Hybrid Unsupervised-Supervised Machine Learning Models for Materials Science

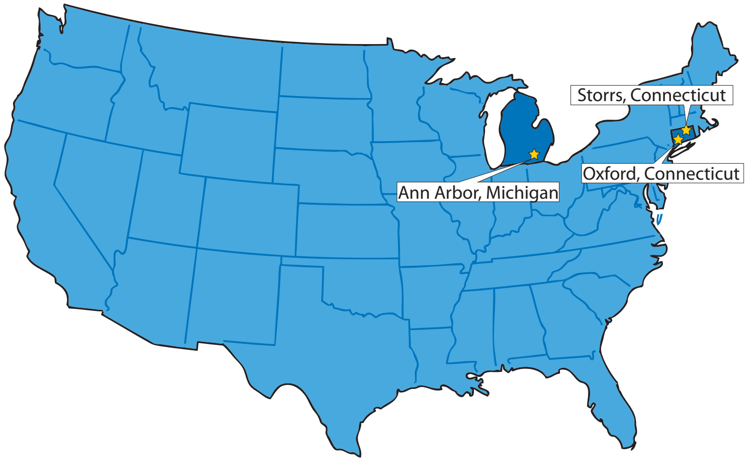
Rose K. Cersonsky

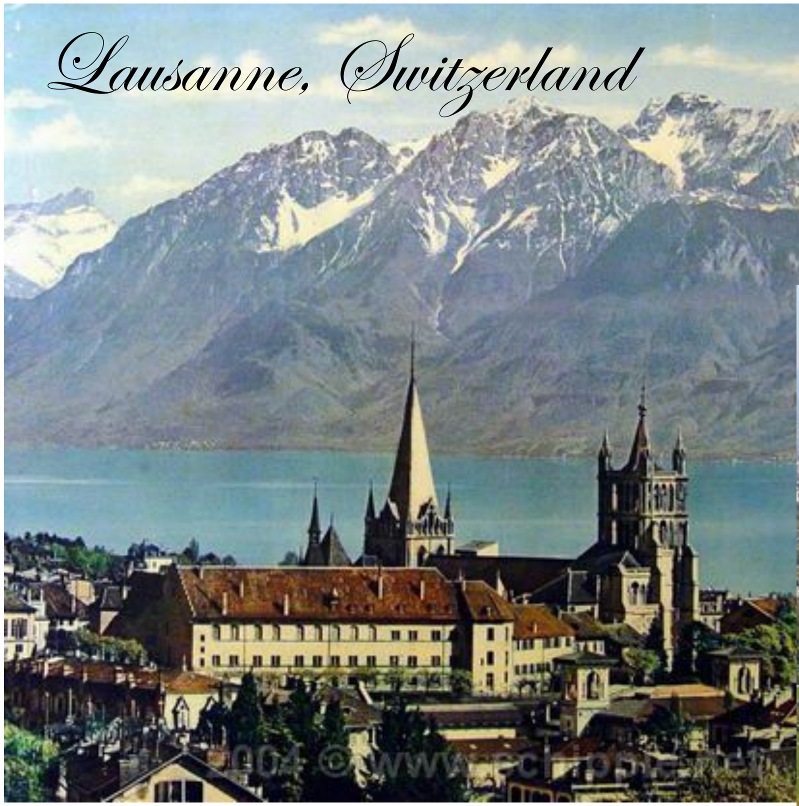
Laboratory of Computational Science and Modeling (COSMO)

École Polytechnique Fédérale de Lausanne (EPFL)

Rose Cersonsky (Sir – sahn – ski)

- Originally from Oxford, Connecticut
- University of Connecticut, Storrs, Connecticut
  - Materials Science and Engineering
  - Minor in Computer Sciences and Engineering
  - Bachelor of Science, 2014
- University of Michigan, Ann Arbor, Michigan
  - Macromolecular Science and Engineering
  - Doctor of Philosophy, 2019



