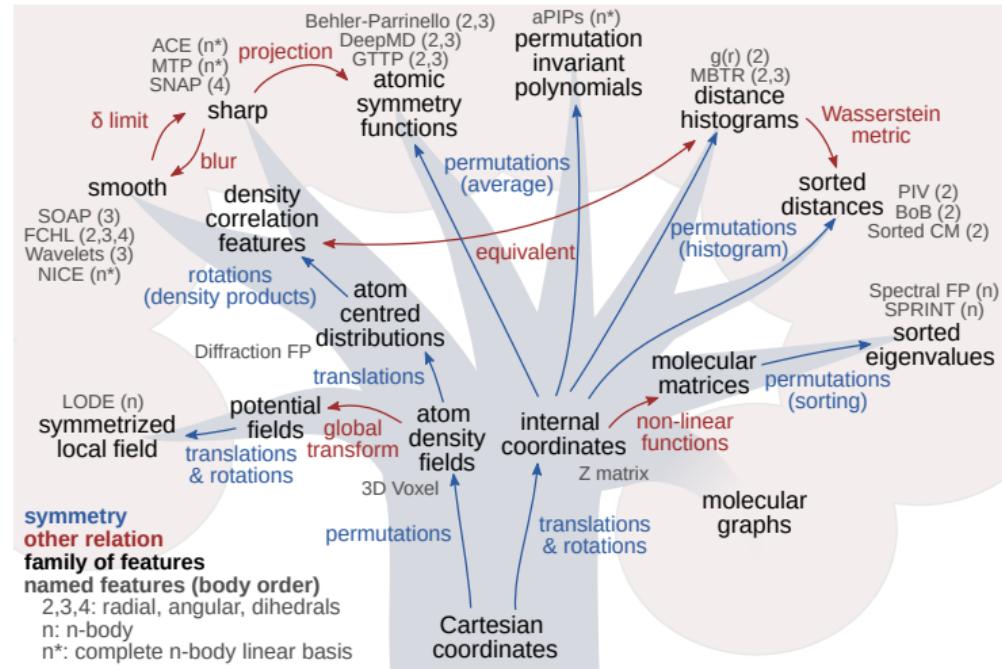


Atom-centered representations for atomistic machine learning

Michele Ceriotti

Outline

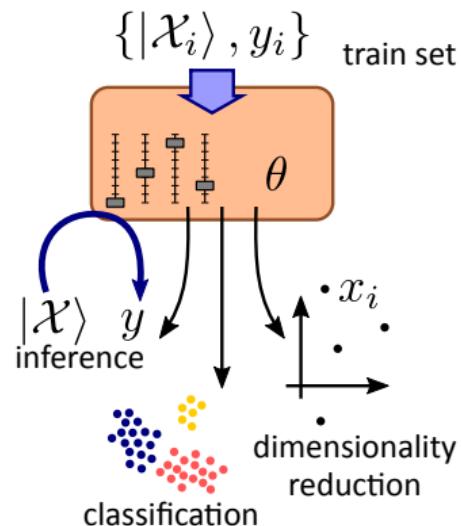
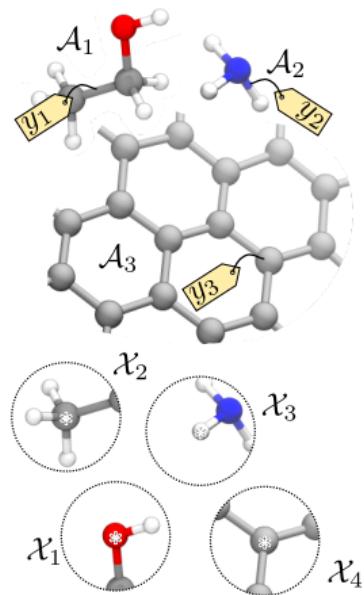
- Let's agree on terminology
- Symmetry and physical constraints
- Density expansion and invariant density correlations
- Feature engineering and feature interpretation



Representations of atomic structures

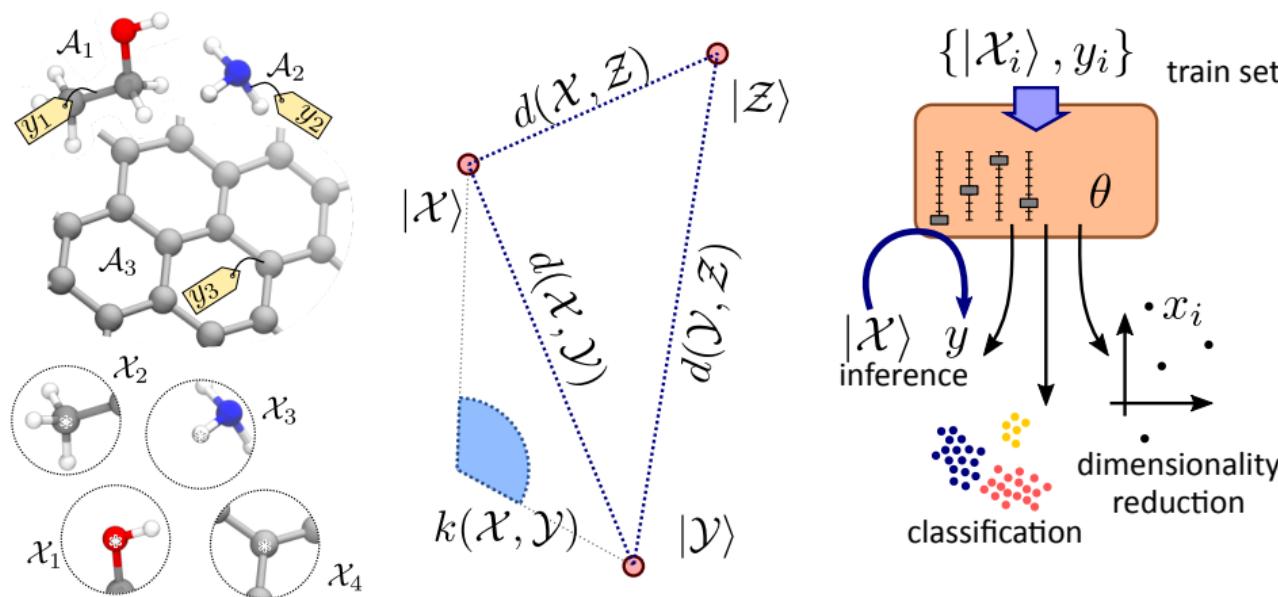
Machine learning at the atomic scale

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



The role of representations

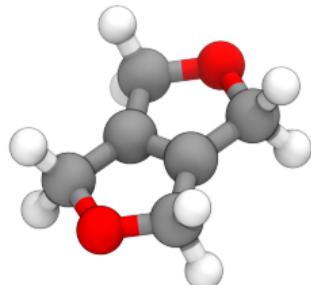
- Structures must be cast to feature vectors $|A\rangle$ / distances $d(A, B)$ / kernels $k(A, B)$ before applying a ML algorithm
- This representation must be concise, complete, and incorporate physical principles



From atoms to properties

- Physical/mathematical requirements are imposed on the structure → property mapping
- Additivity/locality + translation equivariance → atom-centered formalism
- Roto-inversion ($O(3)$) and index permutation → full equivariance

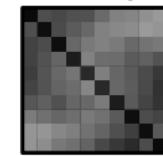
$$A \equiv \{a_i, \mathbf{r}_i\}$$



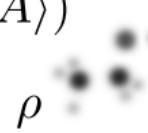
$$|A; \lambda\mu\rangle$$



$$V(\{a_i, \mathbf{r}_i\})$$



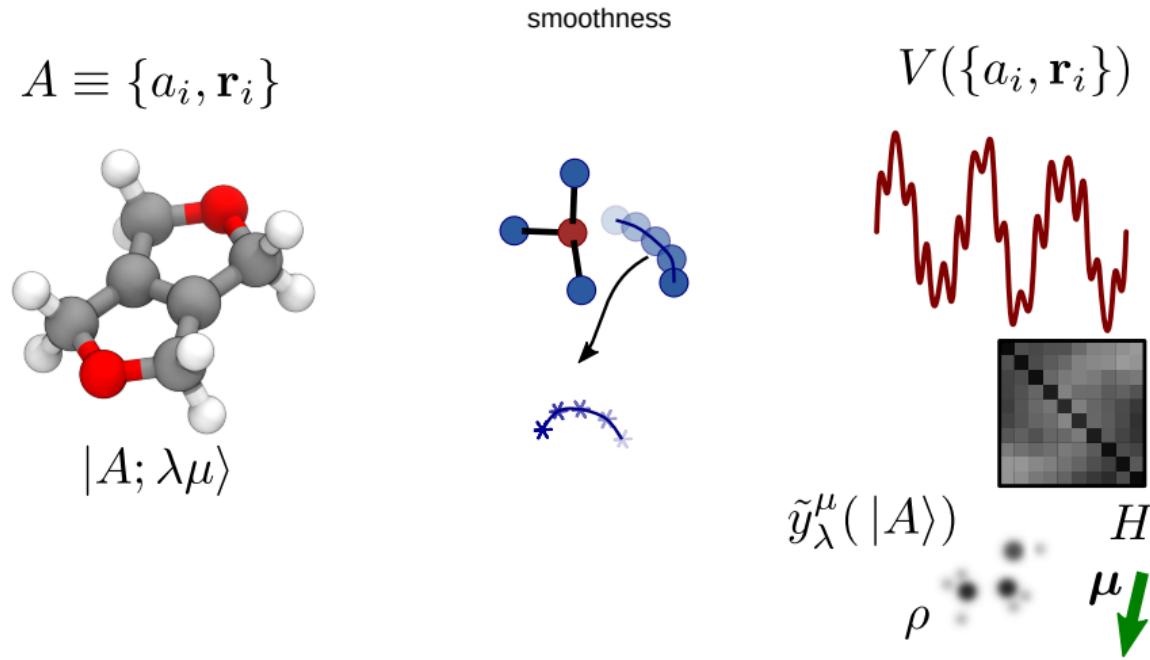
$$\tilde{y}_\lambda^\mu(|A\rangle)$$



$$\mu$$

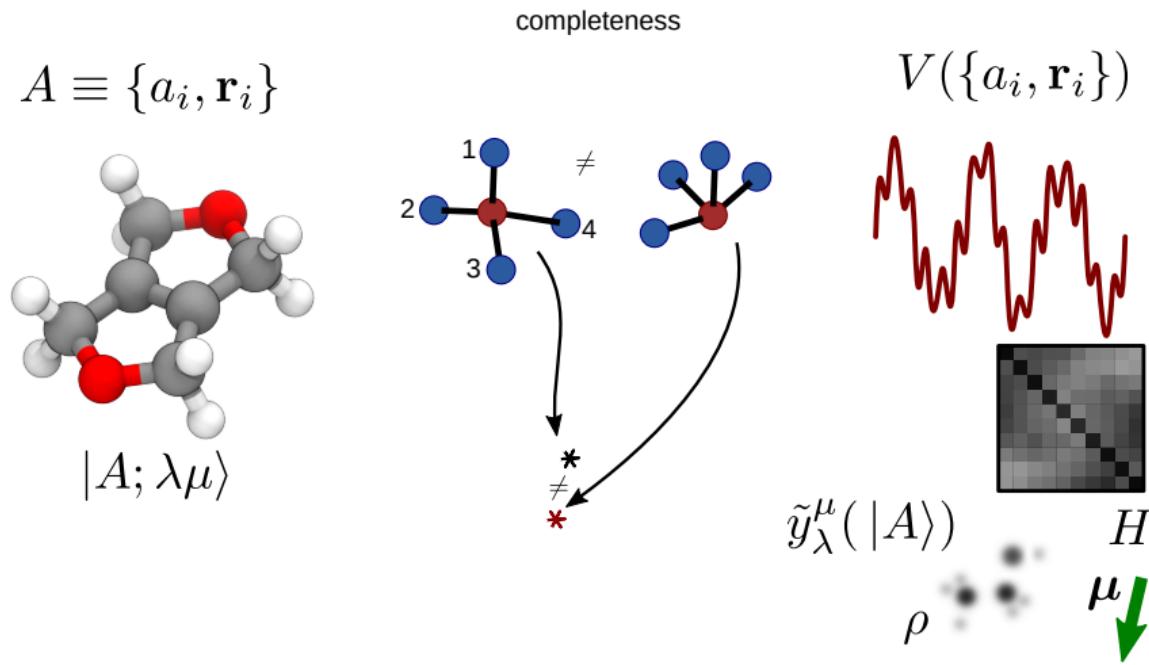

From atoms to properties

- Physical/mathematical requirements are imposed on the structure → property mapping
- Additivity/locality + translation equivariance → atom-centered formalism
- Roto-inversion ($O(3)$) and index permutation → full equivariance



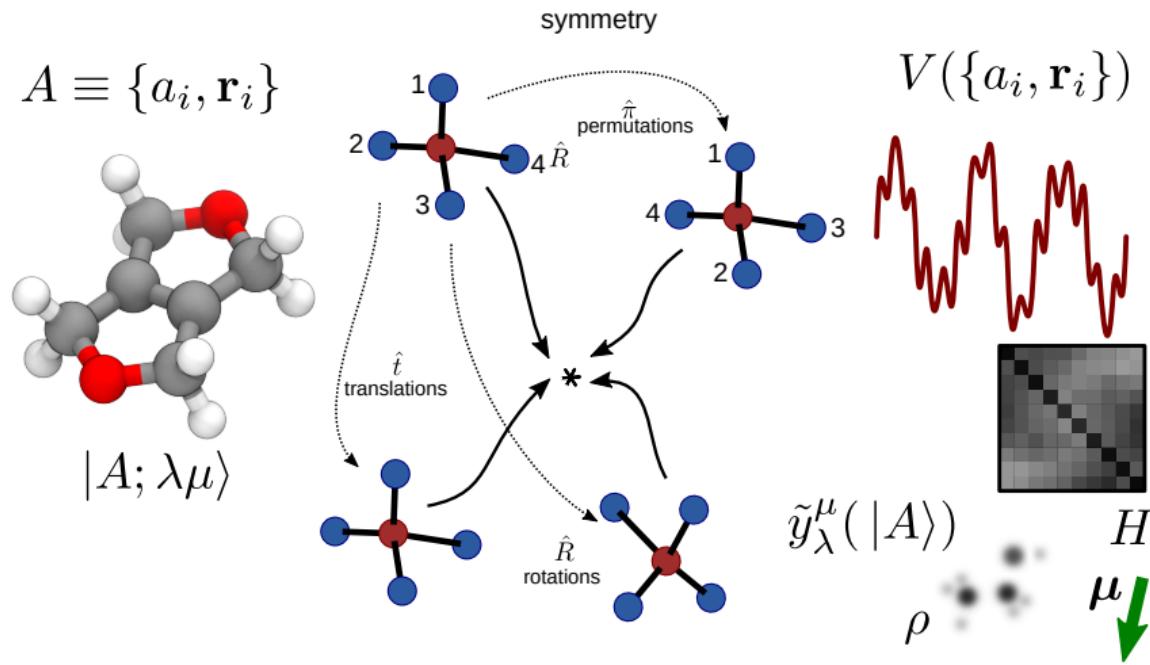
From atoms to properties

- Physical/mathematical requirements are imposed on the structure → property mapping
- Additivity/locality + translation equivariance → atom-centered formalism
- Roto-inversion ($O(3)$) and index permutation → full equivariance



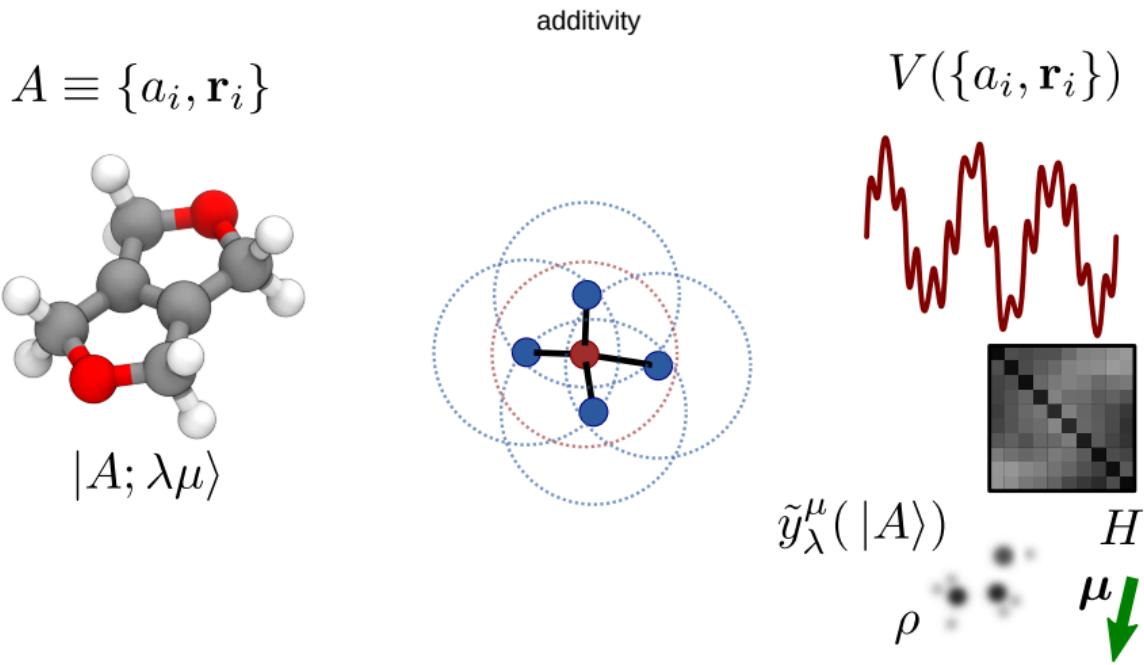
From atoms to properties

- Physical/mathematical requirements are imposed on the structure → property mapping
- Additivity/locality + translation equivariance → atom-centered formalism
- Roto-inversion ($O(3)$) and index permutation → full equivariance



From atoms to properties

- Physical/mathematical requirements are imposed on the structure → property mapping
- Additivity/locality + translation equivariance → atom-centered formalism
- Roto-inversion ($O(3)$) and index permutation → full equivariance



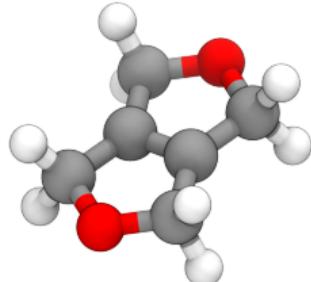
From atoms to properties

- Physical/mathematical requirements are imposed on the structure → property mapping
- Additivity/locality + translation equivariance → atom-centered formalism
- Roto-inversion ($O(3)$) and index permutation → full equivariance

