, 9377 (2021)

Message passing

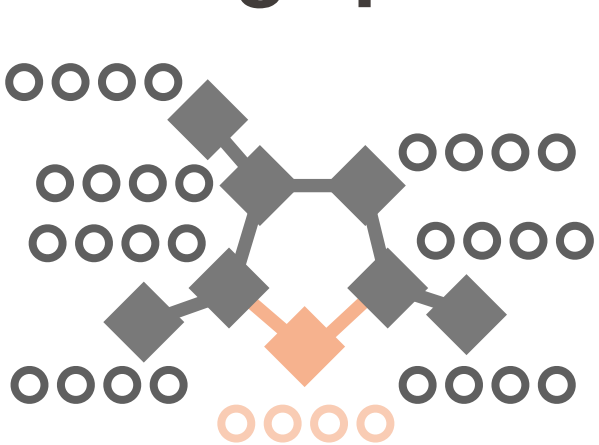


Each node passes its attributes to its neighbours
(according to adjacency matrix) and
learnable update and **interaction function** over the edges

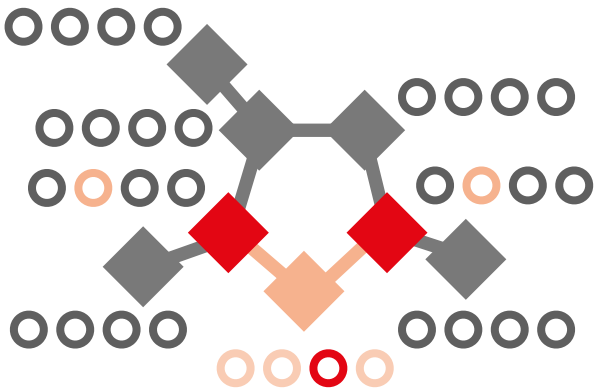
$$\boldsymbol{h}_i^{(l+1)} = f_{\text{update}}\left(\boldsymbol{h}_i^{(l)}, \sum_{j \in \mathcal{N}_i} f_{\text{int}}\left[\boldsymbol{h}_j^{(l)}, \boldsymbol{e}_{(ij)}^{(l)}\right]\right)$$

K. Schütt, O. Unke, M. Gastegger, PMLR 139, 9377 (2021)

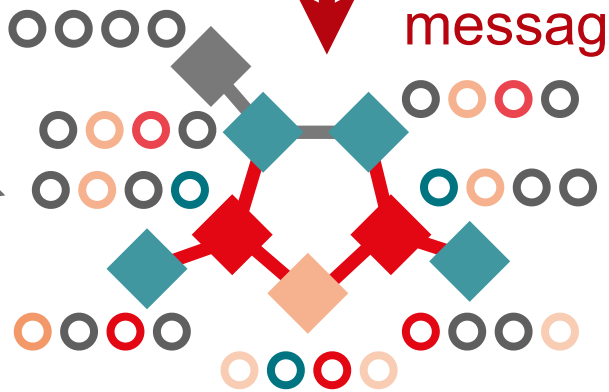
Message passing



first message pass



second message pass



Information from nearest and next nearest neighbour arrives.

The *embedding* learns the atom's local environment.

K. Schütt, O. Unke, M. Gastegger, PMLR 139, 9377 (2021)

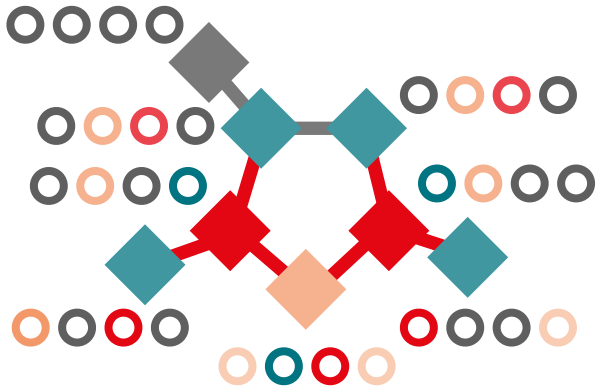
Message passing

Message passing builds up the embeddings,
i.e. local environments, of each atom.