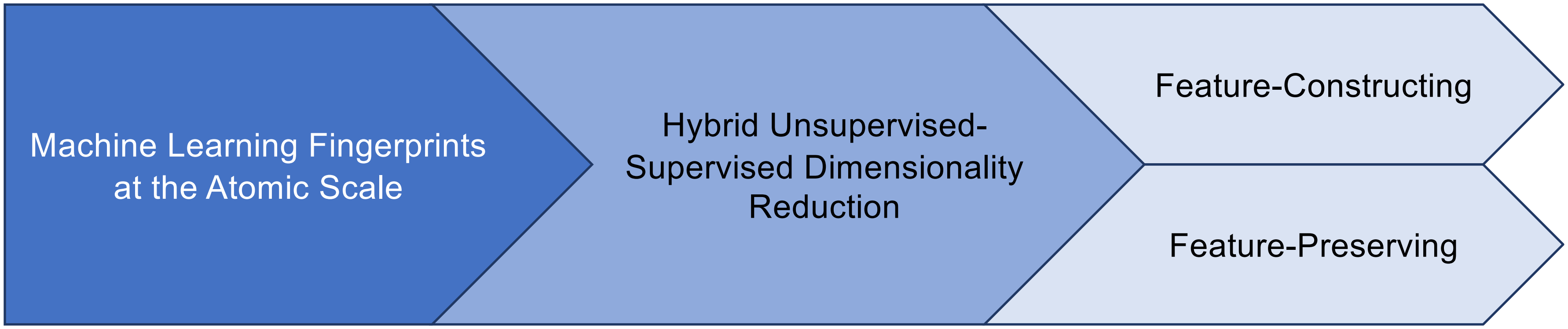


Lausanne, Switzerland

EPFL

PHYSICS-INSPIRED MACHINE LEARNING
ATOM DENSITY REPRESENTATIONS
Error Estimation
NMR
Tensor learning
FEATURE SELECTION
Potentials
predictive materials modeling at the atomic scale
ELECTRON CHARGE DENSITY
dielectric response
librascal
data-driven understanding of complex materials
METALS
interFACES
MOLECULAR
PAMM
PROBABILISTIC ANALYSIS MOLECULAR MOTIFS
sketchmap
QUANTUM ANHARMO FLUCTUAT
THERMODYNAMIC
GCH
COSMO
COMPUTATIONAL SCIENCE AND MODELLING

Hybrid Unsupervised-Supervised Machine Learning Models for Materials Science



Machine Learning Fingerprints at the Atomic Scale

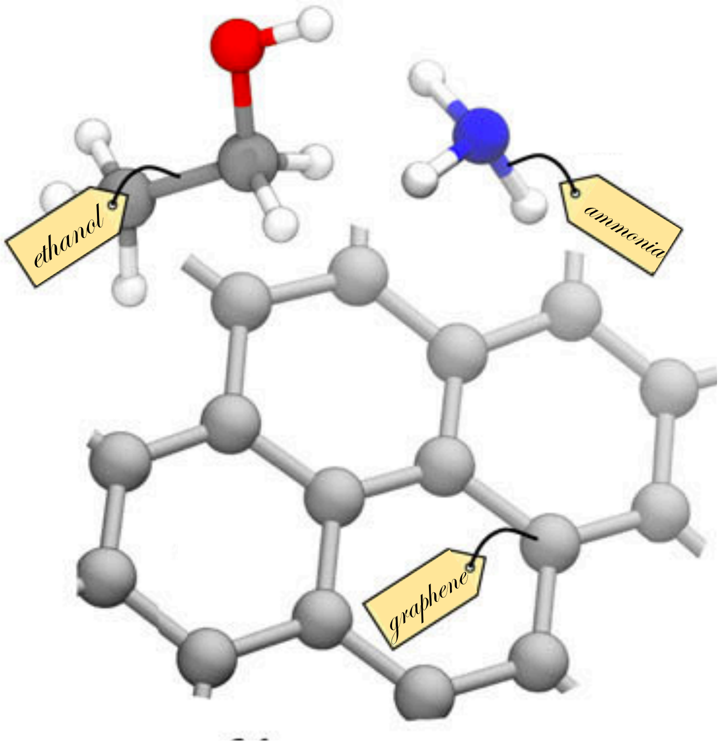
Hybrid Unsupervised-Supervised Dimensionality Reduction