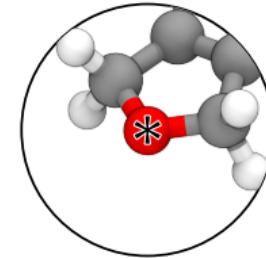
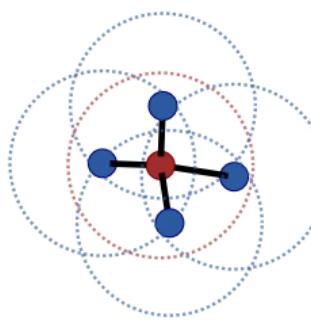
additivity

$$A \equiv \{a_i, \mathbf{r}_i\}$$

$$\mathbf{y}(A) = \sum_i \mathbf{y}(A_i)$$



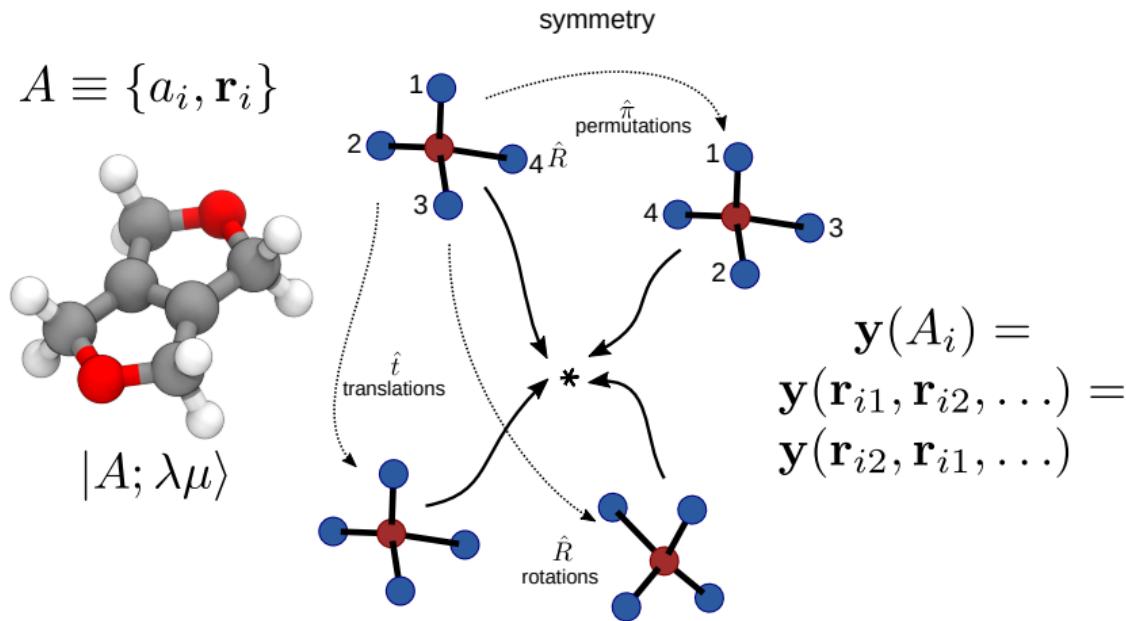
$$|A; \lambda\mu\rangle$$



$$A_i \equiv (a_i, \{a_j, \mathbf{r}_{ij}\})$$

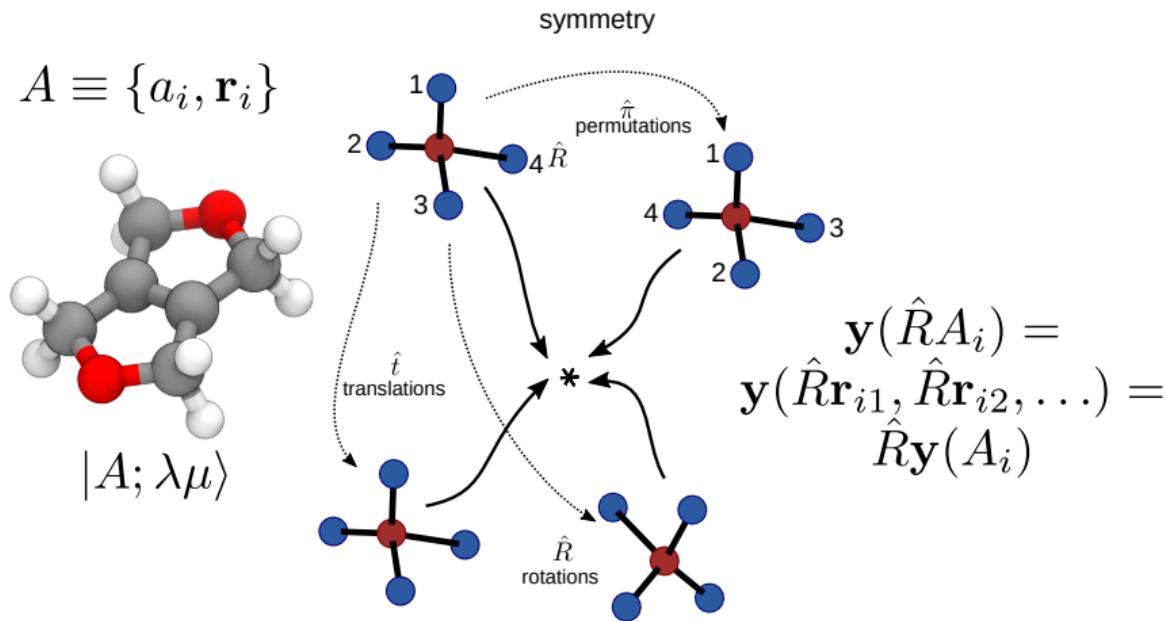
From atoms to properties

- Physical/mathematical requirements are imposed on the structure → property mapping
- Additivity/locality + translation equivariance → atom-centered formalism
- Roto-inversion ($O(3)$) and index permutation → full equivariance



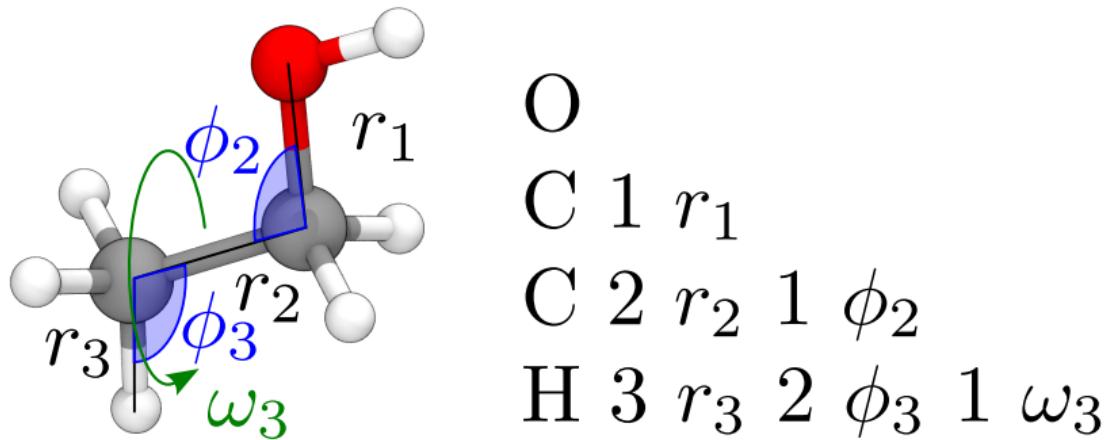
From atoms to properties

- Physical/mathematical requirements are imposed on the structure → property mapping
- Additivity/locality + translation equivariance → atom-centered formalism
- Roto-inversion ($O(3)$) and index permutation → full equivariance



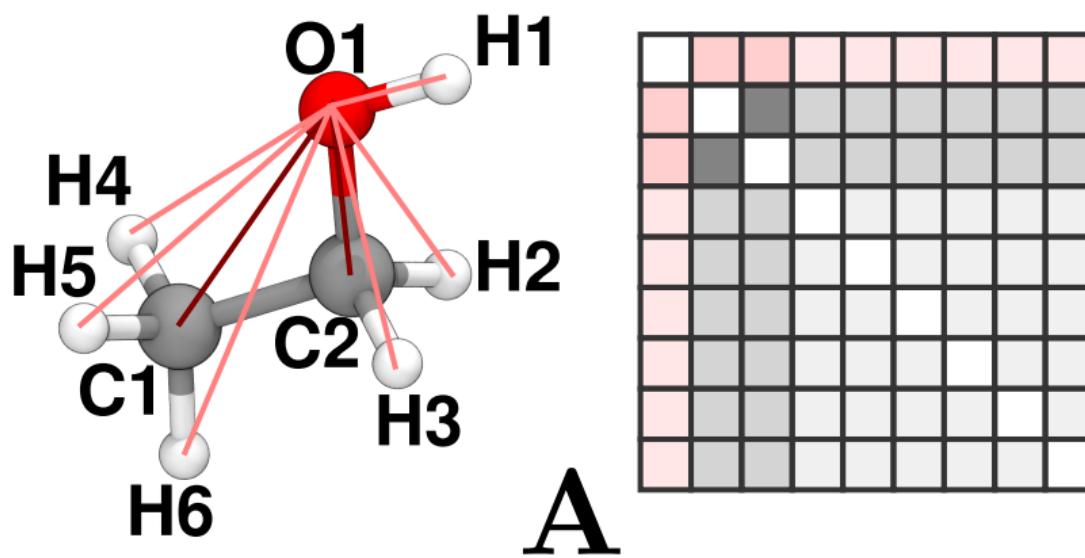
Internal coordinates

- Traditional description in chemistry/cheminformatics: the geometry is translated in a combination of bond lengths, angles, dihedrals (Z-matrix)
- In theory connection with interatomic potentials, based on pairs, triples, torsions
- Non-unique, dependent on atom ordering, and complicated to build and manipulate



Coulomb matrix

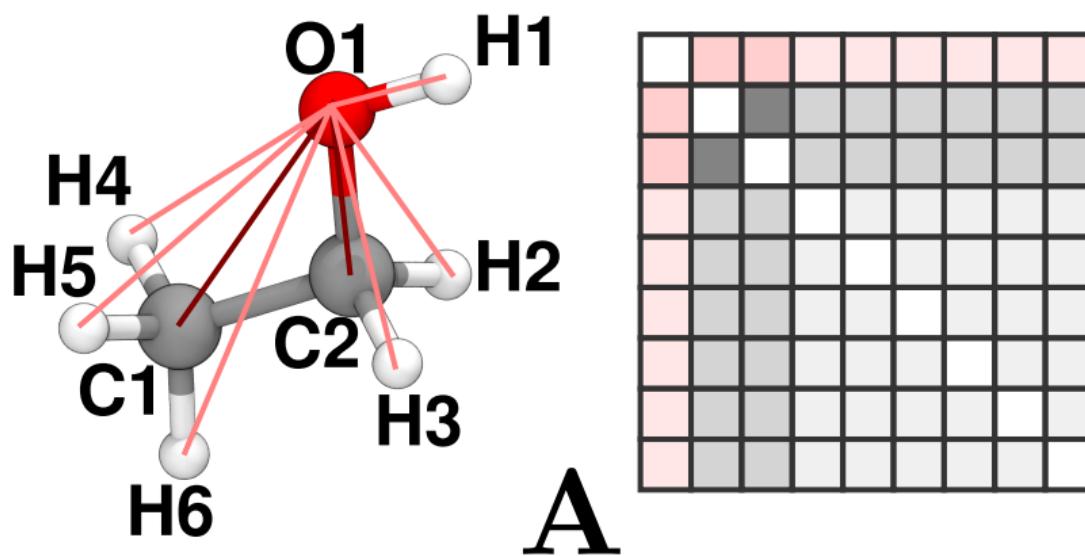
- The matrix of interatomic distances, or the scalar products, provides a full description of a geometry modulo rotations and translation
- Coulomb matrix uses $Z_i Z_j / |\mathbf{r}_i - \mathbf{r}_j|$ to mimic physical interactions
- Depends on atom ordering; several attempts to make invariant by sorting, computing eigenspectrum, random averaging, ...



Rupp et al., PRL 2012

Coulomb matrix

- The matrix of interatomic distances, or the scalar products, provides a full description of a geometry modulo rotations and translation
- Coulomb matrix uses $Z_i Z_j / |\mathbf{r}_i - \mathbf{r}_j|$ to mimic physical interactions
- Depends on atom ordering; several attempts to make invariant by sorting, computing eigenspectrum, random averaging, ...



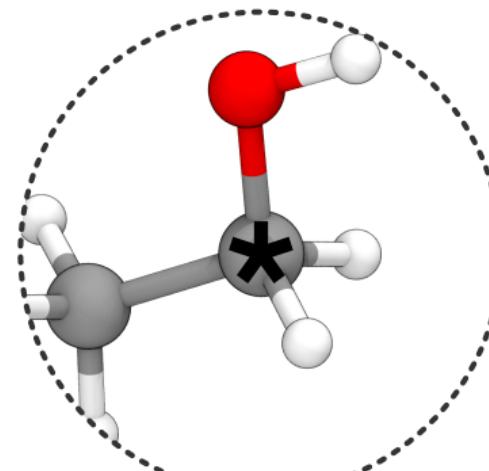
Hansen et al. JCTC (2013); Pietrucci and Andreoni PRL (2011); Sadeghi et al. JCP (2013)

Atom-centered symmetry functions

- Atom-centered construction, summing over tuples of neighbors

$$G_k^{(2)}(A_i) = \sum_j g_k^{(2)}(r_{ij}), \quad G_k^{(3)}(A_i) = \sum_{jj'} g_k^{(3)}(r_{ij}, r_{ij'}, \hat{\mathbf{r}}_{ij} \cdot \hat{\mathbf{r}}_{ij'})$$

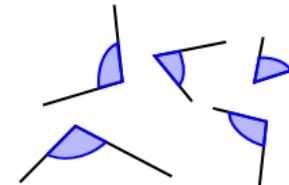
- Invariant to rotations, translations and atom ordering
- Widely used with much success, despite fundamental shortcomings (finite-order, non-orthogonal choice of functions, poor scaling with number of elements and body order)



$$A_i = \{a_i, (a_j, \mathbf{r}_{ij})\}$$

A vertical stack of five horizontal bars of decreasing length from top to bottom, representing a 1D array of distances. An arrow points from this array to the equation.

$$g_k^{(2)}(r) \rightarrow G_k^{(2)}(A_i)$$

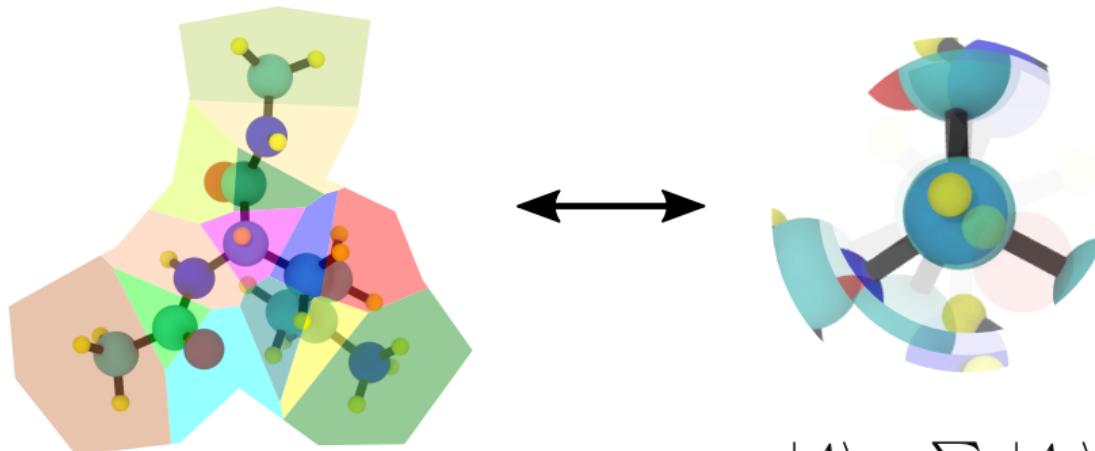


$$g_k^{(3)}(r_{ij}, r_{ij'}, \hat{\mathbf{r}}_{ij} \cdot \hat{\mathbf{r}}_{ij'}) \rightarrow G_k^{(3)}(A_i)$$

Behler and Parrinello (2007); Smith et al. Chem. Sci. (2017); Zhang et al. PRL (2018)

Additivity and locality

- Atom-centered representations are key to transferable models for chemistry



$$V(A) = \sum_{i \in A} V(A_i)$$

$$|A\rangle = \sum_i |A_i\rangle$$
$$K(A, B) = \sum_{i,j} k(A_i, B_j)$$

Additivity and locality

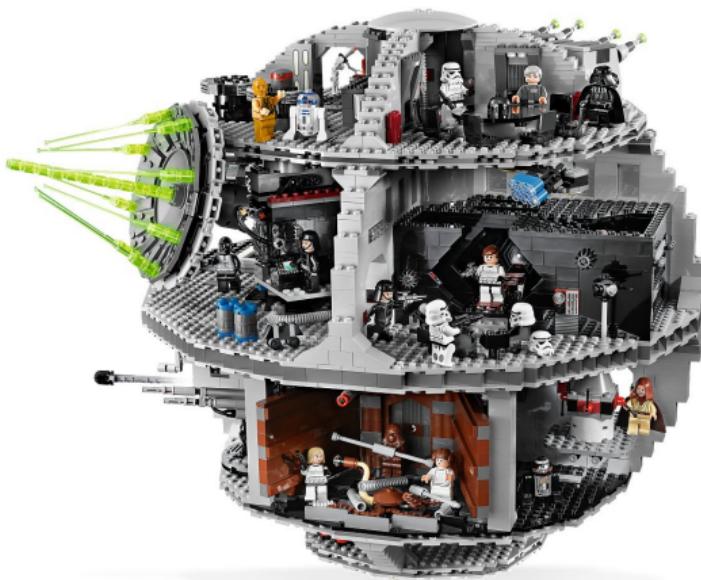
- Atom-centered representations are key to transferable models for chemistry



Prodhān, Kohn, PNAS (2005)