These embeddings are our atomic descriptors.

In descriptor based NN, embeddings were mostly input.
Now they are mostly learned.

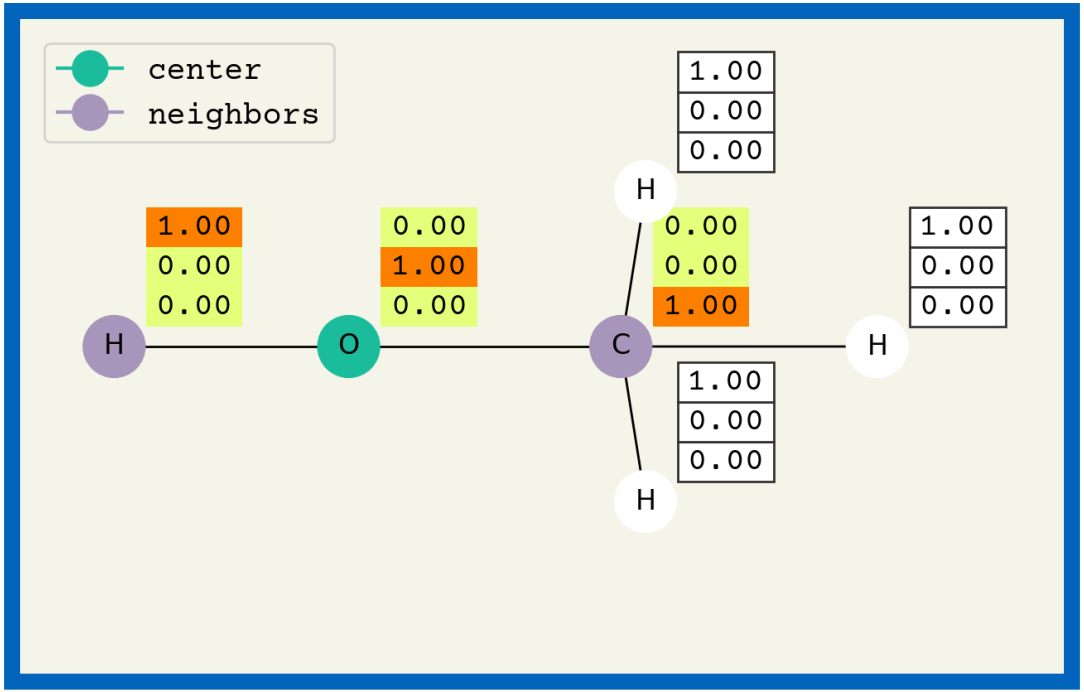
Same approach as before
to get the energy/forces etc.