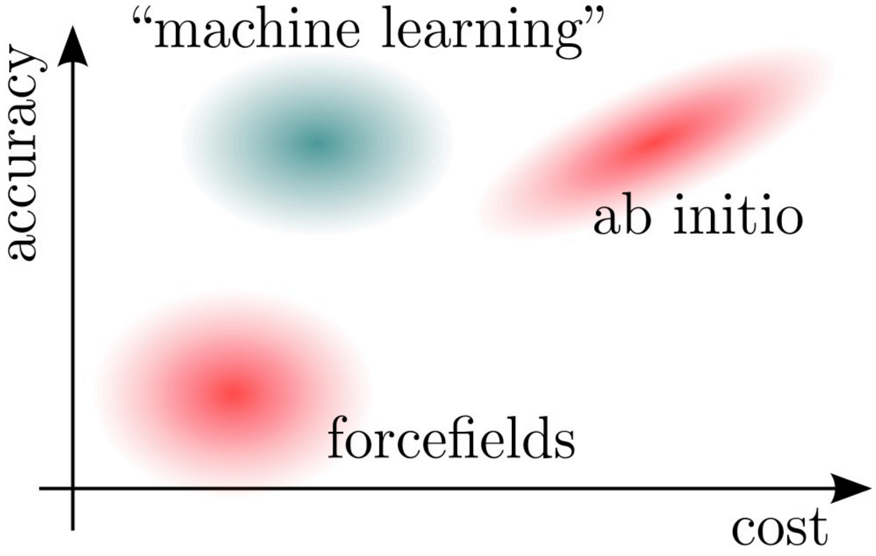
Feature-Constructing

Feature-Preserving

Machine Learning Fingerprints at the Atomic Scale

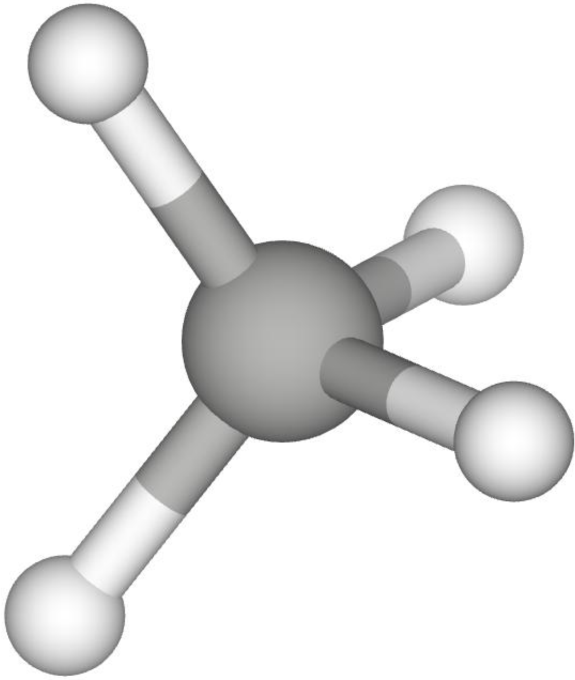
In atomic machine learning, we build models to relate the arrangement of atoms with microscopic properties.





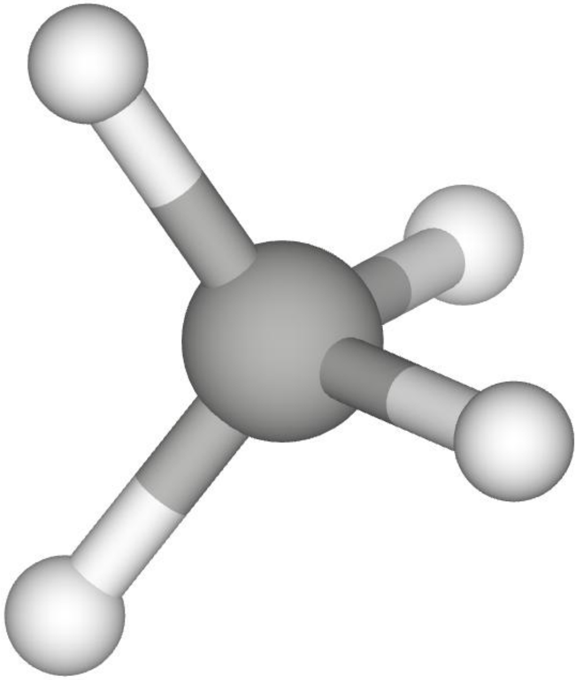
Figures courtesy of M. Ceriotti

When learning on a collection of atoms, we must encode the geometry in a numerical fingerprint which contains all relevant information.



$$\rightarrow \mathcal{X} = \{x_1, x_2, ...x_n\}$$

ML representations vary based upon the goal of the ML model, and in many cases there is a simple representation that will suffice.



$$\longrightarrow \mathcal{X} = \{4, 0, 0, 0, 0, 1, ...0\}$$

$$\underset{\text{Molar Mass}}{M} = \mathcal{X} \cdot \underset{\text{Atomic Mass}}{m_a}$$

What are the important aspects of an ML descriptor for typical atom-centered quantities?