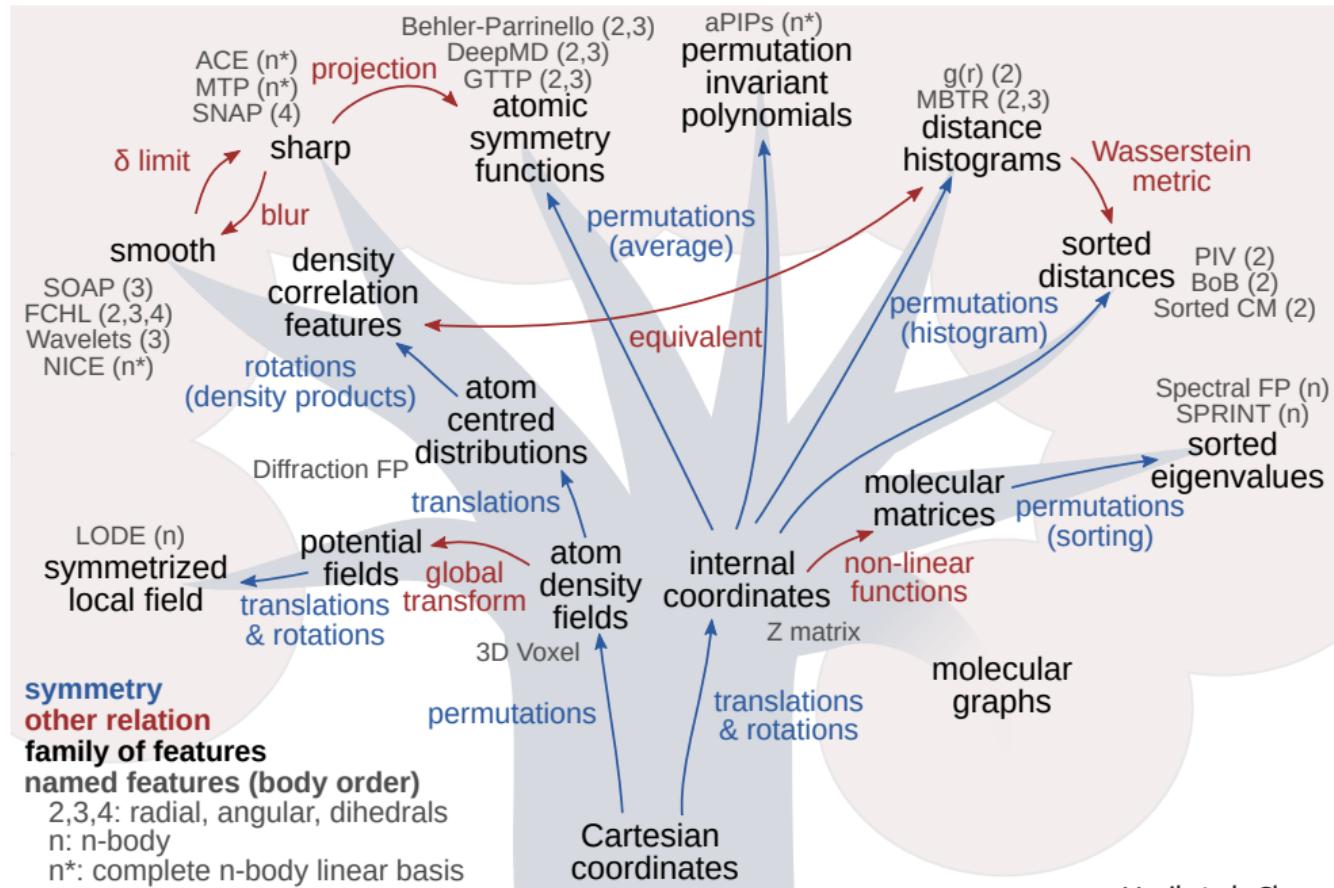
Additivity and locality

- Atom-centered representations are key to transferable models for chemistry



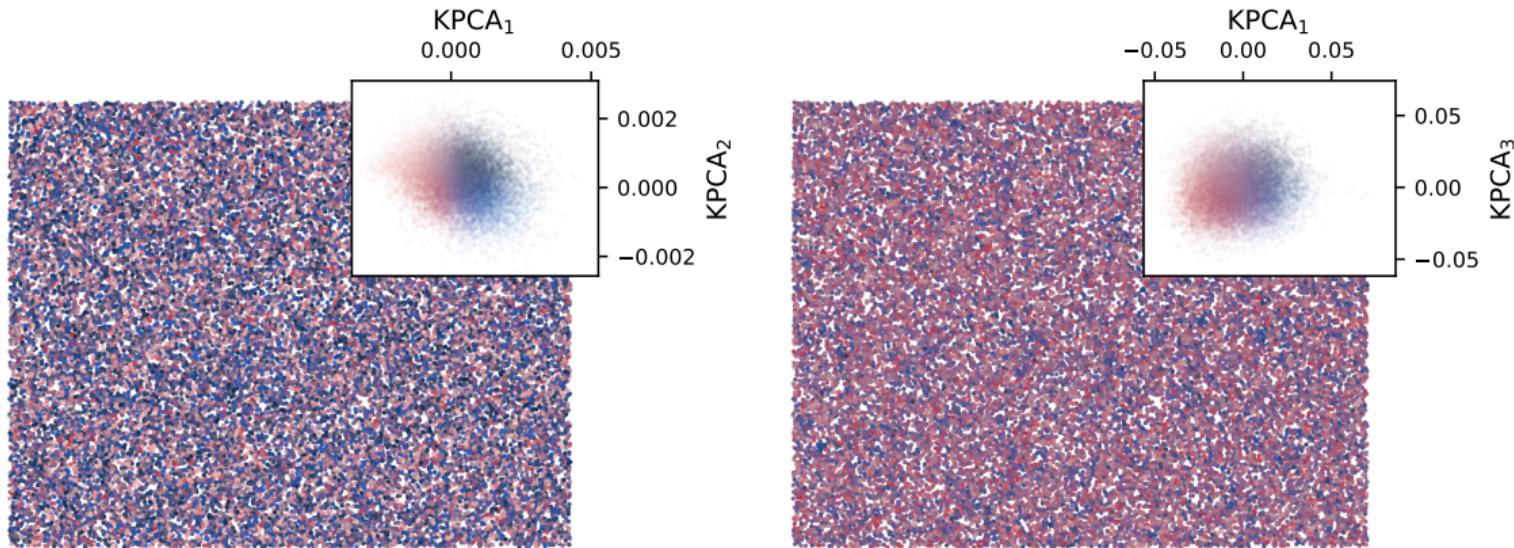
Prodhān, Kohn, PNAS (2005)

A phylogenetic tree of ML representations



What you ask is what you get

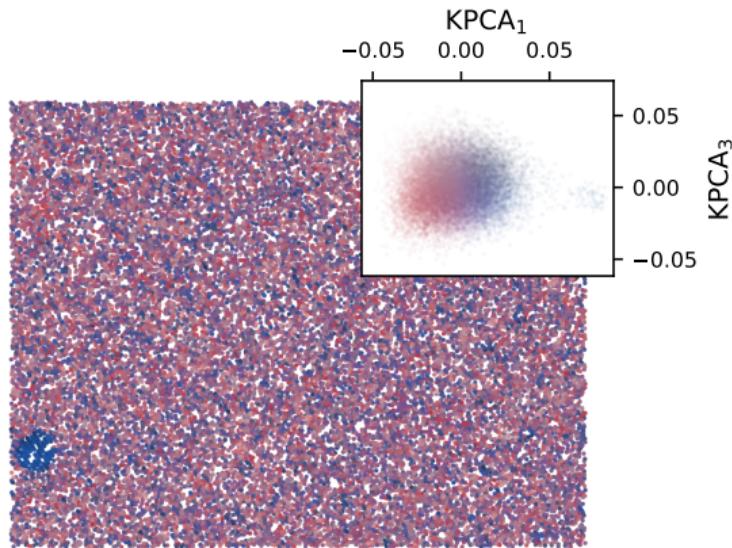
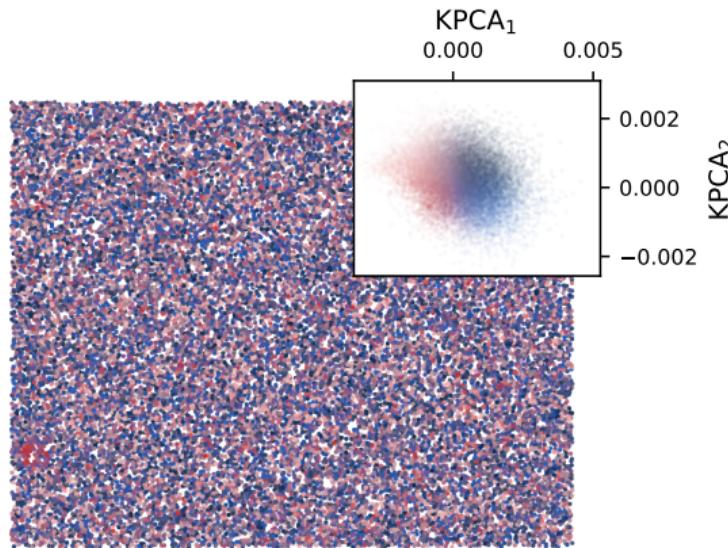
- Understanding what goes into a representation is key to achieve meaningful results from automated data analytics
- Example: you don't *always* want to have rotational invariance



data: Shibuta, Sakane, Takaki, Ohno, Acta Mat. (2016)

What you ask is what you get

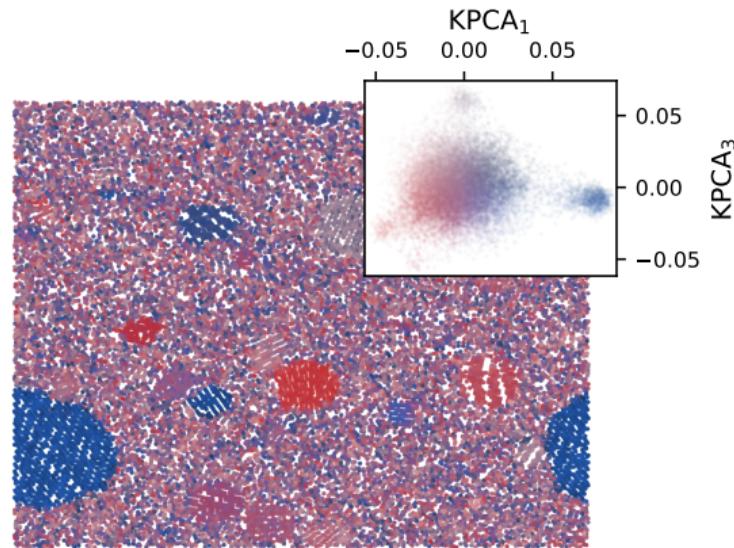
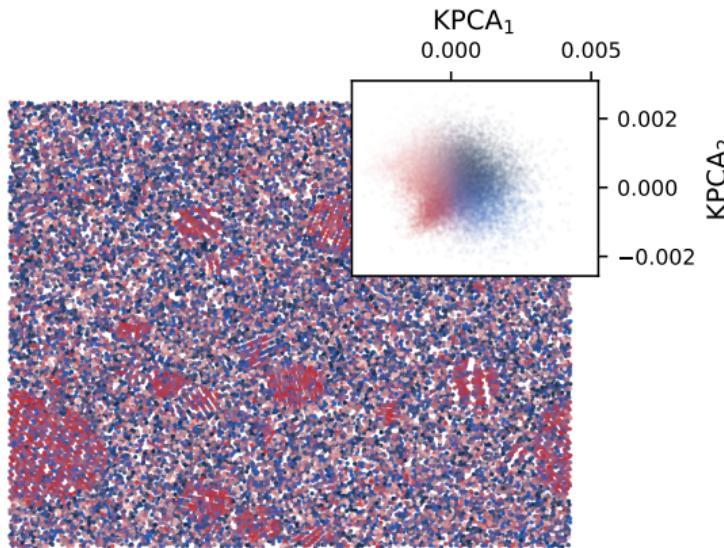
- Understanding what goes into a representation is key to achieve meaningful results from automated data analytics
- Example: you don't *always* want to have rotational invariance



data: Shibuta, Sakane, Takaki, Ohno, Acta Mat. (2016)

What you ask is what you get

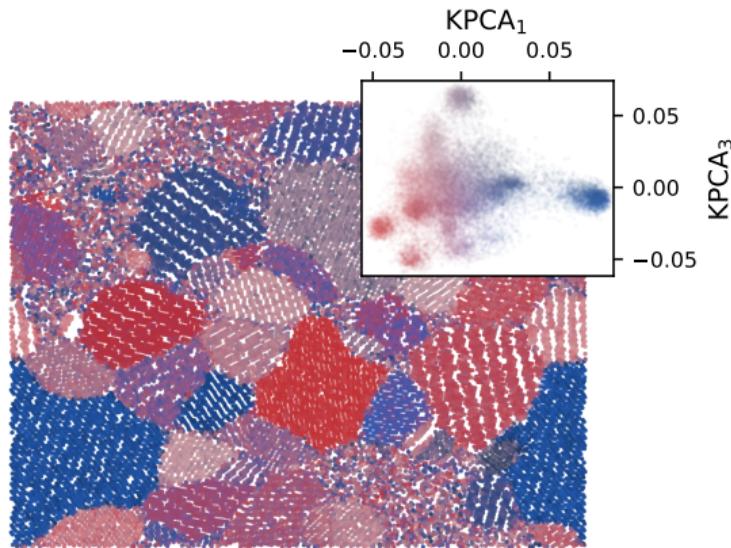
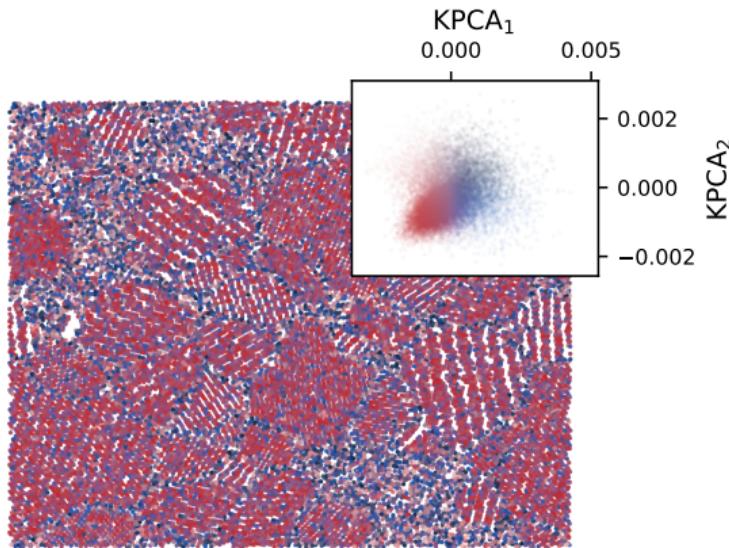
- Understanding what goes into a representation is key to achieve meaningful results from automated data analytics
- Example: you don't *always* want to have rotational invariance



data: Shibuta, Sakane, Takaki, Ohno, Acta Mat. (2016)

What you ask is what you get

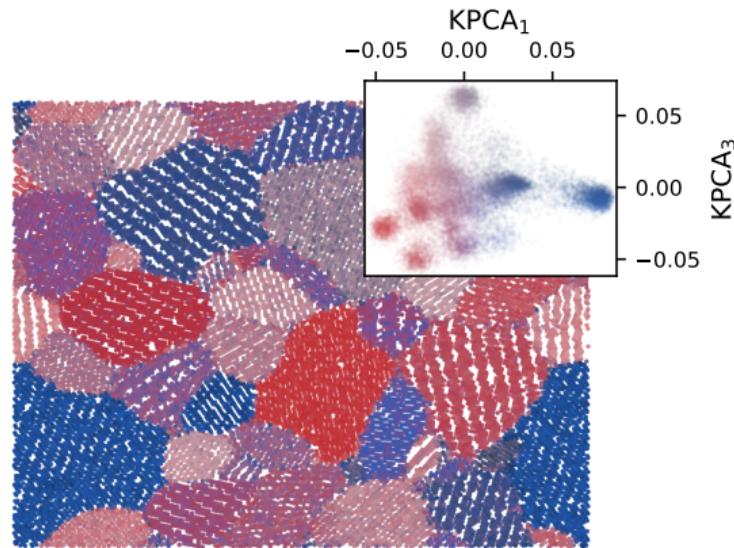
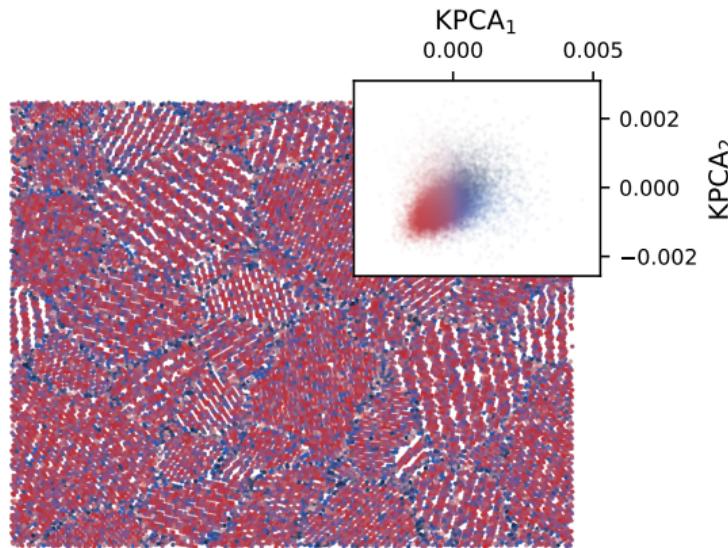
- Understanding what goes into a representation is key to achieve meaningful results from automated data analytics
- Example: you don't *always* want to have rotational invariance



data: Shibuta, Sakane, Takaki, Ohno, Acta Mat. (2016)

What you ask is what you get

- Understanding what goes into a representation is key to achieve meaningful results from automated data analytics
- Example: you don't *always* want to have rotational invariance

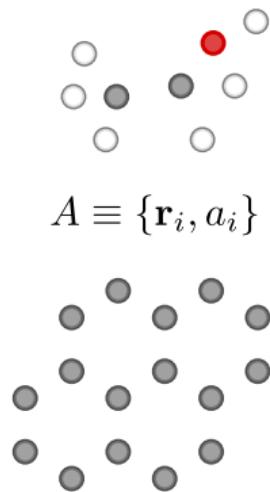
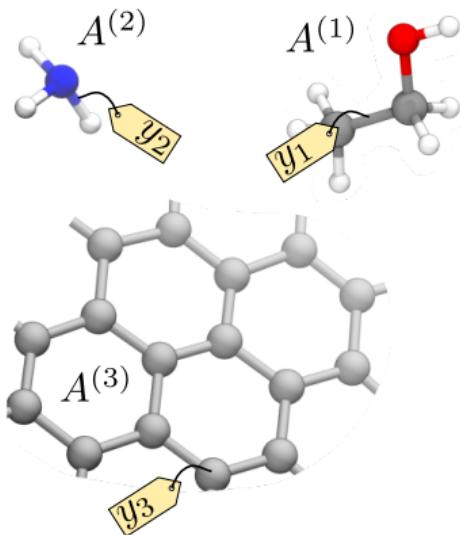


data: Shibuta, Sakane, Takaki, Ohno, Acta Mat. (2016)