K. Schütt, O. Unke, M. Gastegger, PMLR 139, 9377 (2021)

A very simple graph neural network

Node features: one-hot based on the element

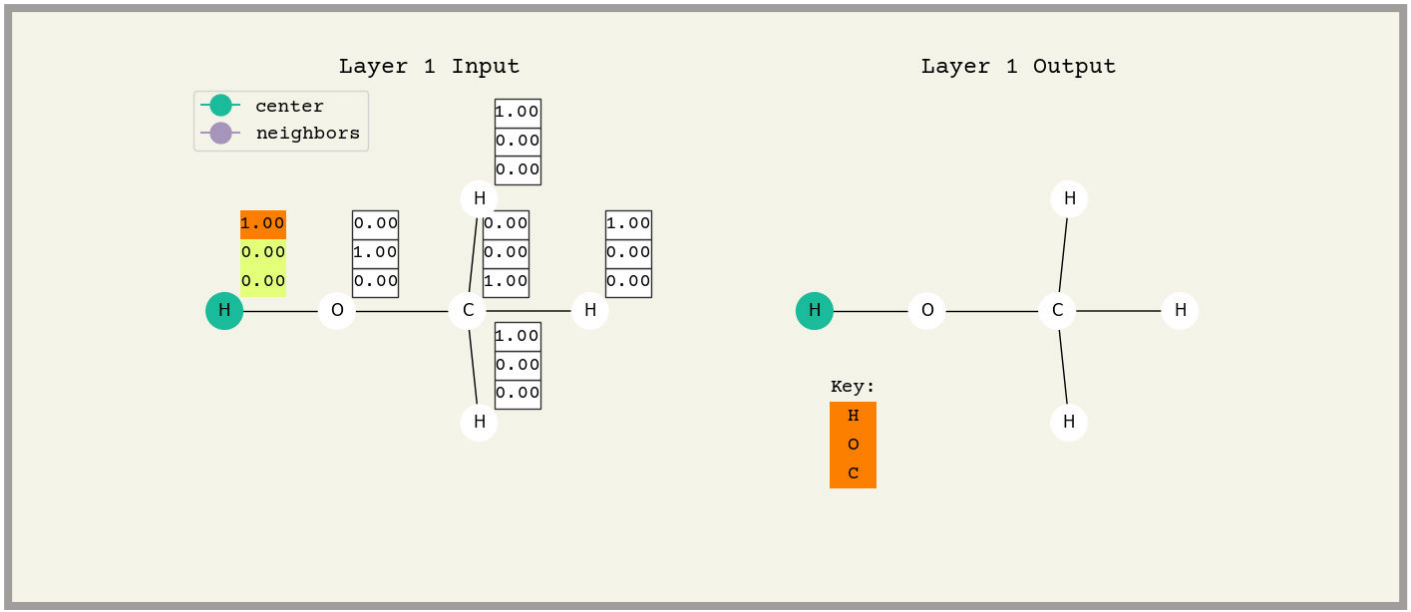


https://dmol.pub/dl/gnn.html

A very simple graph neural network

Node features: one-hot based on the element

Inference: Update via the averaging over neighbors



https://dmol.pub/dl/gnn.html

How successful are GNNs models

Matbench Discovery

Model	CPS ↑	Acc ↑	F1 ↑	DAF ↑	Prec ↑	MAE ↓	R^2 ↑	κ_{SRME} ↓	RMSD ↓	Training Set	Params	Targets	Date Added	r_{cut}	Org
eSEN-30M-OAM	0.888	0.977	0.925	6.069	0.928	0.018	0.866	0.170	0.061	6.6M (113M) OMat24+MPtrj+sAlex	30.2M	EFS_G	2025-03-17	6 Å	
Nequip-OAM-L	0.870	0.967	0.893	5.823	0.890	0.022	0.865	0.166	0.065	6.6M (113M) OMat24+sAlex+MPtrj	9.6M	EFS_G	2025-09-08	6 Å	
ORB v3	0.861	0.971	0.905	5.912	0.904	0.024	0.821	0.210	0.075	6.47M (133M) MPtrj+Alex+OMat24	25.5M	EFS_G	2025-04-05	Å	
SevenNet-MF-ompa	0.845	0.969	0.901	5.825	0.890	0.021	0.867	0.317	0.064	6.6M (113M) OMat24+sAlex+MPtrj	25.7M	EFS_G	2025-03-13	6 Å	
Allegro-OAM-L	0.840	0.966	0.895	5.674	0.867	0.022	0.868	0.319	0.065	6.6M (113M) OMat24+sAlex+MPtrj	9.7M	EFS_G	2025-09-08	7 Å	
GRACE-2L-OAM	0.837	0.963	0.880	5.774	0.883	0.023	0.862	0.294	0.067	6.6M (113M) OMat24+sAlex+MPtrj	12.6M	EFS_G	2025-02-06	6 Å	ICAMS
DPA-3.1-3M-FT	0.802	0.963	0.884	5.667	0.866	0.023	0.869	0.469	0.069	163M OpenLAM	3.27M	EFS_G	2025-06-05	6 Å	AISI
eSEN-30M-MP	0.797	0.946	0.831	5.260	0.804	0.033	0.822	0.340	0.075	146k (1.58M) MPtrj	30.1M	EFS_G	2025-03-17	6 Å	
MACE-MPA-0	0.795	0.954	0.852	5.582	0.853	0.028	0.842	0.412	0.073	3.37M (12M) MPtrj+sAlex	9.06M	EFS_G	2024-12-09	6 Å	