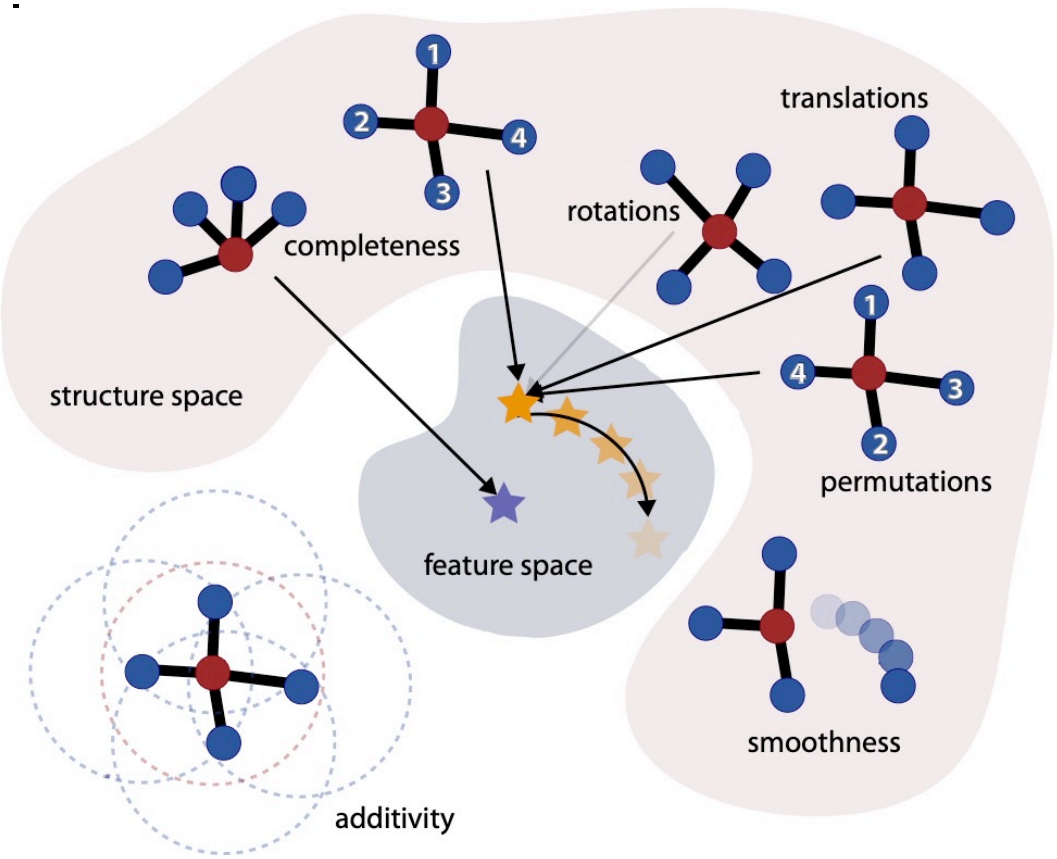


Figure adapted from: F. Musil, et al. Chem. Rev. 2021.

Many structure representations have been developed for ML of atomic-scale data.