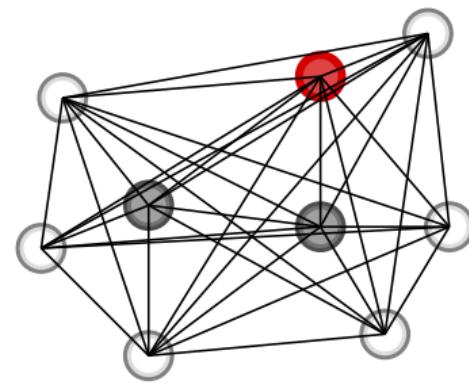
Atom density representations

Notation and nomenclature

- Molecules and materials as point clouds (or graphs), decorated by the atom type
- Will use a different language, more reminiscent of statistical mechanics



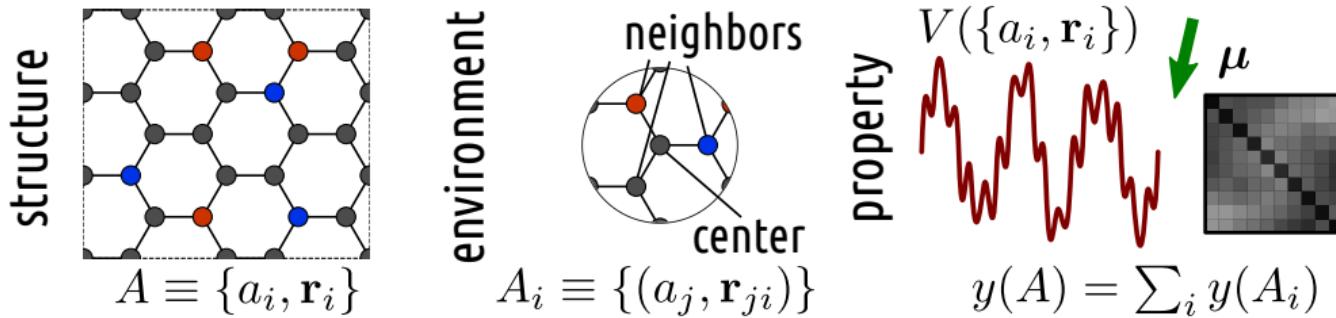
$$A \equiv \{\mathbf{r}_i, a_i\}$$



$$A \equiv \{a_i, \mathbf{r}_{ji}\}$$

Notation and nomenclature

- Molecules and materials as point clouds (or graphs), decorated by the atom type
- Will use a different language, more reminiscent of statistical mechanics



Notation and nomenclature

- Molecules and materials as point clouds (or graphs), decorated by the atom type
- Will use a different language, more reminiscent of statistical mechanics

structure		$A \equiv \{a_i, \mathbf{r}_i\}$	environment		$A_i \equiv \{(a_j, \mathbf{r}_{ji})\}$	property		$V(\{a_i, \mathbf{r}_i\})$	μ
representations	$\xi_q(A_i)$	neighbor density		$\rho_i^a(\mathbf{x}) = \sum_{j \in A_i} \delta_{aa_j} g(\mathbf{x} - \mathbf{r}_{ji})$	$\langle a\mathbf{x} \rho_i \rangle \equiv \sum_{j \in A_i} \delta_{aa_j} \langle \mathbf{x} \mathbf{r}_{ji} \rangle$				
	$\langle q A_i \rangle$	density coefficients		$c_{i;nlm}^a \equiv \int d\mathbf{x} R_{nl}(x) Y_m^l(\hat{\mathbf{x}}) \rho_i^a(\mathbf{x})$	$\langle anlm \rho_i \rangle \equiv \sum_{j \in A_i} \delta_{aa_j} \langle nlm \mathbf{r}_{ji} \rangle$				

A Dirac notation for ML representations

$$\langle Q | A \rangle$$

features index representation target & nature

radial indices structure correlation order parity

$\langle n_1 l_1; \dots n_\nu l_\nu k_\nu | A; \rho_i^{\otimes \nu}; \sigma; \lambda \mu \rangle$

angular channels field center rot. / symmetry

- A representation maps a structure A (or environment A_i) to a vector
- Continuous (or discrete) feature index Q
- Bra-ket notation $\langle Q | A; \text{rep.} \rangle$ indicates in an abstract way this mapping
- Plenty of room to express the details of a representation
- Notation reflects naturally a change of basis, the construction of a kernel, or a linear model

$$\langle Y | A \rangle = \int dQ \langle Y | Q \rangle \langle Q | A \rangle$$

Willatt, Musil, **MC**, JCP (2019); Musil et al., Chem. Rev. (2021); <https://tinyurl.com/dirac-rep>

A Dirac notation for ML representations

$$\begin{array}{c} \text{features} \\ \text{index} \end{array} \left\langle Q \mid A \right\rangle \begin{array}{c} \text{representation} \\ \text{target \& nature} \end{array}$$
$$\left\langle \begin{array}{c} \text{radial indices} \\ n_1 l_1; \dots n_\nu l_\nu k_\nu \\ \text{angular channels} \end{array} \mid \begin{array}{c} \text{structure} \\ A; \rho_i^{\otimes \nu}; \sigma; \lambda \mu \\ \text{field center} \end{array} \right\rangle \begin{array}{c} \text{correlation} \\ \text{order} \\ \text{parity} \\ \hline \text{rot. /} \\ \text{symmetry} \end{array}$$

- A representation maps a structure A (or environment A_i) to a vector
- Continuous (or discrete) feature index Q
- Bra-ket notation $\langle Q | A; \text{rep.} \rangle$ indicates in an abstract way this mapping
- Plenty of room to express the details of a representation
- Notation reflects naturally a change of basis, the construction of a kernel, or a linear model

$$k(A, A') = \langle A | A' \rangle \approx \int dQ \langle A | Q \rangle \langle Q | A' \rangle$$

Willatt, Musil, **MC**, JCP (2019); Musil et al., Chem. Rev. (2021); <https://tinyurl.com/dirac-rep>

A Dirac notation for ML representations

$$\begin{array}{c} \text{features} \\ \text{index} \end{array} \left\langle Q \mid A \right\rangle \begin{array}{c} \text{representation} \\ \text{target \& nature} \end{array}$$
$$\left\langle \begin{array}{c} \text{radial indices} \\ n_1 l_1; \dots n_\nu l_\nu k_\nu \\ \text{angular channels} \end{array} \mid \begin{array}{c} \text{structure} \\ A; \rho_i^{\otimes \nu}; \sigma; \lambda \mu \\ \text{field center} \end{array} \right\rangle \begin{array}{c} \text{correlation} \\ \text{order} \\ \text{parity} \\ \hline \text{rot. /} \\ \text{symmetry} \end{array}$$

- A representation maps a structure A (or environment A_i) to a vector
- Continuous (or discrete) feature index Q
- Bra-ket notation $\langle Q|A; \text{rep.}\rangle$ indicates in an abstract way this mapping
- Plenty of room to express the details of a representation
- Notation reflects naturally a change of basis, the construction of a kernel, or a linear model

$$E(A) = \langle E|A \rangle \approx \int dQ \langle E|Q \rangle \langle Q|A \rangle$$

Atoms, density and symmetries

- A 1D toy model: compute properties in a permutation and inversion-invariant way

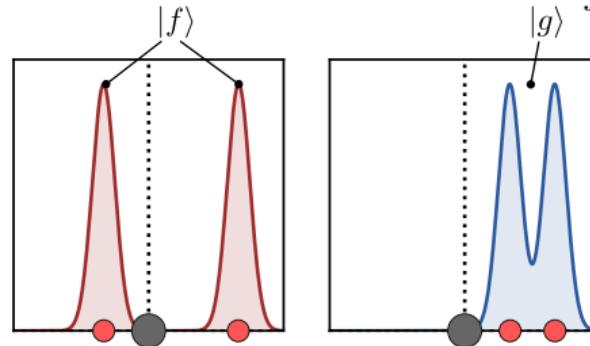
$$y = \tilde{y}(x_1, x_2)$$



Atoms, density and symmetries

- A 1D toy model: compute properties in a permutation and inversion-invariant way

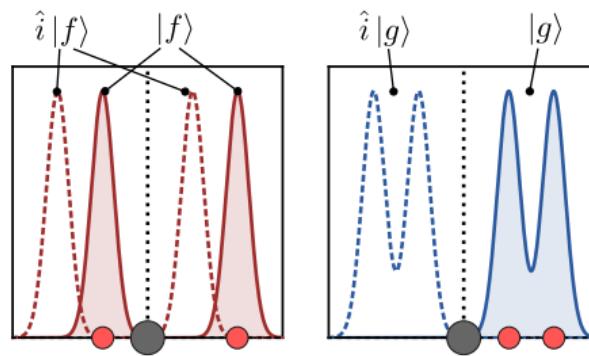
$$y = \tilde{y}^{(1)}(x_1) + \tilde{y}^{(1)}(x_2) = \int dx \tilde{y}^{(1)}(x) [\delta(x - x_1) + \delta(x - x_2)] = \int dx \tilde{y}^{(1)}(x) \langle x | f \rangle \quad \text{density trick}$$



Atoms, density and symmetries

- A 1D toy model: compute properties in a permutation and inversion-invariant way

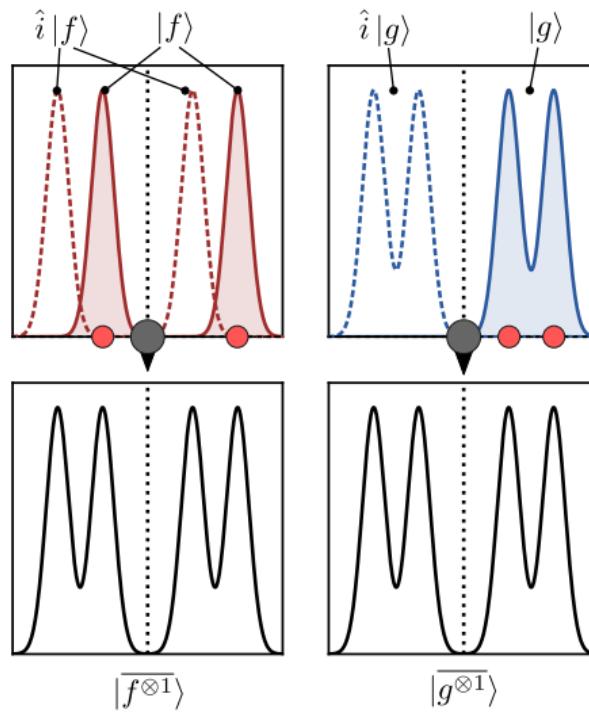
$$y = \int dx \tilde{y}^{(1)}(x) [\langle x|f\rangle + \langle -x|f\rangle] \quad \text{Haar integration}$$



Atoms, density and symmetries

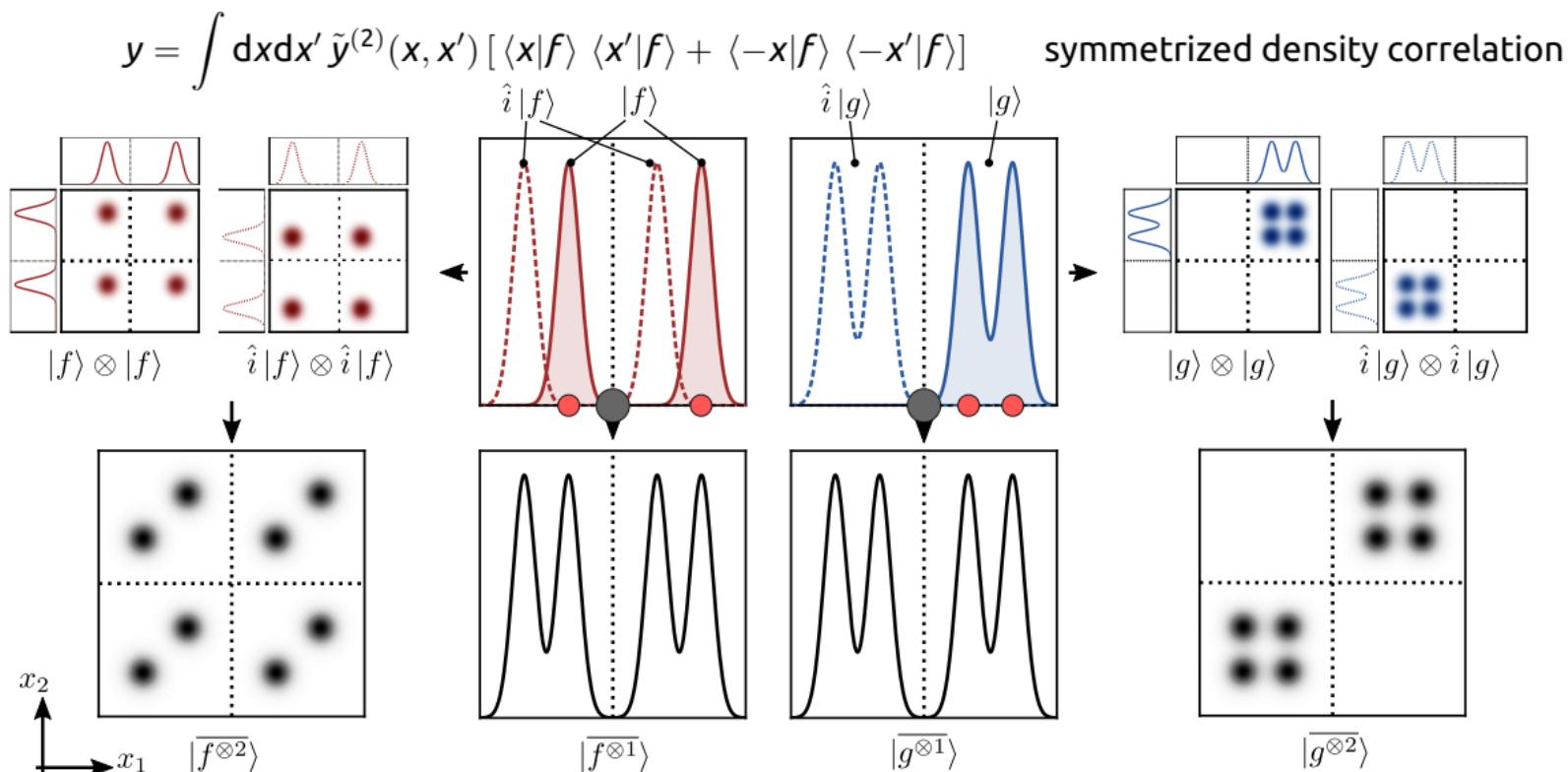
- A 1D toy model: compute properties in a permutation and inversion-invariant way

$$y = \int dx \tilde{y}^{(1)}(x) [\langle x|f\rangle + \langle -x|f\rangle] \quad \text{Haar integration}$$



Atoms, density and symmetries

- A 1D toy model: compute properties in a permutation and inversion-invariant way



Atom-Centered Density Correlations

- Any permutation-invariant function of neighbor positions can be linearly expanded in terms of tensor products of the neighbor density
- Symmetrized correlations \leftrightarrow invariant models

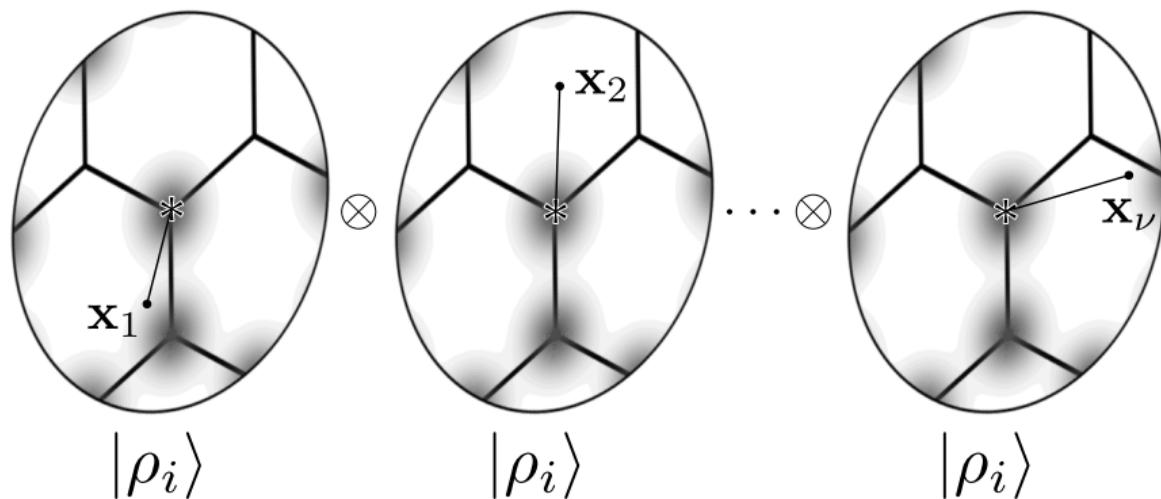
$$y(A_i) \approx \sum_{i_1 i_2 \in A_i} \tilde{y}(\mathbf{r}_{ii_1}, \mathbf{r}_{ii_2}) = \sum_{i_1 i_2 \in A_i} \int d\mathbf{x}_1 d\mathbf{x}_2 \tilde{y}(\mathbf{x}_1, \mathbf{x}_2) \delta(\mathbf{x}_1 - \mathbf{r}_{ii_1}) \delta(\mathbf{x}_2 - \mathbf{r}_{ii_2})$$

$$= \int d\mathbf{x}_1 d\mathbf{x}_2 \tilde{y}(\mathbf{x}_1, \mathbf{x}_2) \rho_i(\mathbf{x}_1) \rho_i(\mathbf{x}_2) = \int d\mathbf{x}_1 d\mathbf{x}_2 \tilde{y}(\mathbf{x}_1, \mathbf{x}_2) \langle \mathbf{x}_1 | A_i \rangle \langle \mathbf{x}_2 | A_i \rangle$$

$$y(A_i) \approx \sum_{q_1 q_2} y_{q_1 q_2} \langle q_1 | A_i \rangle \langle q_2 | A_i \rangle$$