Ususally models trained on very large datasets

The field moves very fast

Big tech has entered your game

The Open Catalyst 2025 (OC25) Dataset and Models for Solid-Liquid Interfaces

Meta

Sushree Jagriti Sahoo[1], **Mikael Maraschin**[2], **Daniel S. Levine**[1], **Zachary Ulissi**[1], **C. Lawrence Zitnick**[1], **Joel B Varley**[4], **Joseph A. Gauthier**[2,†], **Nitish Govindarajan**[3,4†], **Muhammed Shuaibi**[1,†]

[1]FAIR at Meta, [2]Department of Chemical Engineering, Texas Tech University, Lubbock, TX 79409, USA, [3]School of Chemistry, Chemical Engineering and Biotechnology, Nanyang Technological University, 21 Nanyang Link, Singapore 637371, Singapore, [4]Materials Science Division, Lawrence Livermore National Laboratory, Livermore, CA 94550, USA

[†]Co-corresponding Author

SimpleFold: Folding Proteins is Simpler than You Think

Yuyang Wang, Jiarui Lu*, Navdeep Jaitly, Josh Susskind, Miguel Angel Bautista

Apple

Protein folding models have achieved groundbreaking results typically via a combination of integrating domain knowledge into the architectural blocks and training pipelines. Nonetheless, given the success of generative models across different but related problems, it is natural to question whether these architectural designs are a necessary condition to build performant models. In this paper, we introduce *SimpleFold, the first flow-matching based protein folding model that solely uses general purpose transformer blocks*. Protein folding models typically employ computationally expensive modules involving triangular updates, explicit pair representations or multiple training objectives curated for this specific domain.

ByteDance | Seed

Designing quantum chemistry algorithms with just-in-time compilation

Xiaojie Wu[1,†], **Qiming Sun**[1], **Yuanheng Wang**[1]

[1]ByteDance Seed

[†]Corresponding authors

MatterSim: A Deep Learning Atomistic Model Across Elements, Temperatures and Pressures

Han Yang[1*†], Chenxi Hu[1†], Yichi Zhou[1†], Xixian Liu[1†], Yu Shi[1†], Jielan Li[1*†], Guanzhi Li[1†], Zekun Chen[1†], Shuizhou Chen[1†], Claudio Zeni[1], Matthew Horton[1], Robert Pinsler[1], Andrew Fowler[1], Daniel Zügner[1], Tian Xie[1], Jake Smith[1], Lixin Sun[1], Qian Wang[1], Lingyu Kong[1], Chang Liu[1], Hongxia Hao[1*], Ziheng Lu[1*]

Microsoft

*Corresponding author(s). E-mail(s): hanyang@microsoft.com; jielanli@microsoft.com; hongxiahao@microsoft.com; zihenglu@microsoft.com;

[†]These authors contributed equally to this work.

End-to-end NNPs - advantages

1. **invariant & equivariant:**
   energy invariance and force equivariance built in
2. **flexible functional form**
   representations are learned internally
3. **locality**
   allows to be trained on smaller structures, but applied to larger, more complex ones
4. **accuracy**
   significantly more accurate than conventional force fields
5. **versatility**
   works for molecules and periodic structures
6. **efficiency**
   quite data efficient due to complex architectures

End-to-end NNPs - disadvantages

1. **training data:**
   require significantly more training data than conventional force fields, but less than descriptor based NNPs

2. **run time:**
   execution time is higher than descriptor based NNPs

3. **long range:**
   long range effects (like van der Waals or electrostatics) were not included in early generations, but are being added now.

4. **charge:**
   explicit charging was not included in early generations, but is being added now.

Representations for Atomistic Machine Learning

1. Descriptors (attribute or structure based) are *currently* a requirement to perform atomistic machine learning

2. Descriptors need certain properties to be expressive: atom-centered, invariant, equivariant, …

3. In combination with neural networks, descriptors allow the exact modeling of the PES

4. Modern state-of-the-art descriptors are learnable