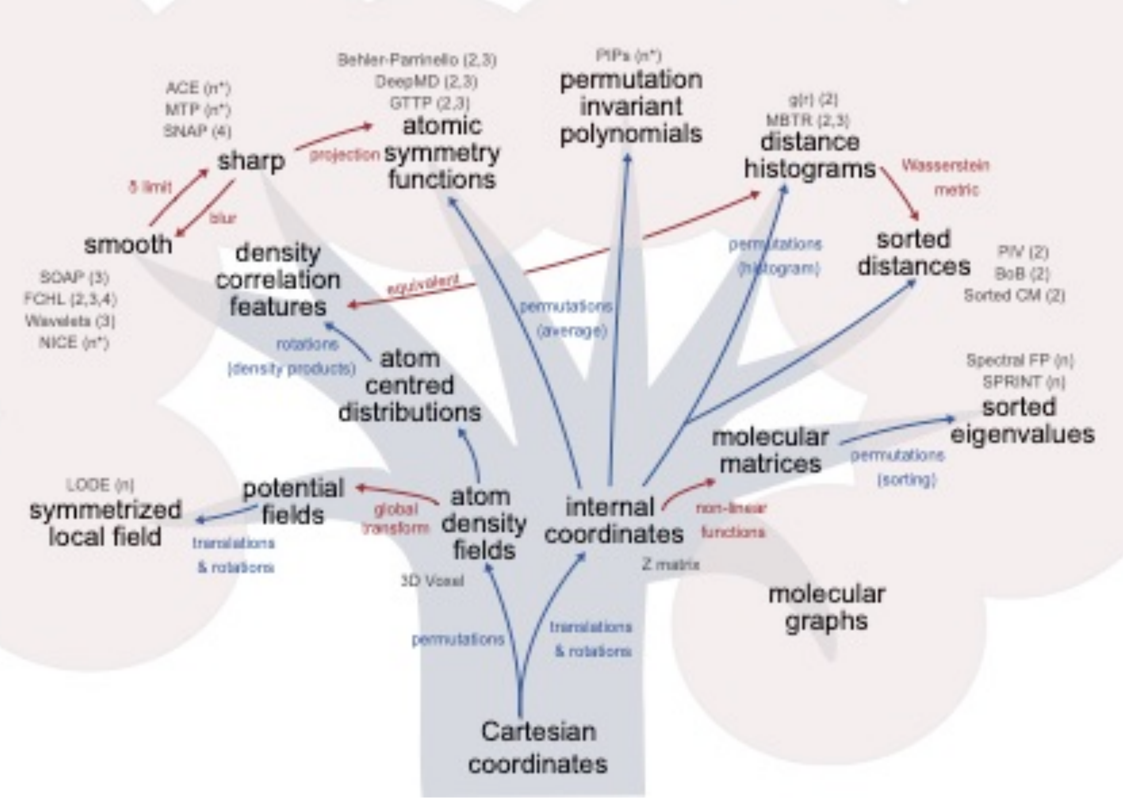


Figure adapted from: F. Musil, et al. Chem. Rev. 2021.

One way to encode the molecular geometry is by assuming a Gaussian centered on each atom.

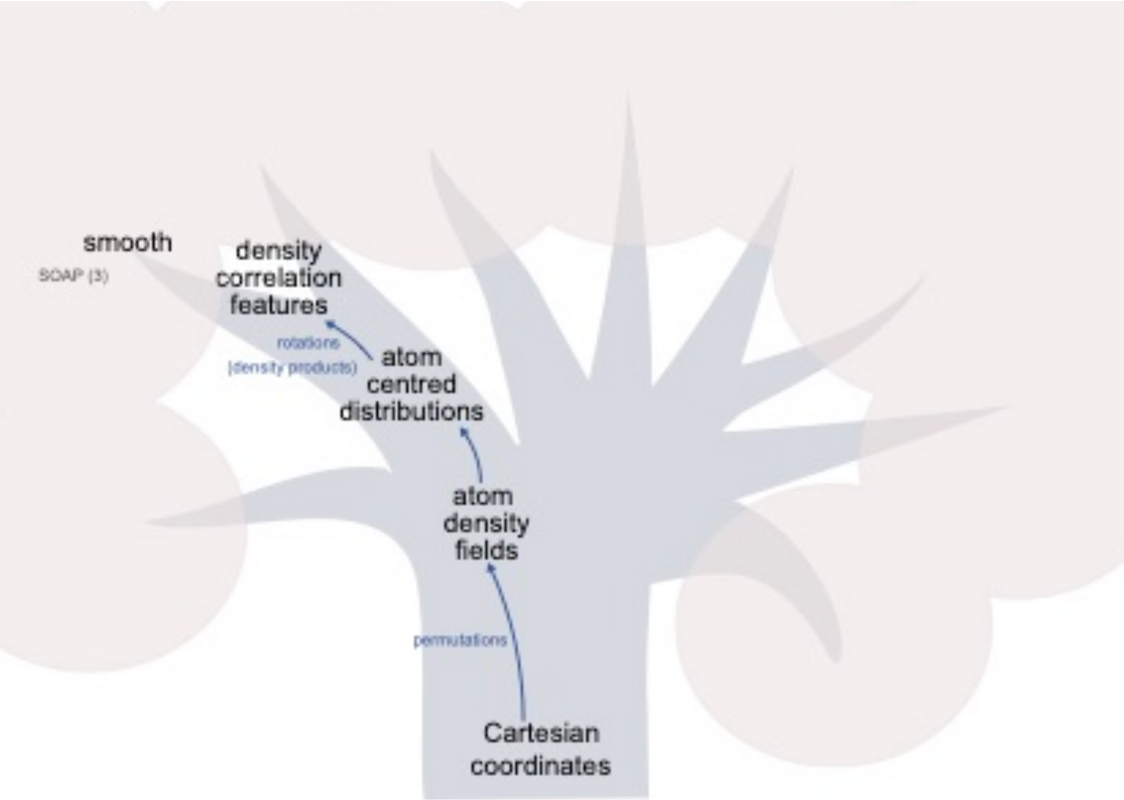
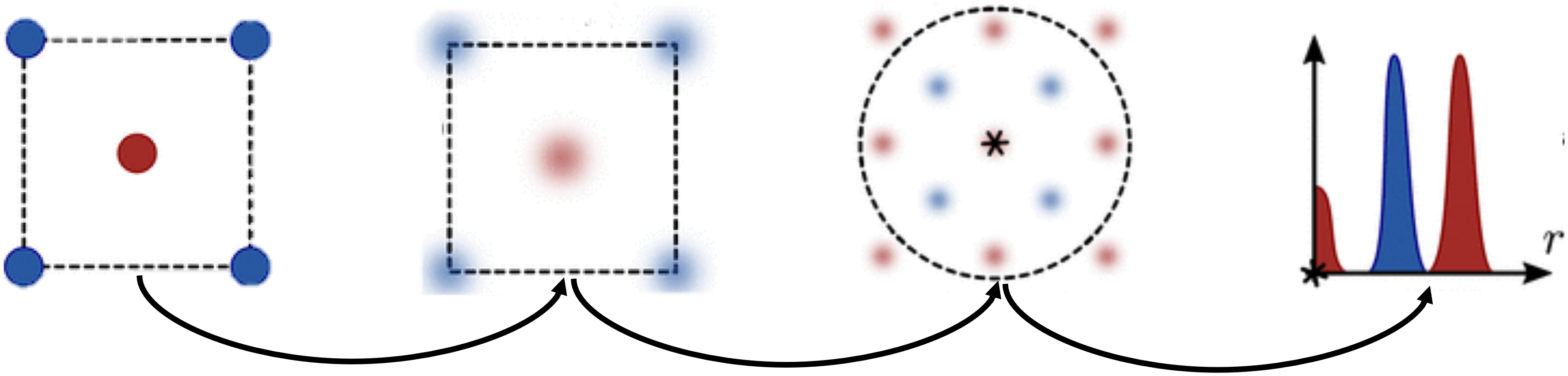