Atom-Centered Density Correlations

- Any permutation-invariant function of neighbor positions can be linearly expanded in terms of tensor products of the neighbor density
- Symmetrized correlations \leftrightarrow invariant models



Atom-Centered Density Correlations

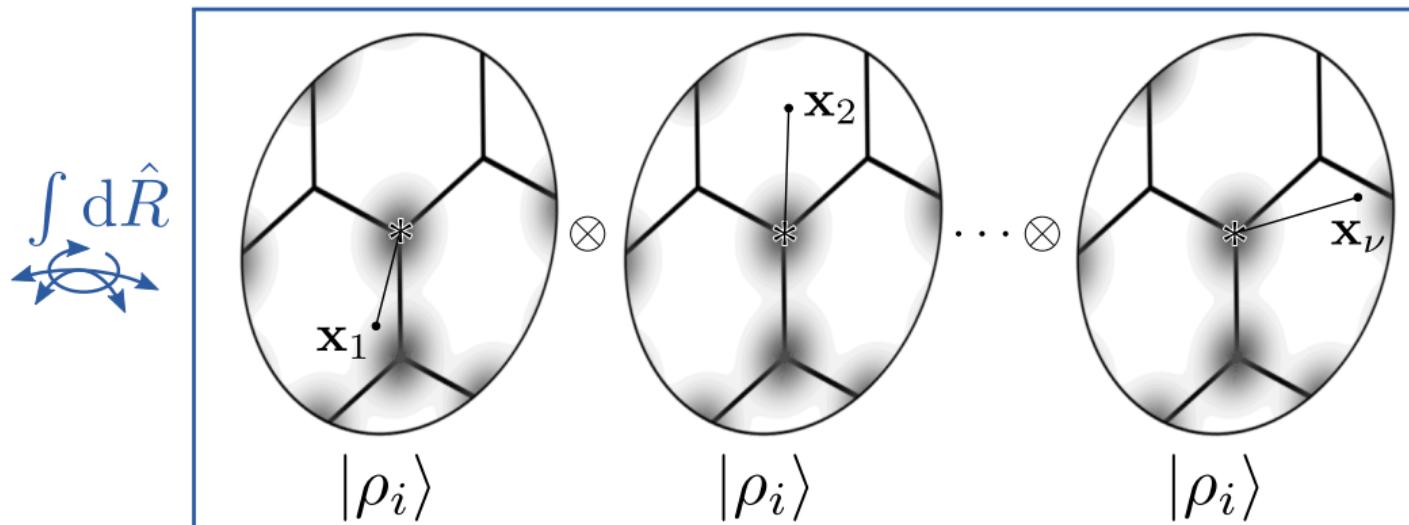
- Any permutation-invariant function of neighbor positions can be linearly expanded in terms of tensor products of the neighbor density
- Symmetrized correlations \leftrightarrow invariant models

$$y(A_i) \approx \sum_{i_1 i_2 \in A_i} \int d\hat{\mathbf{R}} \tilde{y} \left(\hat{\mathbf{R}} \mathbf{r}_{ii_1}, \hat{\mathbf{R}} \mathbf{r}_{ii_2} \right) = \int d\mathbf{x}_1 d\mathbf{x}_2 \tilde{y}(\mathbf{x}_1, \mathbf{x}_2) \int d\hat{\mathbf{R}} \langle \mathbf{x}_1 | \hat{\mathbf{R}} A_i \rangle \langle \mathbf{x}_2 | \hat{\mathbf{R}} A_i \rangle$$

$$y(A_i) \approx \sum_{nn'l} y_{nn'l} \sum_m \langle nlm | A_i \rangle \langle n'lm | A_i \rangle$$

Atom-Centered Density Correlations

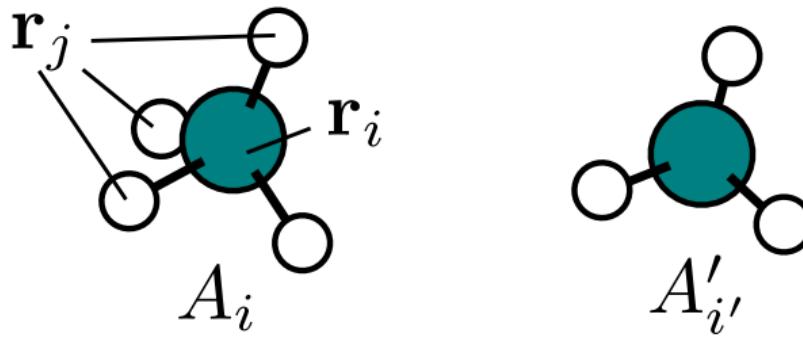
- Any permutation-invariant function of neighbor positions can be linearly expanded in terms of tensor products of the neighbor density
- Symmetrized correlations \leftrightarrow invariant models



Smooth overlap of atomic positions a worked example

Representing chemical environments

- Smooth overlap of atomic densities (SOAP): a kernel to compare atomic environments
 - Atomic environments are defined by the relative position of neighbors (*translation-invariant*)
 - Positions are transformed in a neighbor density (*permutation invariant*)
 - Similarity between environments → overlap **kernel**
 - Averaged over rotations (*rotation invariant*)

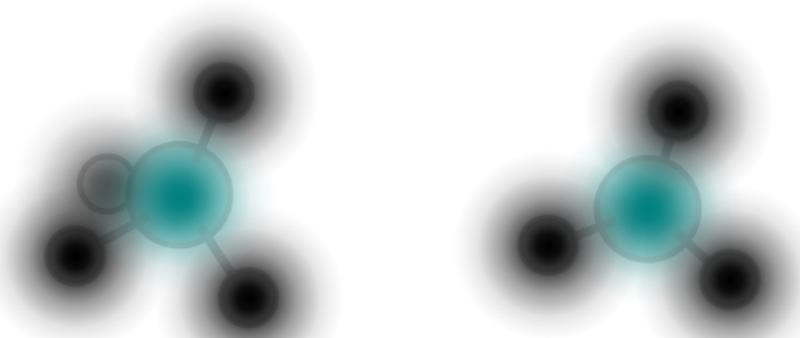


$$\{\mathbf{r}_{ji} = \mathbf{r}_j - \mathbf{r}_i\} \leftrightarrow A_i$$

Bartók, Kondor, Csányi, PRB (2013)

Representing chemical environments

- Smooth overlap of atomic densities (SOAP): a kernel to compare atomic environments
 - ① Atomic environments are defined by the relative position of neighbors (*translation-invariant*)
 - ② Positions are transformed in a neighbor density (*permutation invariant*)
 - ③ Similarity between environments → overlap **kernel**
 - ④ Averaged over rotations (*rotation invariant*)

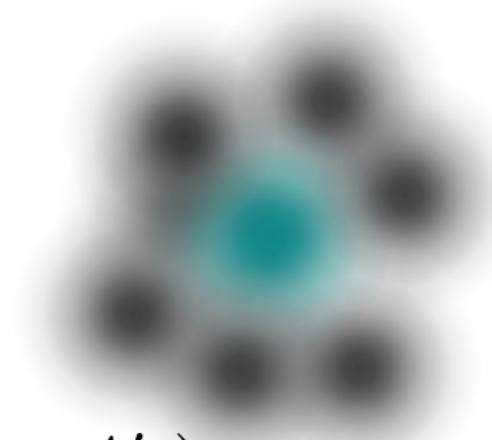


$$\begin{aligned}\langle a\mathbf{x} | \rho_i \rangle &= \sum_{j \in A_i} \delta_{aa_j} \langle \mathbf{x} | \mathbf{r}_{ji}; g \rangle \\ \langle \mathbf{x} | \mathbf{r}_{ji}; g \rangle &\equiv g(\mathbf{x} - \mathbf{r}_{ji})\end{aligned}$$

Bartók, Kondor, Csányi, PRB (2013)

Representing chemical environments

- Smooth overlap of atomic densities (SOAP): a kernel to compare atomic environments
 - ① Atomic environments are defined by the relative position of neighbors (*translation-invariant*)
 - ② Positions are transformed in a neighbor density (*permutation invariant*)
 - ③ Similarity between environments → overlap **kernel**
 - ④ Averaged over rotations (*rotation invariant*)

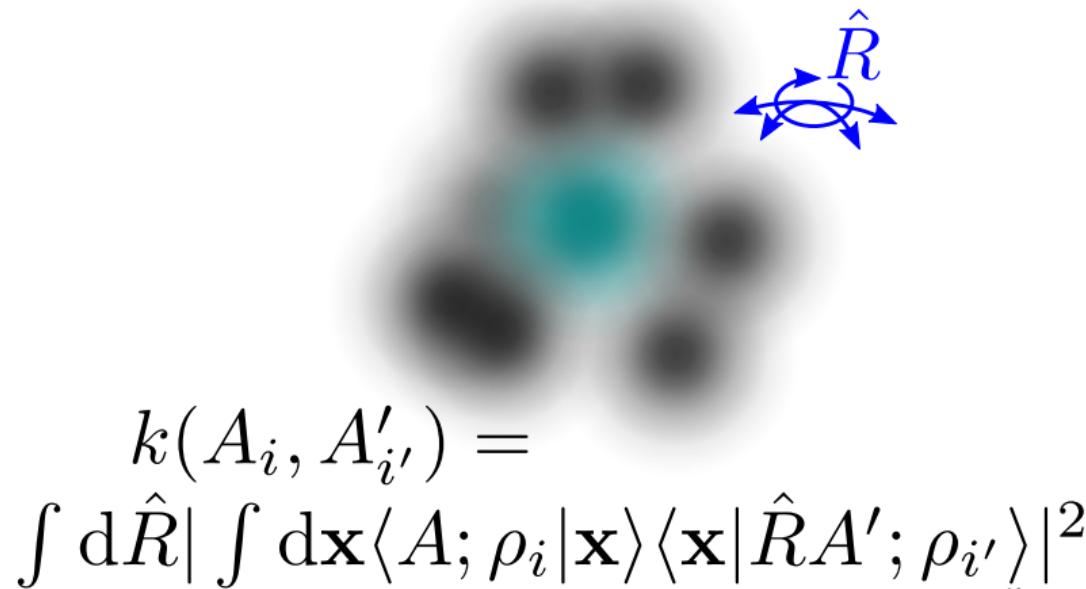


$$k(A_i, A'_{i'}) = \int d\mathbf{x} \langle A; \rho_i | \mathbf{x} \rangle \langle \mathbf{x} | A'; \rho_{i'} \rangle$$

Bartók, Kondor, Csányi, PRB (2013)

Representing chemical environments

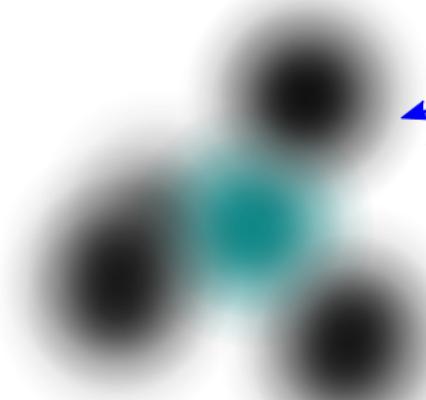
- Smooth overlap of atomic densities (SOAP): a kernel to compare atomic environments
 - ① Atomic environments are defined by the relative position of neighbors (*translation-invariant*)
 - ② Positions are transformed in a neighbor density (*permutation invariant*)
 - ③ Similarity between environments → overlap **kernel**
 - ④ Averaged over rotations (*rotation invariant*)


$$k(A_i, A'_{i'}) = \int d\hat{R} \left| \int d\mathbf{x} \langle A; \rho_i | \mathbf{x} \rangle \langle \mathbf{x} | \hat{R} A'; \rho_{i'} \rangle \right|^2$$

Bartók, Kondor, Csányi, PRB (2013)

Representing chemical environments

- Smooth overlap of atomic densities (SOAP): a kernel to compare atomic environments
 - ① Atomic environments are defined by the relative position of neighbors (*translation-invariant*)
 - ② Positions are transformed in a neighbor density (*permutation invariant*)
 - ③ Similarity between environments → overlap **kernel**
 - ④ Averaged over rotations (*rotation invariant*)


$$k(A_i, A'_{i'}) = \int d\hat{R} \left| \int d\mathbf{x} \langle A; \rho_i | \mathbf{x} \rangle \langle \mathbf{x} | \hat{R} A'; \rho_{i'} \rangle \right|^2$$

Bartók, Kondor, Csányi, PRB (2013)

Symmetrized density correlations

- The same information can be encoded in features, equivalent to symmetrized correlations of the neighbor density

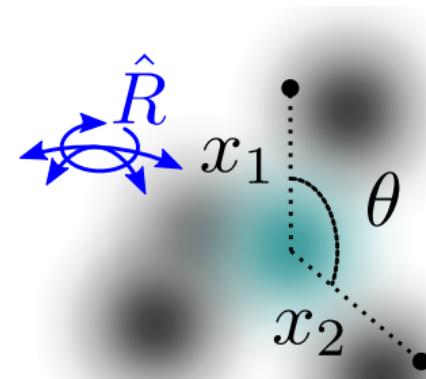
- 1 Symmetrize over rotations a tensor product of the neighbor densities
- 2 This is equivalent to a function of two distances and one angle
- 3 In the limit of sharp Gaussians, this is equivalent to a list of 2-neighbors tuples $(r_{j_1 i}, r_{j_2 i}, \hat{r}_{j_1 i} \cdot \hat{r}_{j_2 i})$
- 4 Linear model \rightarrow 3-body potential!

$$\langle \mathbf{x}; \mathbf{x}' | A; \overline{\rho_i^{\otimes 2}} \rangle = \int d\hat{R} \langle \mathbf{x} | \hat{R}A; \rho_i \rangle \langle \mathbf{x}' | \hat{R}A; \rho_i \rangle$$

Willatt, Musil, MC, JCP (2019)

Symmetrized density correlations

- The same information can be encoded in features, equivalent to symmetrized correlations of the neighbor density
 - 1 Symmetrize over rotations a tensor product of the neighbor densities
 - 2 This is equivalent to a function of two distances and one angle
 - 3 In the limit of sharp Gaussians, this is equivalent to a list of 2-neighbors tuples $(r_{j_1 i}, r_{j_2 i}, \hat{\mathbf{r}}_{j_1 i} \cdot \hat{\mathbf{r}}_{j_2 i})$
 - 4 Linear model \rightarrow 3-body potential!



$$\begin{aligned}\langle x_1; x_2; \theta | A; \overline{\rho_i^{\otimes 2}} \rangle &= \\ \int d\hat{R} \langle x_1 \hat{R} \hat{\mathbf{e}}_z | A; \rho_i \rangle & \\ \langle x_2 \hat{R} (\hat{\mathbf{e}}_z \cos \theta + \hat{\mathbf{e}}_x \sin \theta) | A; \rho_i \rangle &\end{aligned}$$

Willatt, Musil, **MC**, JCP (2019)

Symmetrized density correlations

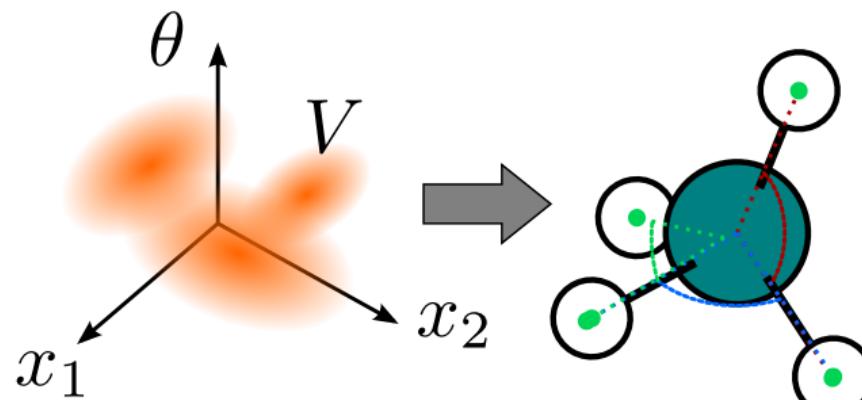
- The same information can be encoded in features, equivalent to symmetrized correlations of the neighbor density
 - 1 Symmetrize over rotations a tensor product of the neighbor densities
 - 2 This is equivalent to a function of two distances and one angle
 - 3 In the limit of sharp Gaussians, this is equivalent to a list of 2-neighbors tuples $(r_{j_1 i}, r_{j_2 i}, \hat{\mathbf{r}}_{j_1 i} \cdot \hat{\mathbf{r}}_{j_2 i})$
 - 4 Linear model \rightarrow 3-body potential!

$$\begin{aligned}\langle x_1; x_2; \theta | A; \overline{\delta_i^{\otimes 2}} \rangle = \\ \sum_{j_1 j_2} \delta(x_1 - r_{j_1 i}) \delta(x_2 - r_{j_2 i}) \\ \delta(\cos \theta - \hat{\mathbf{r}}_{j_1 i} \cdot \hat{\mathbf{r}}_{j_2 i})\end{aligned}$$

Willatt, Musil, **MC**, JCP (2019)

Symmetrized density correlations

- The same information can be encoded in features, equivalent to symmetrized correlations of the neighbor density
 - 1 Symmetrize over rotations a tensor product of the neighbor densities
 - 2 This is equivalent to a function of two distances and one angle
 - 3 In the limit of sharp Gaussians, this is equivalent to a list of 2-neighbors tuples $(r_{j_1 i}, r_{j_2 i}, \hat{\mathbf{r}}_{j_1 i} \cdot \hat{\mathbf{r}}_{j_2 i})$
 - 4 Linear model \rightarrow 3-body potential!

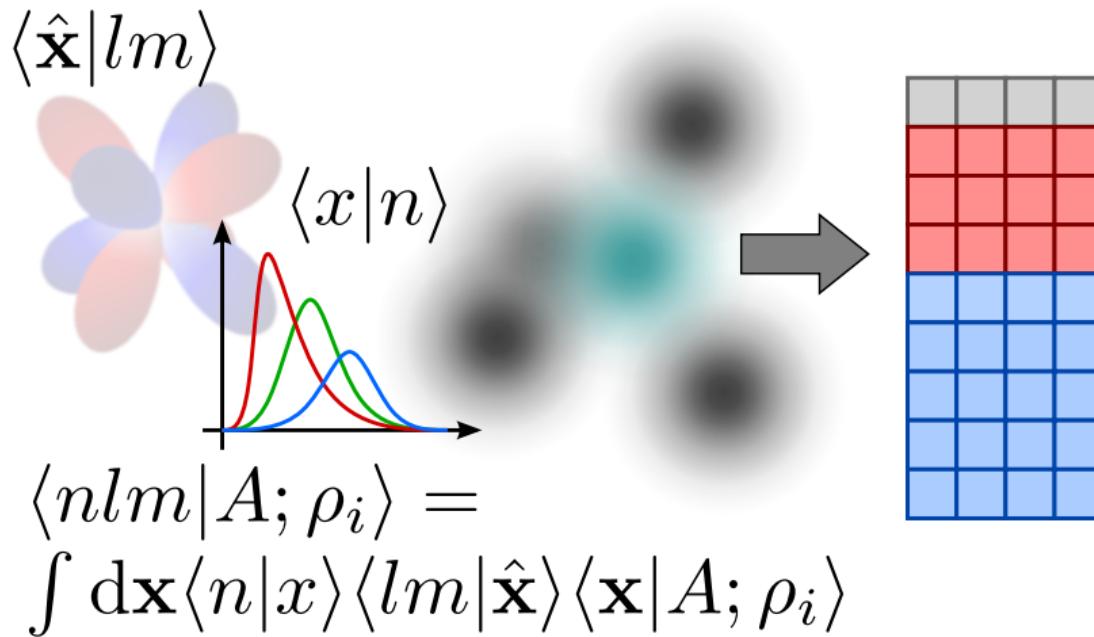


$$\int \langle V | x_1; x_2; \theta \rangle \langle x_1; x_2; \theta | A; \overline{\delta_i^{\otimes 2}} \rangle = \\ \sum_{j_1 j_2} V(r_{j_1 i}, r_{j_2 i}, \hat{\mathbf{r}}_{j_1 i} \cdot \hat{\mathbf{r}}_{j_2 i})$$

Willatt, Musil, **MC**, JCP (2019)

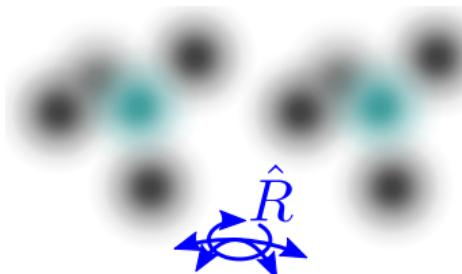
Density trick in an $\langle nlm |$ basis

- The symmetrized correlations can be computed in closed form using a discrete basis
 - The neighbor density can be expanded on a basis of radial functions $\langle x | n \rangle \equiv R_n(x)$ and spherical harmonics $\langle \hat{x} | lm \rangle \equiv Y_l^m(\hat{x})$
 - Spherical harmonics transform linearly under rotations based on Wigner rotation matrices $\mathbf{D}^l(\hat{R})$
 - Orthogonality of Wigner matrices yields the SOAP powerspectrum



Density trick in an $\langle nlm |$ basis

- The symmetrized correlations can be computed in closed form using a discrete basis
 - The neighbor density can be expanded on a basis of radial functions $\langle x | n \rangle \equiv R_n(x)$ and spherical harmonics $\langle \hat{x} | lm \rangle \equiv Y_l^m(\hat{x})$
 - Spherical harmonics transform linearly under rotations based on Wigner rotation matrices $\mathbf{D}^l(\hat{R})$
 - Orthogonality of Wigner matrices yields the SOAP powerspectrum

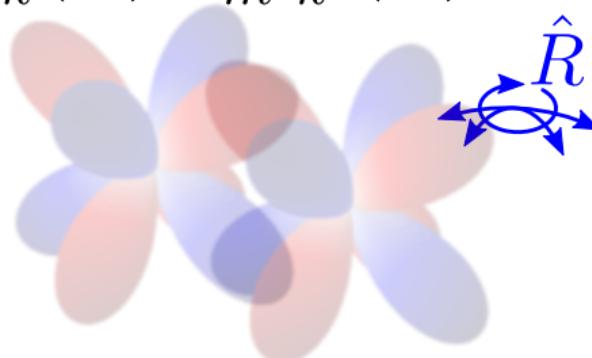


$$\begin{aligned}\hat{R}|lm\rangle &= \sum_{m'} D_{mm'}^l(\hat{R})|lm'\rangle \\ \langle nlm; n'l'm' | A; \rho_i^{\otimes 2} \rangle &= \int d\hat{R} \langle nlm | \hat{R}A; \rho_i \rangle \langle n'l'm' | \hat{R}A; \rho_i \rangle\end{aligned}$$

Density trick in an $\langle nlm |$ basis

- The symmetrized correlations can be computed in closed form using a discrete basis
 - The neighbor density can be expanded on a basis of radial functions $\langle x | n \rangle \equiv R_n(x)$ and spherical harmonics $\langle \hat{x} | lm \rangle \equiv Y_l^m(\hat{x})$
 - Spherical harmonics transform linearly under rotations based on Wigner rotation matrices $\mathbf{D}^l(\hat{R})$
 - Orthogonality of Wigner matrices yields the SOAP powerspectrum

$$\int d\hat{R} \sum_{kk'} D_{mk}^l(\hat{R}) D_{m'k'}^{l'}(\hat{R}) \propto \delta_{ll'} \delta_{mm'} \delta_{kk'}$$

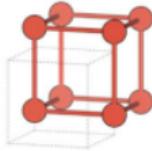


$$\begin{aligned} \langle nn'l | A; \rho_i^{\otimes 2} \rangle &= \\ \sum_m \langle nlm | A; \rho_i \rangle \langle n'lm | A; \rho_i \rangle & \end{aligned}$$

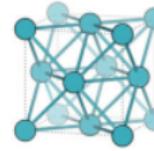
Variations on a theme

- Most of the existing density-based representations and kernels emerge as special cases of this construction
 - Basis set choice - e.g. plane waves basis for $|\overline{\rho^{\otimes 2}}\rangle$ (Ziletti et al. N.Comm 2018)
 - Projection on symmetry functions (Behler-Parrinello, ANI, DeepMD)

$$\langle \mathbf{k} | A; \overline{\rho^{\otimes 2}} \rangle = \sum_{ij \in A} e^{i\mathbf{k} \cdot \mathbf{r}_{ij}}$$



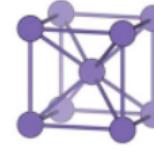
Simple cubic
(sc) structure
spgroup = 221



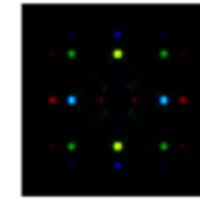
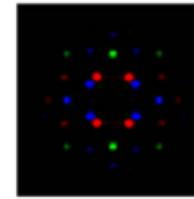
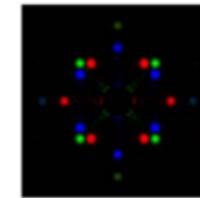
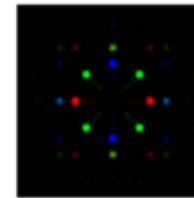
Face-centered-cubic
(fcc) structure
spgroup = 225



Diamond
(diam) structure
spgroup = 227



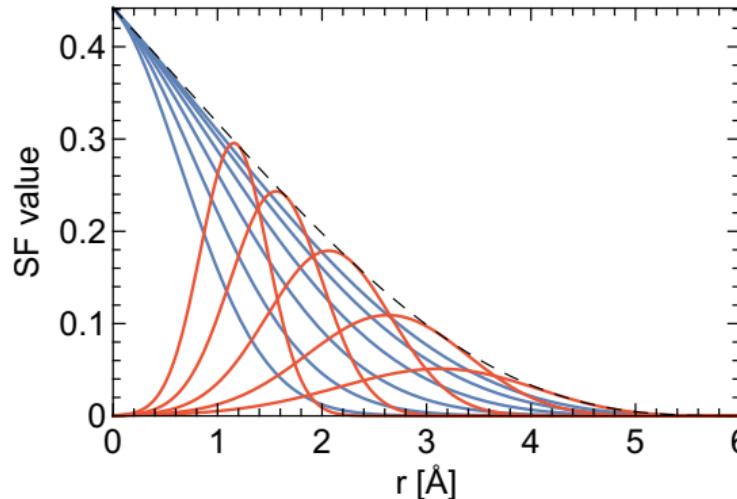
Body-centered-cubic
(bcc) structure
spgroup = 229



Variations on a theme

- Most of the existing density-based representations and kernels emerge as special cases of this construction
 - Basis set choice - e.g. plane waves basis for $|\overline{\rho_i^{\otimes 2}}\rangle$ (Ziletti et al. N.Comm 2018)
 - Projection on symmetry functions (Behler-Parrinello, ANI, DeepMD)

$$\langle abG_2|\overline{\rho_i^{\otimes 1}}\rangle = \delta_{aa} \int dr G_2(r) \langle br|\overline{\rho_i^{\otimes 1}}; g \rightarrow \delta \rangle$$



Willatt, Musil, **MC**, JCP (2019), <https://arxiv.org/pdf/1807.00408>

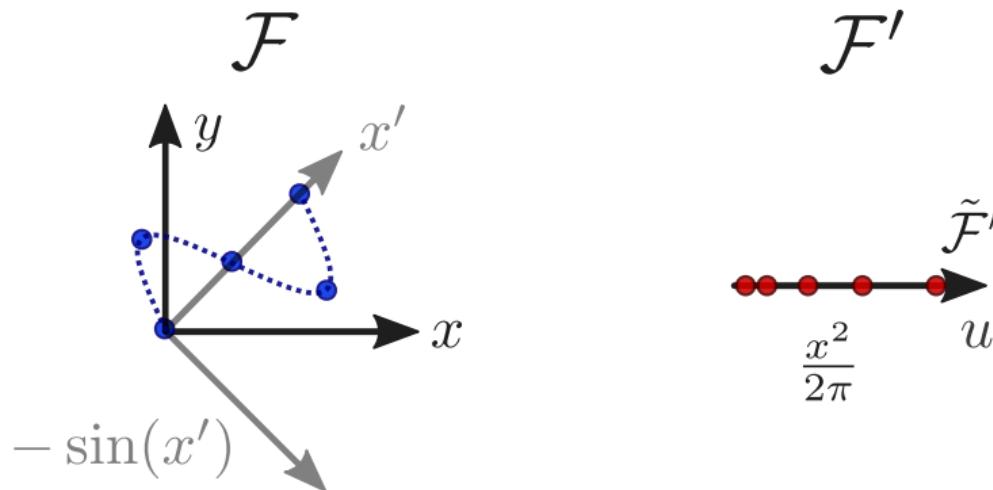
Basis set engineering



Measuring feature spaces

- General problem: how do we compare information content of different choices of features?
How to compare metrics and kernels?
- Feature space Reconstruction Error (**FRE**) measures the linearly-embeddable mutual information. Locally-linear and kernelized extensions also available

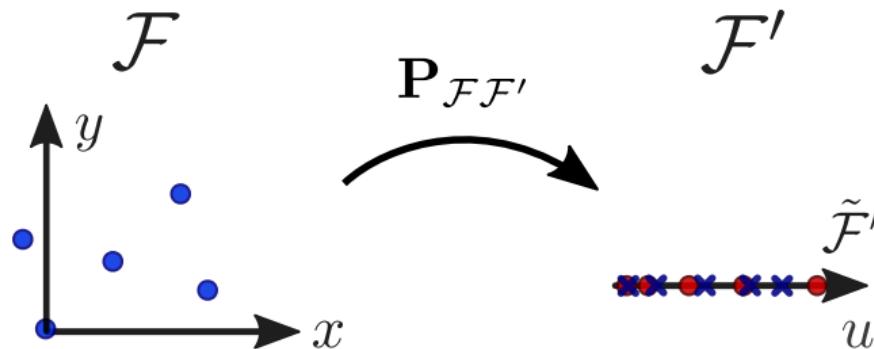
$$\text{GFRE}(\mathcal{F} \rightarrow \mathcal{F}') = \min_{\mathbf{P} \in \mathbb{R}^{n_{\mathcal{F}} \times n_{\mathcal{F}'}}} \|\mathbf{X}_{\mathcal{F}'} - \mathbf{X}_{\mathcal{F}} \mathbf{P}\|$$



Measuring feature spaces

- General problem: how do we compare information content of different choices of features?
How to compare metrics and kernels?
- Feature space Reconstruction Error (**FRE**) measures the linearly-embeddable mutual information. Locally-linear and kernelized extensions also available

$$\text{GFRE}(\mathcal{F} \rightarrow \mathcal{F}') = \min_{\mathbf{P} \in \mathbb{R}^{n_{\mathcal{F}} \times n_{\mathcal{F}'}}} \|\mathbf{X}_{\mathcal{F}'} - \mathbf{X}_{\mathcal{F}} \mathbf{P}\|$$

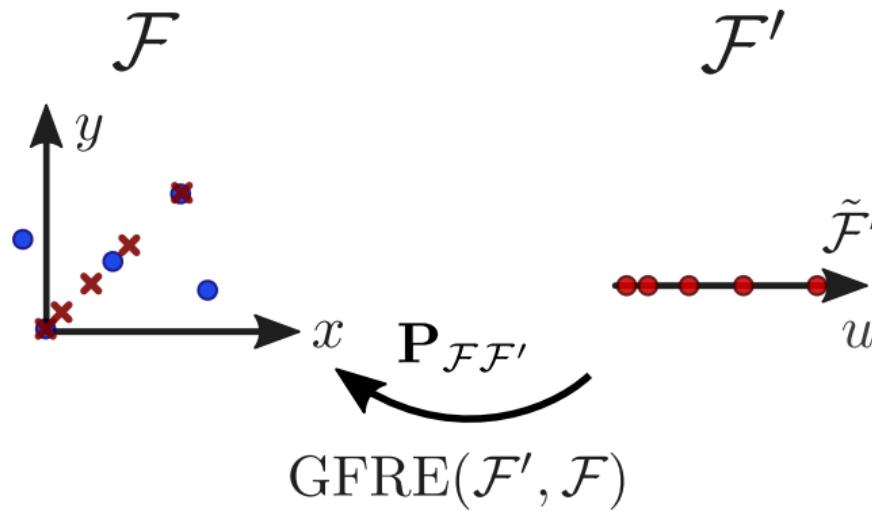


$$\text{GFRE}(\mathcal{F}, \mathcal{F}')$$

Measuring feature spaces

- General problem: how do we compare information content of different choices of features?
How to compare metrics and kernels?
- Feature space Reconstruction Error (**FRE**) measures the linearly-embeddable mutual information. Locally-linear and kernelized extensions also available

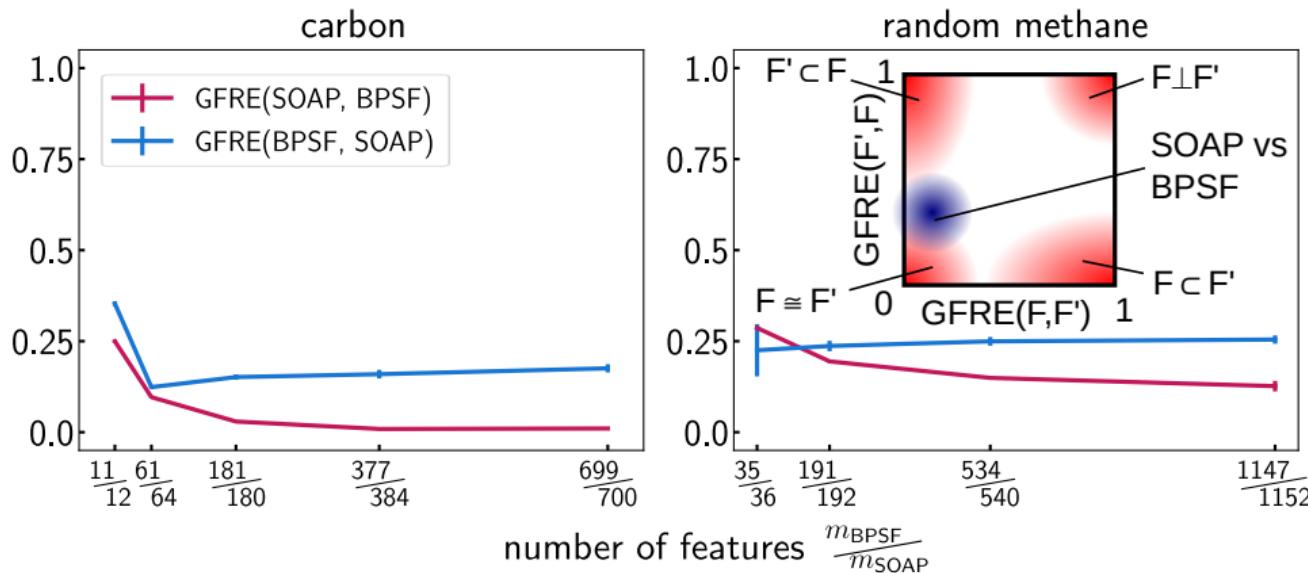
$$\text{GFRE}(\mathcal{F} \rightarrow \mathcal{F}') = \min_{\mathbf{P} \in \mathbb{R}^{n_{\mathcal{F}} \times n_{\mathcal{F}'}}} \|\mathbf{X}_{\mathcal{F}'} - \mathbf{X}_{\mathcal{F}} \mathbf{P}\|$$



Measuring feature spaces

- General problem: how do we compare information content of different choices of features?
How to compare metrics and kernels?
- Feature space Reconstruction Error (**FRE**) measures the linearly-embeddable mutual information. Locally-linear and kernelized extensions also available