Figure adapted from: F. Musil, et al. Chem. Rev. 2021.

One way to encode the molecular geometry is by assuming a Gaussian centered on each atom, and then integrating over all translations and rotations.



density field

$\sum_i g(\mathbf{x} - \mathbf{r}_i)$

translation average

$\int d\hat{t}$

rotation average

$\int d\hat{R}$

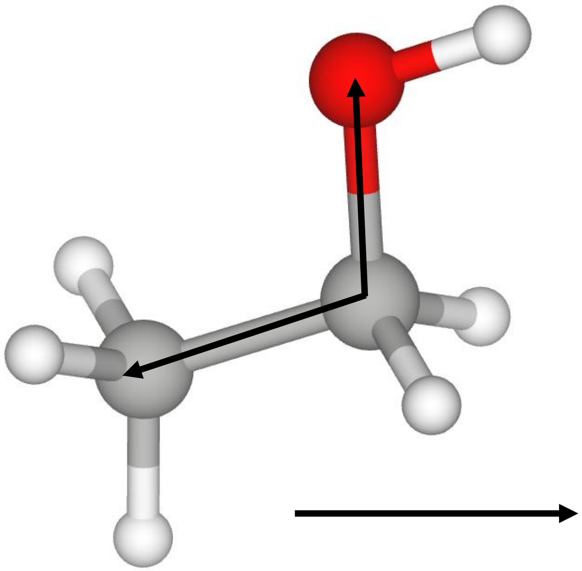
Atom Centered Density: $\langle anlm|\rho_i\rangle = \int \mathrm{d}\mathbf{r}\, \boxed{\langle \mathbf{r}|\rho_i\rangle}\,\boxed{R_n(r)}\,\boxed{Y_m^l(\hat{\mathbf{r}})}$

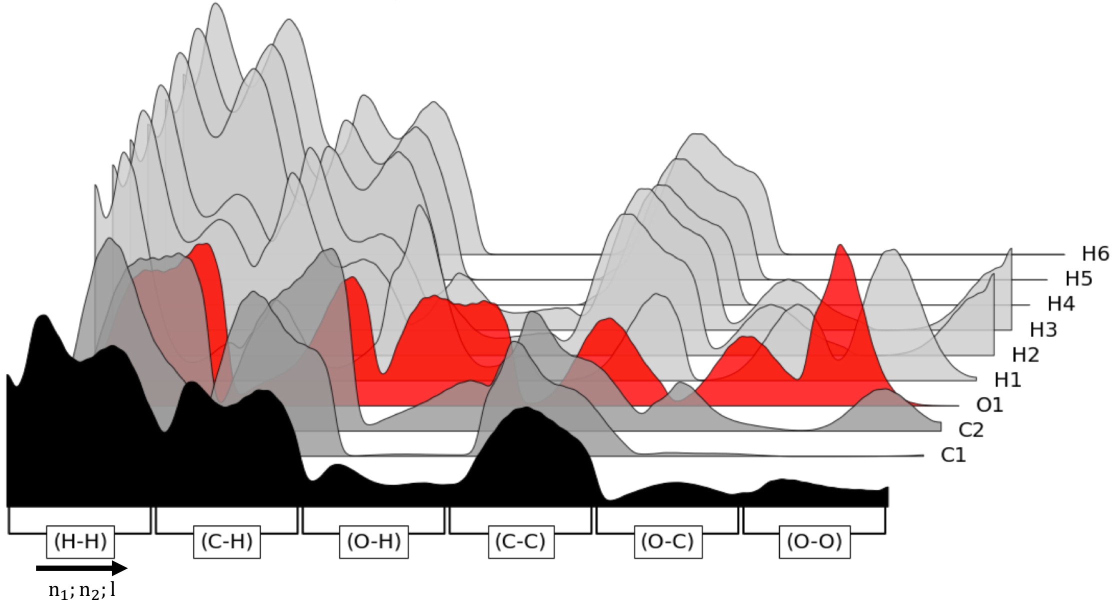
Figure adapted from: F. Musil, et al. Chem. Rev. 2021.

A popular schema for ML models of materials is the three-body SOAP (smooth overlap of atomic positions).

SOAP Vector $(3-\text{body correlation function})$:

$$\frac{1}{\sqrt{2l+1}}\sum_{m}(-1)^{m}\langle a_1 n_1 lm|\rho_i\rangle\langle a_2\, n_2 l(-m)|\rho_i\rangle$$