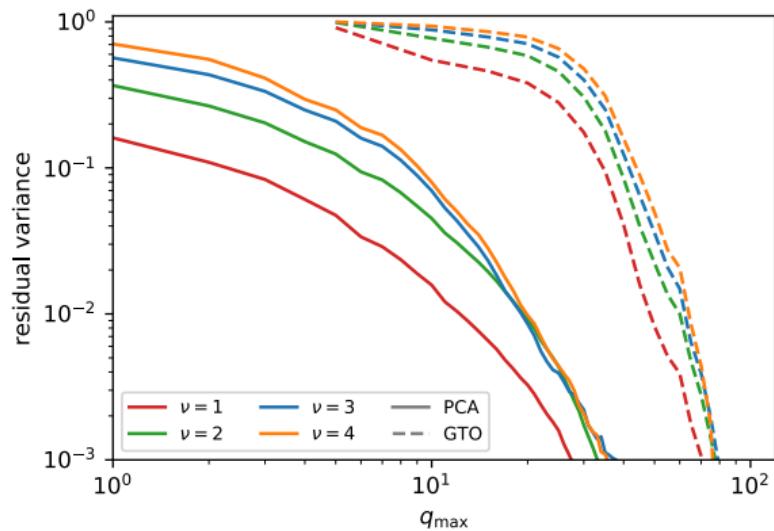
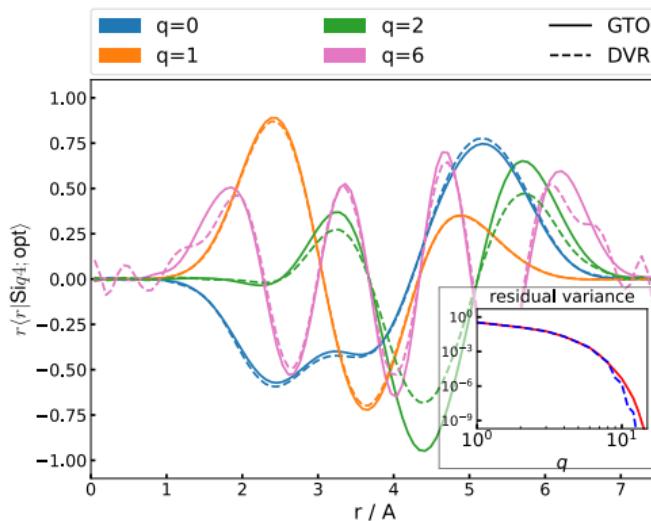
$$\text{GFRE}(\mathcal{F} \rightarrow \mathcal{F}') = \min_{\mathbf{P} \in \mathbb{R}^{n_{\mathcal{F}} \times n_{\mathcal{F}'}}} \|\mathbf{X}_{\mathcal{F}'} - \mathbf{X}_{\mathcal{F}} \mathbf{P}\|$$



Basis set optimization - data-driven vs smoothness

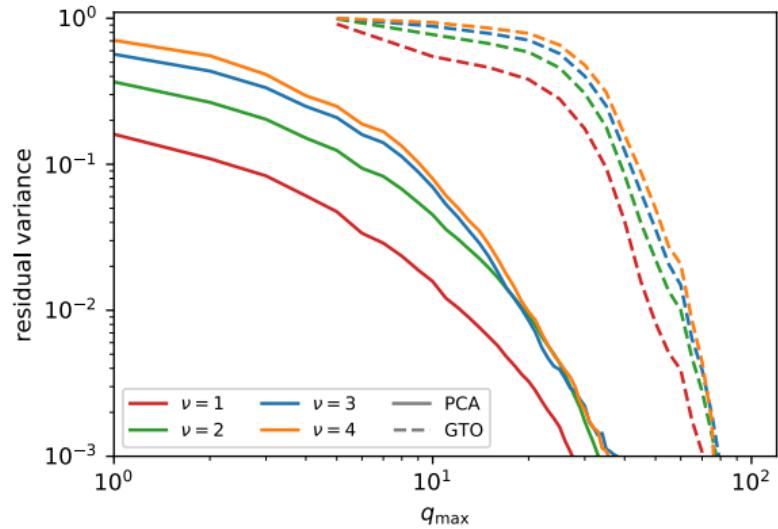
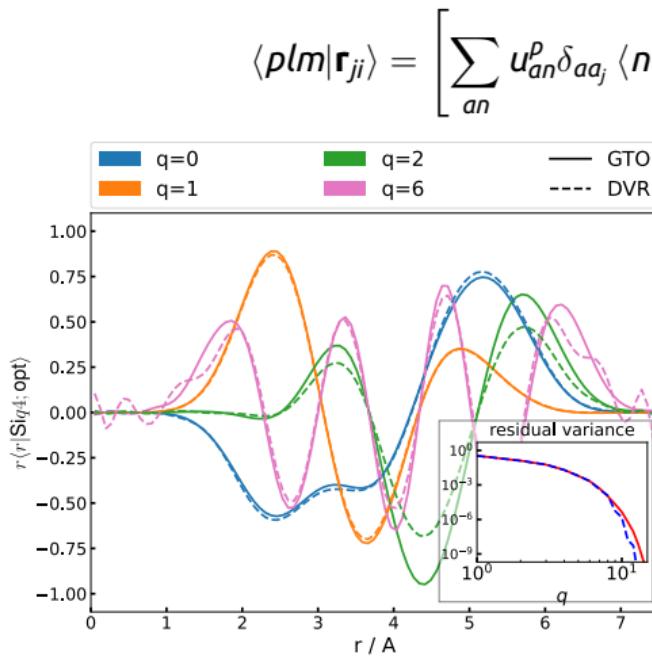
- Can we make a better choice of basis with an optimized contraction?
- Data-driven optimal contraction: PCA! Direct evaluation of contracted basis using splines
- Quite surprising: a general smoothness criterion works better in practice for potentials

$$\langle anlm | \rightarrow \langle plm | = \sum u_{an}^p \langle anlm |$$



Basis set optimization - data-driven vs smoothness

- Can we make a better choice of basis with an optimized contraction?
- Data-driven optimal contraction: PCA! Direct evaluation of contracted basis using splines
- Quite surprising: a general smoothness criterion works better in practice for potentials

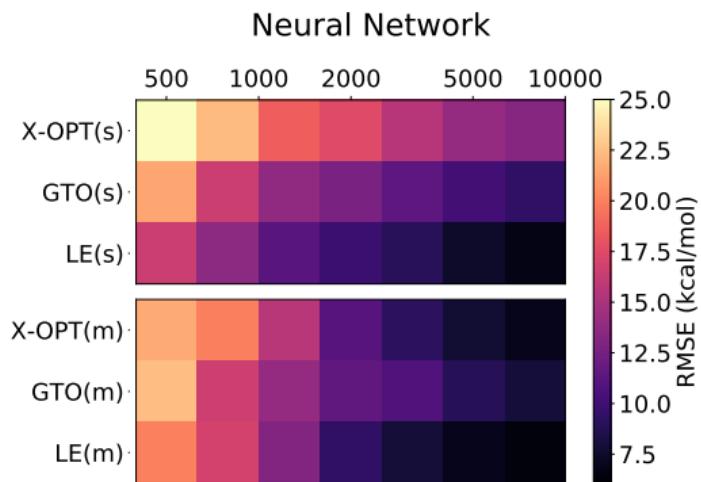
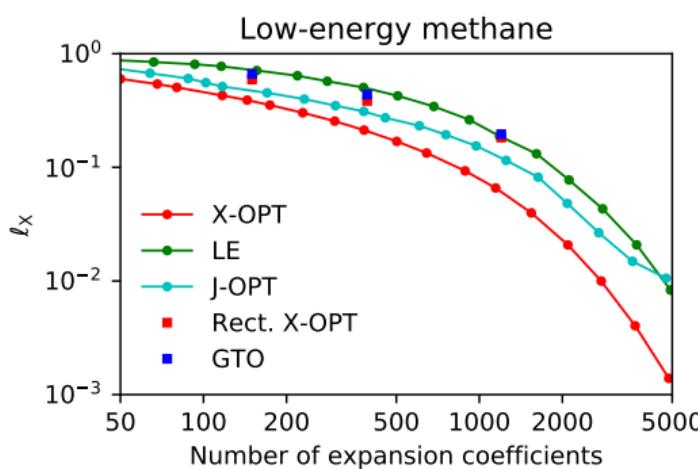


Goscinski et al., JCP (2021)

Basis set optimization - data-driven vs smoothness

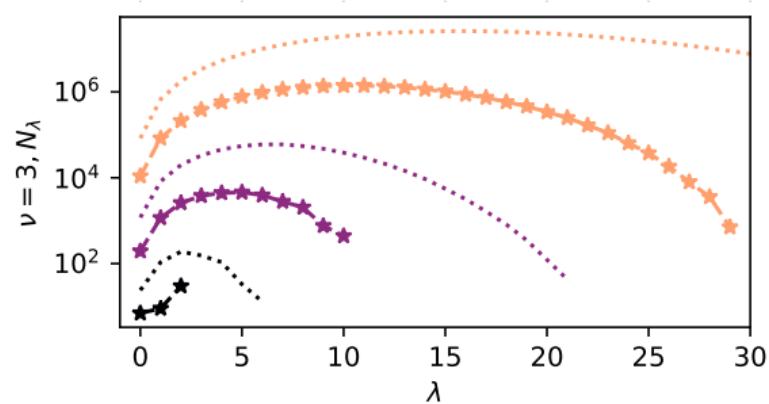
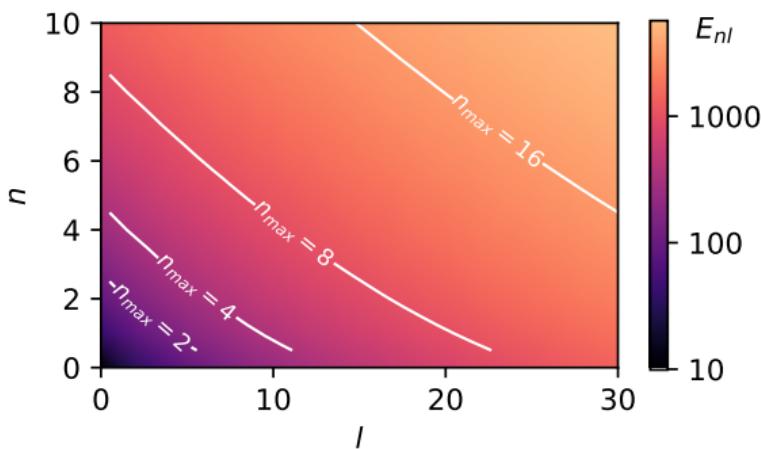
- Can we make a better choice of basis with an optimized contraction?
- Data-driven optimal contraction: PCA! Direct evaluation of contracted basis using splines
- Quite surprising: a general smoothness criterion works better in practice for potentials

$$u(r) = \operatorname{argmin} Q(u) \quad Q = \int_{\Omega} |\nabla u(\mathbf{x})|^2 \, d\mathbf{x} \Big/ \int_{\Omega} |u(\mathbf{x})|^2 \, d\mathbf{x}$$



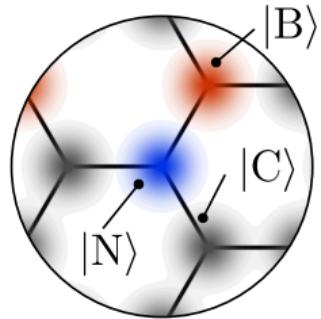
Basis set optimization - data-driven vs smoothness

- Can we make a better choice of basis with an optimized contraction?
- Data-driven optimal contraction: PCA! Direct evaluation of contracted basis using splines
- Quite surprising: a general smoothness criterion works better in practice for potentials



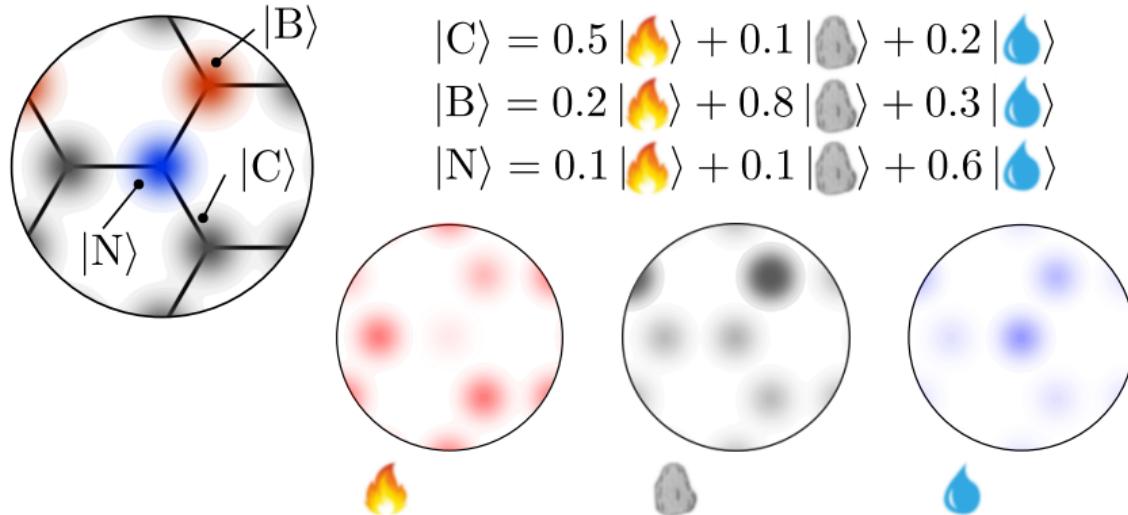
Machine-learning with pseudoelements

- How to learn with multiple species? Decorate atomic Gaussian with elemental kets $|C\rangle$, $|N\rangle$, ...
- Tensor-product features lead to exponential scaling with the number of elements
- Expand each ket in a finite basis, $|b\rangle = \sum_a u_{ba} |a\rangle$
- Optimize coefficients for “alchemical learning”



Machine-learning with pseudoelements

- How to learn with multiple species? Decorate atomic Gaussian with elemental kets $|C\rangle$, $|N\rangle$, ...
- Tensor-product features lead to exponential scaling with the number of elements
- Expand each ket in a finite basis, $|b\rangle = \sum_a u_{ba} |a\rangle$
- Optimize coefficients for “alchemical learning”



Empedocles et al. (ca 360BC). Metaphor courtesy of Albert Bartók.
Willatt, Musil, MC, PCCP (2018)

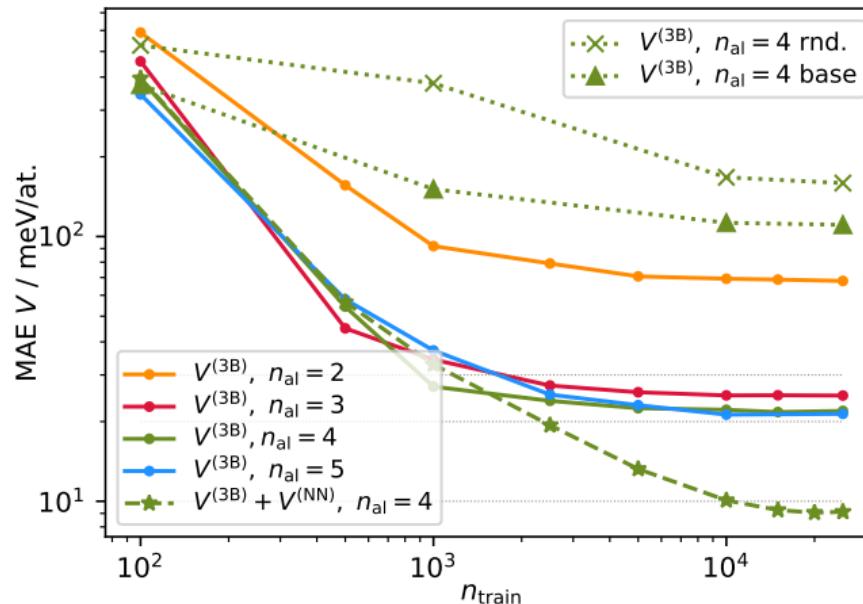
A potential for transition metal alloys

- 25 *d*-block elements that are used in high-entropy alloys.
- ~25'000 training structures with distorted *fcc* and *bcc* structures
- Alchemical compression with 3-body potential saturates at 3-4 pseudoelements.
- A non-linear layer brings error below 10meV MAE
- Interpretability: alchemical weights reflect the periodic table ordering; interpolation!

Group →	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	
↓ Period	1															
1	H															
2	Li	Be											B	C	N	
3	Na	Mg											Al	Si	P	
4	K	Ca	21 Sc	22 Ti	23 V	24 Cr	25 Mn	26 Fe	27 Co	28 Ni	29 Cu	30 Zn		Ga	Ge	As
5	Rb	Sr	39 Y	40 Zr	41 Nb	42 Mo	Tc	44 Ru	45 Rh	46 Pd	47 Ag		Cd	In	Sn	Sb
6	Cs	Ba	71 Lu	72 Hf	73 Ta	74 W	Re	Os	77 Ir	78 Pt	79 Au		Hg	Tl	Pb	Bi

A potential for transition metal alloys

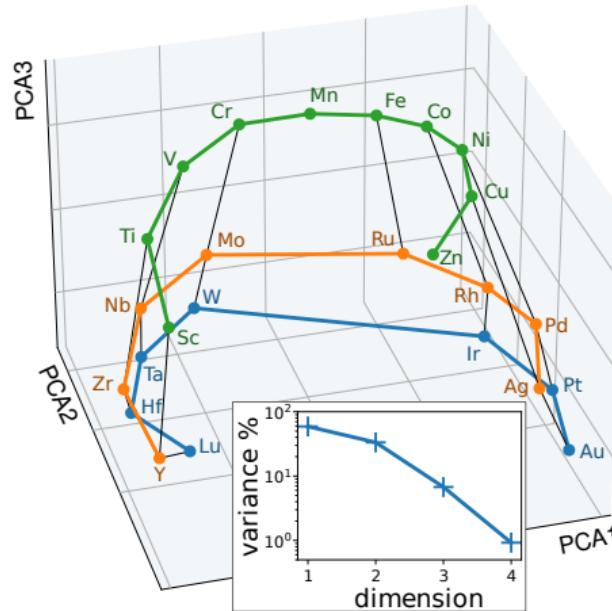
- 25 *d*-block elements that are used in high-entropy alloys.
- ~25'000 training structures with distorted *fcc* and *bcc* structures
- Alchemical compression with 3-body potential saturates at 3-4 pseudoelements.
- A non-linear layer brings error below 10meV MAE
- Interpretability: alchemical weights reflect the periodic table ordering; interpolation!



Lopanitsyna, Fraux, Springer, De, MC, Phys. Rev. Materials (2023)

A potential for transition metal alloys

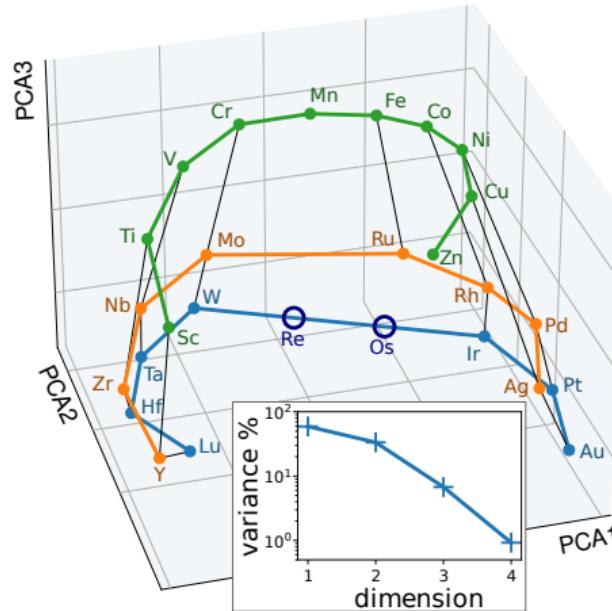
- 25 *d*-block elements that are used in high-entropy alloys.
- ~25'000 training structures with distorted *fcc* and *bcc* structures
- Alchemical compression with 3-body potential saturates at 3-4 pseudoelements.
- A non-linear layer brings error below 10meV MAE
- Interpretability: alchemical weights reflect the periodic table ordering; interpolation!



Lopanitsyna, Fraux, Springer, De, MC, Phys. Rev. Materials (2023)

A potential for transition metal alloys

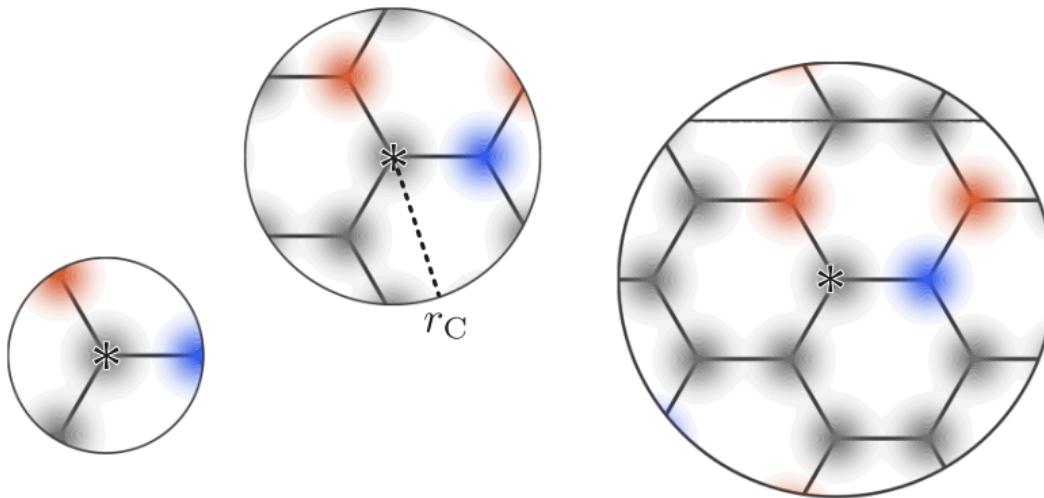
- 25 *d*-block elements that are used in high-entropy alloys.
- ~25'000 training structures with distorted *fcc* and *bcc* structures
- Alchemical compression with 3-body potential saturates at 3-4 pseudoelements.
- A non-linear layer brings error below 10meV MAE
- Interpretability: alchemical weights reflect the periodic table ordering; interpolation!



Lopanitsyna, Fraux, Springer, De, MC, Phys. Rev. Materials (2023)

Interpretable representations: range and body order

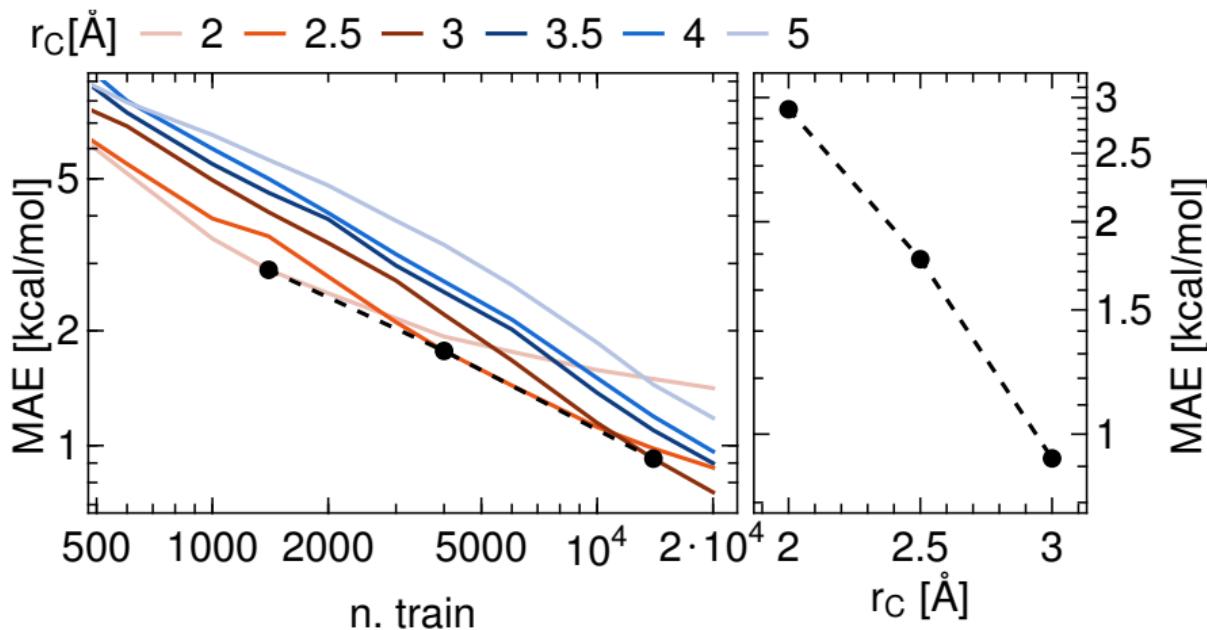
- Environment kernels can be built for different cutoff radii: dimensionality/accuracy tradeoff, a measure of the range of interactions



Bartók, De, Poelking, Kermode, Bernstein, Csányi, **MC**, Science Advances (2017); [data: QM9, von Lilienfeld&C]

Interpretable representations: range and body order

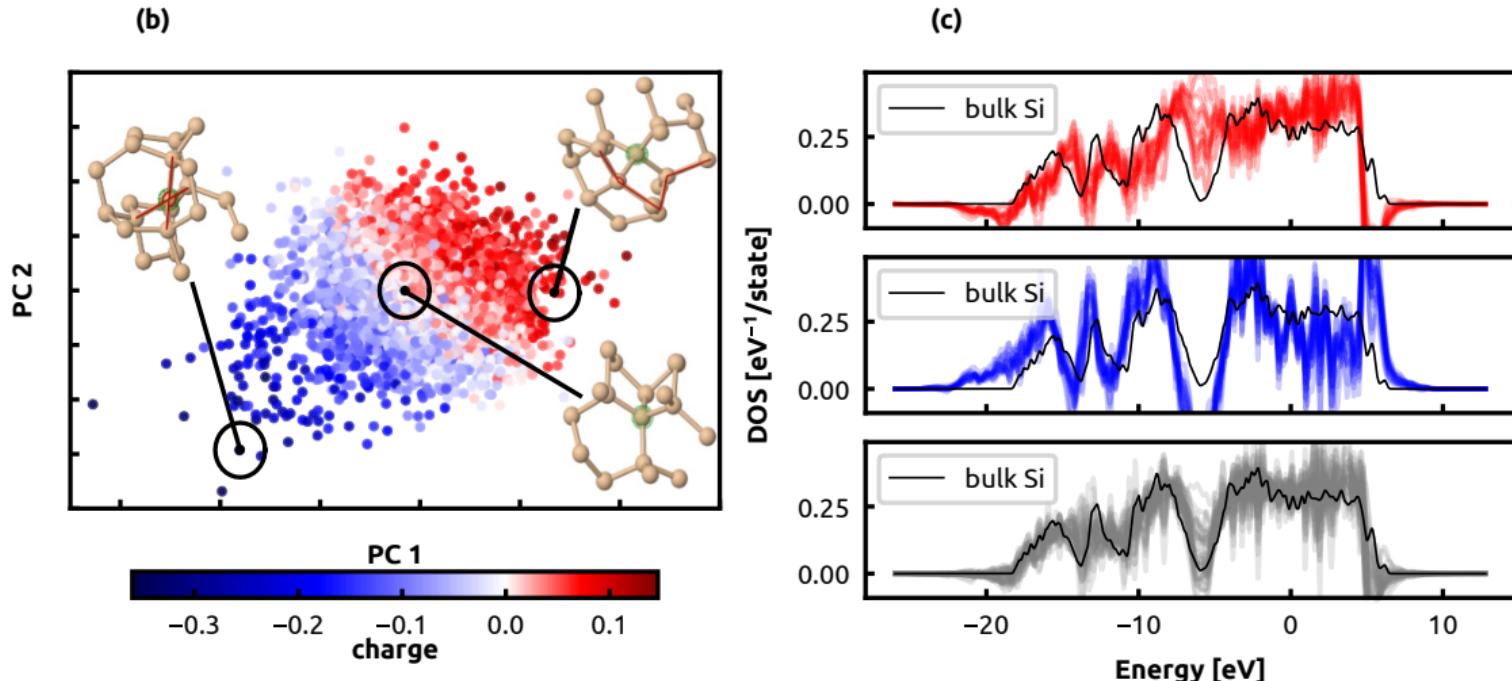
- Environment kernels can be built for different cutoff radii: dimensionality/accuracy tradeoff, a measure of the range of interactions



Bartók, De, Poelking, Kermode, Bernstein, Csányi, **MC**, Science Advances (2017); [data: QM9, von Lilienfeld&C]

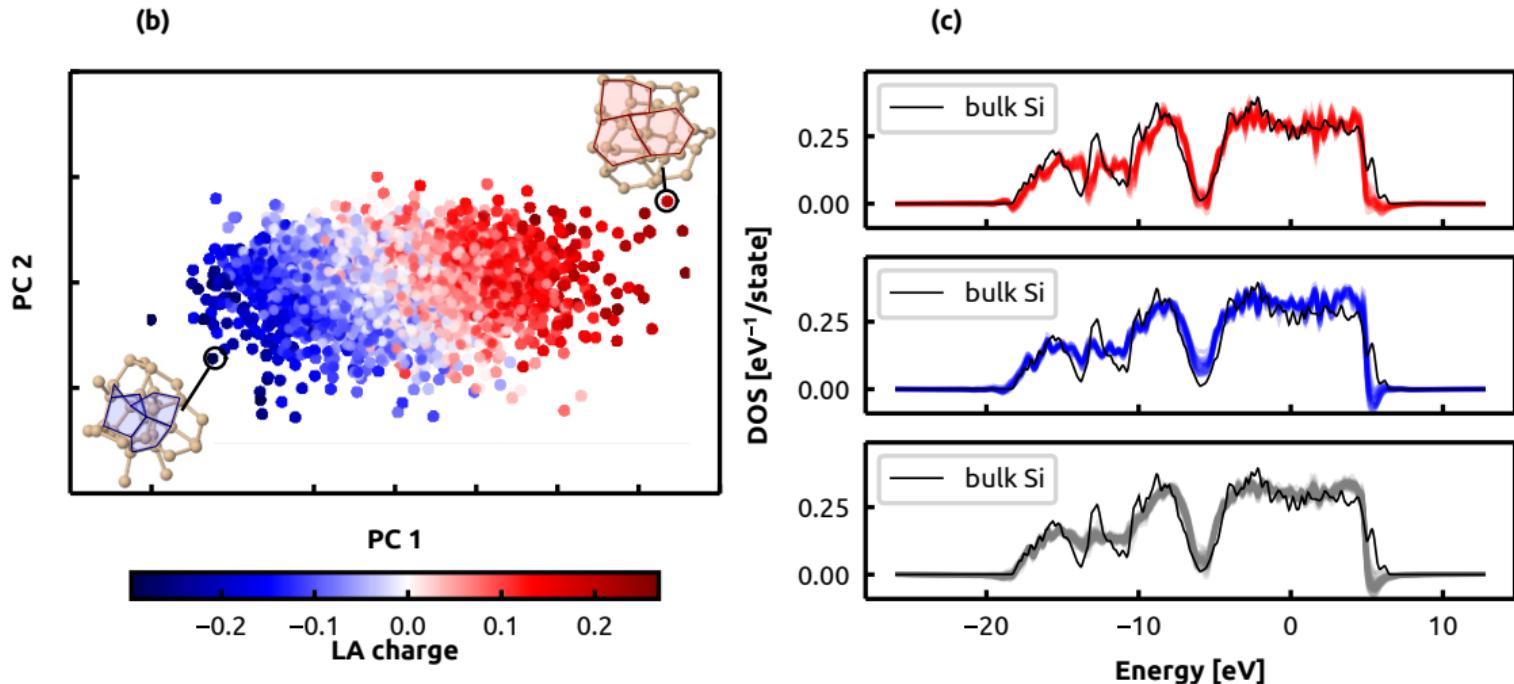
Interpretable representations: local predictions

- Atom-centered decomposition lends themselves naturally to be interpreted in terms of local atomic motifs. Is the decompositions meaningful?
- Local prediction rigidity: assess (and enhance) robustness of decomposition to model details



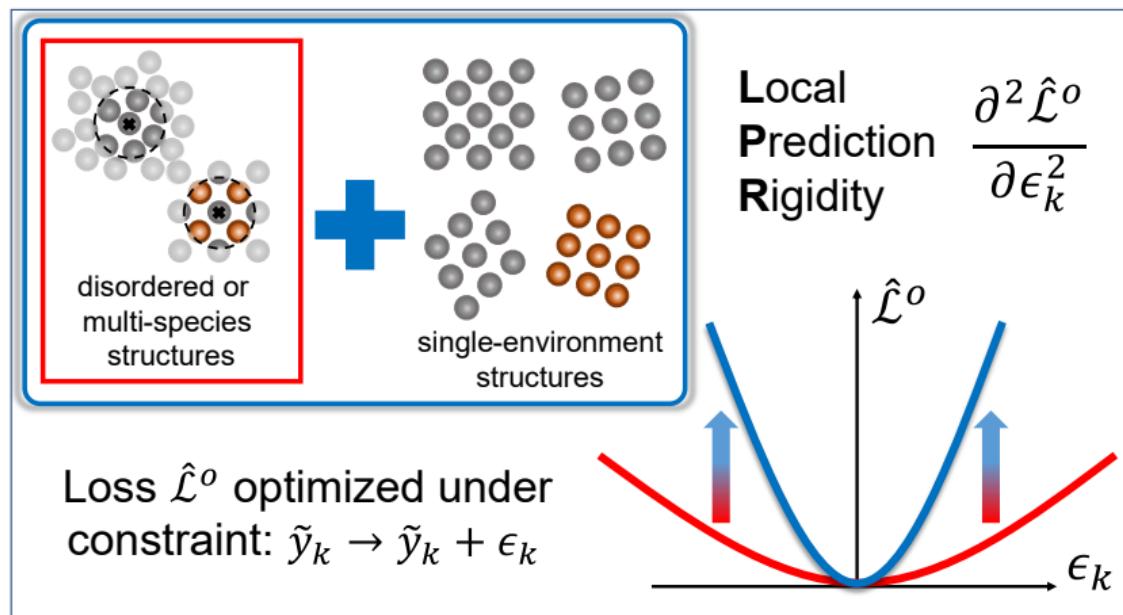
Interpretable representations: local predictions

- Atom-centered decomposition lends themselves naturally to be interpreted in terms of local atomic motifs. Is the decompositions meaningful?
- Local prediction rigidity: assess (and enhance) robustness of decomposition to model details



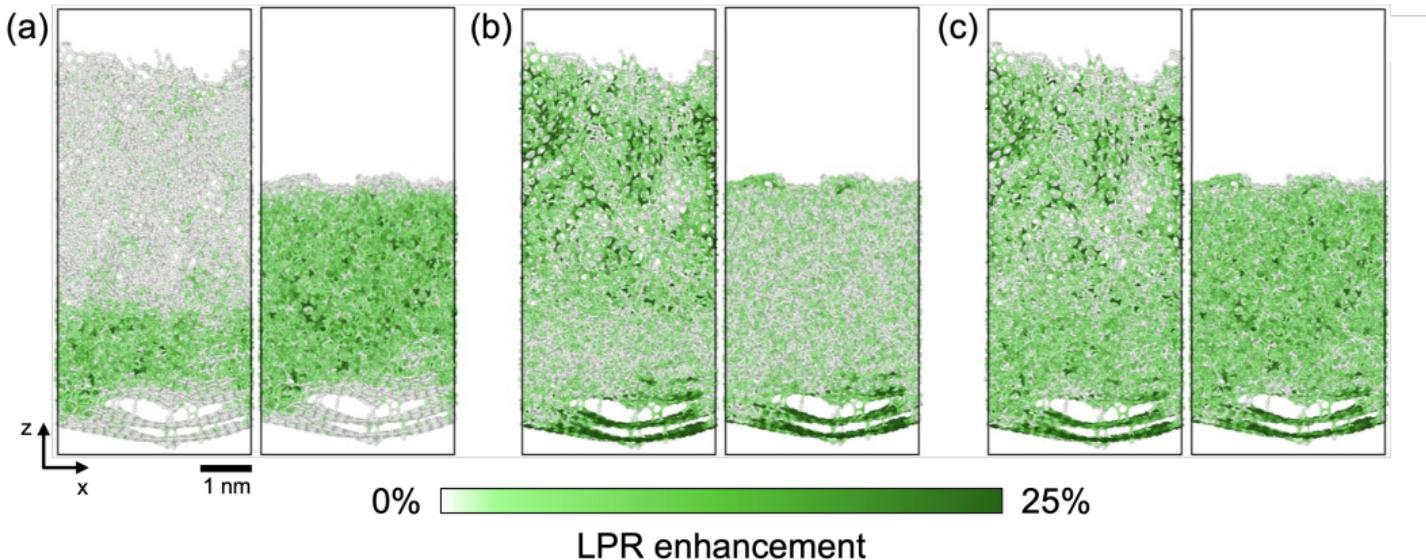
Interpretable representations: local predictions

- Atom-centered decomposition lends themselves naturally to be interpreted in terms of local atomic motifs. Is the decompositions meaningful?
- Local prediction rigidity: assess (and enhance) robustness of decomposition to model details



Interpretable representations: local predictions

- Atom-centered decomposition lends themselves naturally to be interpreted in terms of local atomic motifs. Is the decompositions meaningful?
- Local prediction rigidity: assess (and enhance) robustness of decomposition to model details



Wrapping up

- Many frameworks have been proposed to provide a mathematical description of a molecular geometry
- Descriptors formed by symmetry-adapted correlations of the neighbor density provide a complete linear basis to expand scalar properties.
- The overwhelming majority of known ML frameworks for matter can be seen as a variation on this theme.
- Systematic, principled, interpretable models. Model optimization by basis set engineering

Behler, J. Chem. Phys. 134(7), 074106 (2011).
Musil, Grisafi, Bartók, Ortner, Csányi, MC, "Physics-Inspired Structural Representations
for Molecules and Materials," Chem. Rev. 121(16), 9759-9815 (2021).
rascaline <https://github.com/Luthaf/rascaline>
metatensor <https://github.com/lab-cosmo/metatensor>