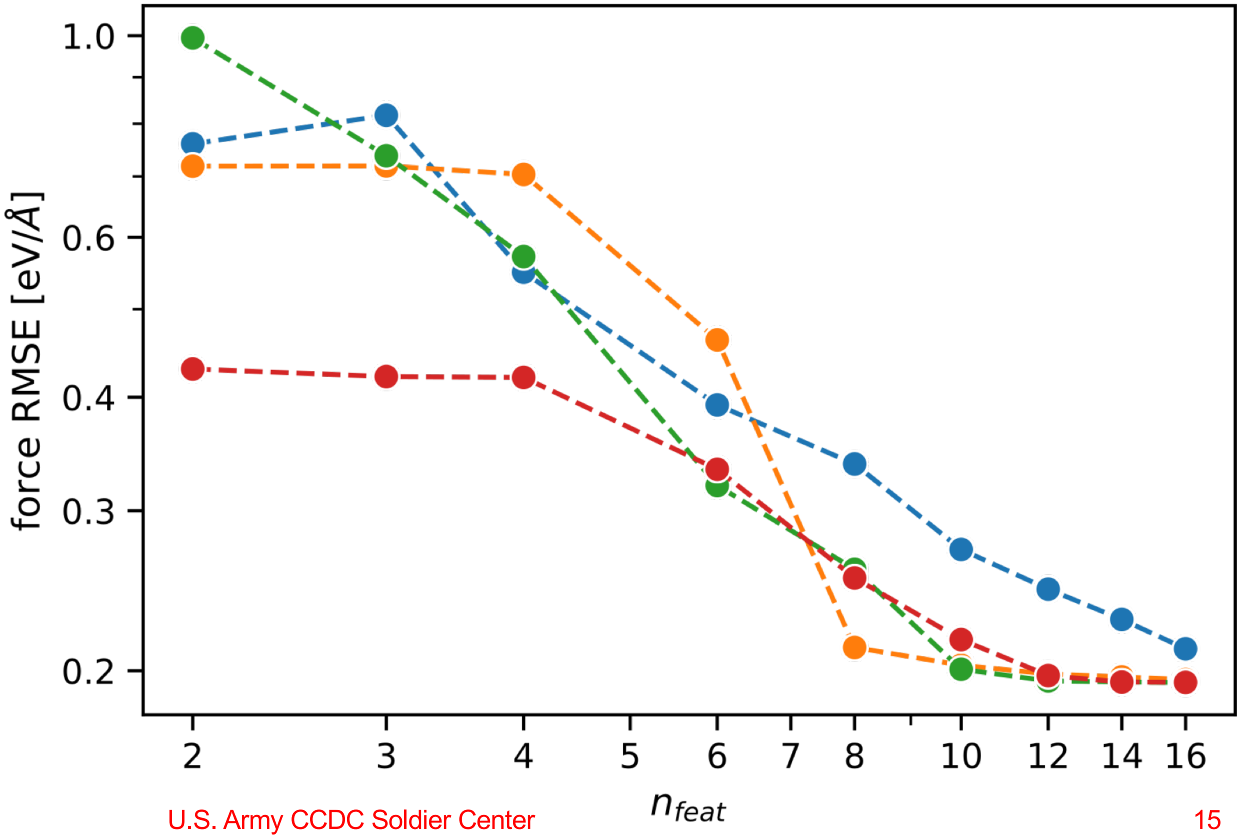


How do we know which featurization to use?

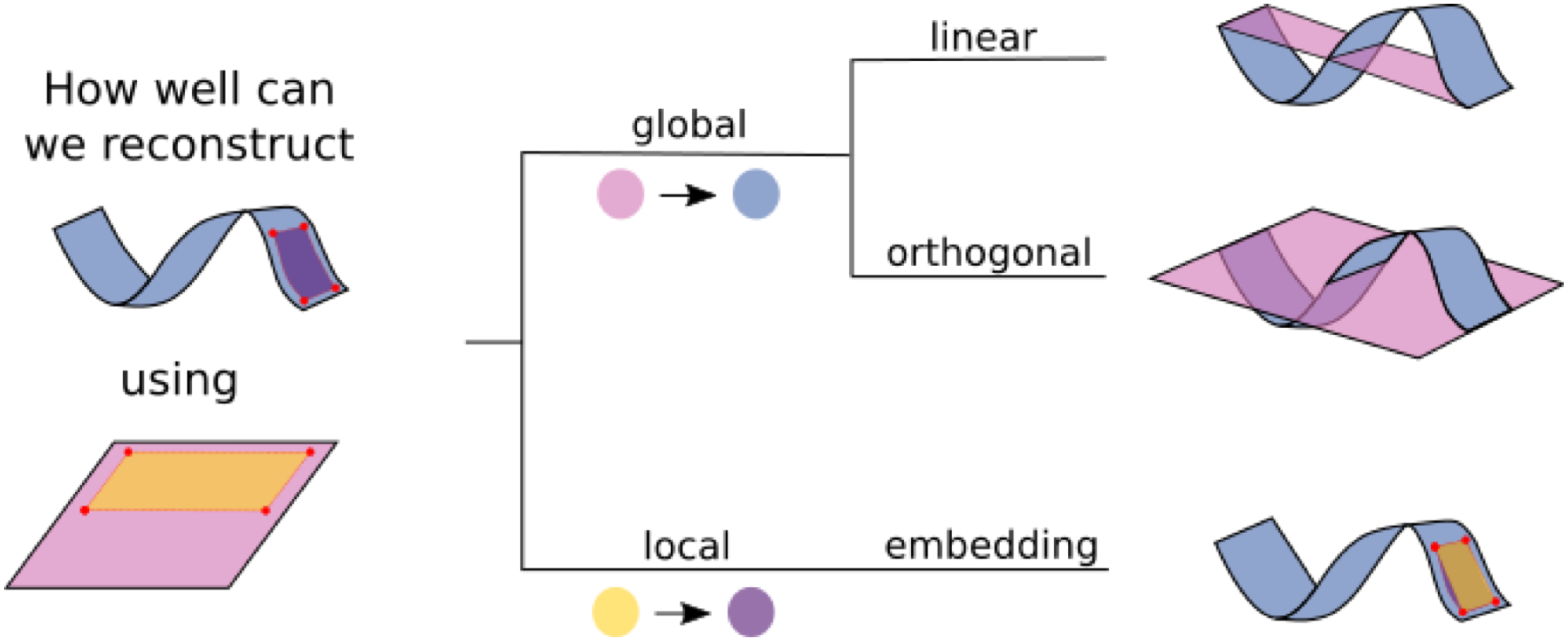
Roughly speaking, better features lead to better predictions

We can compare features with respect to a property like forces

But how do we compare features independent from properties?

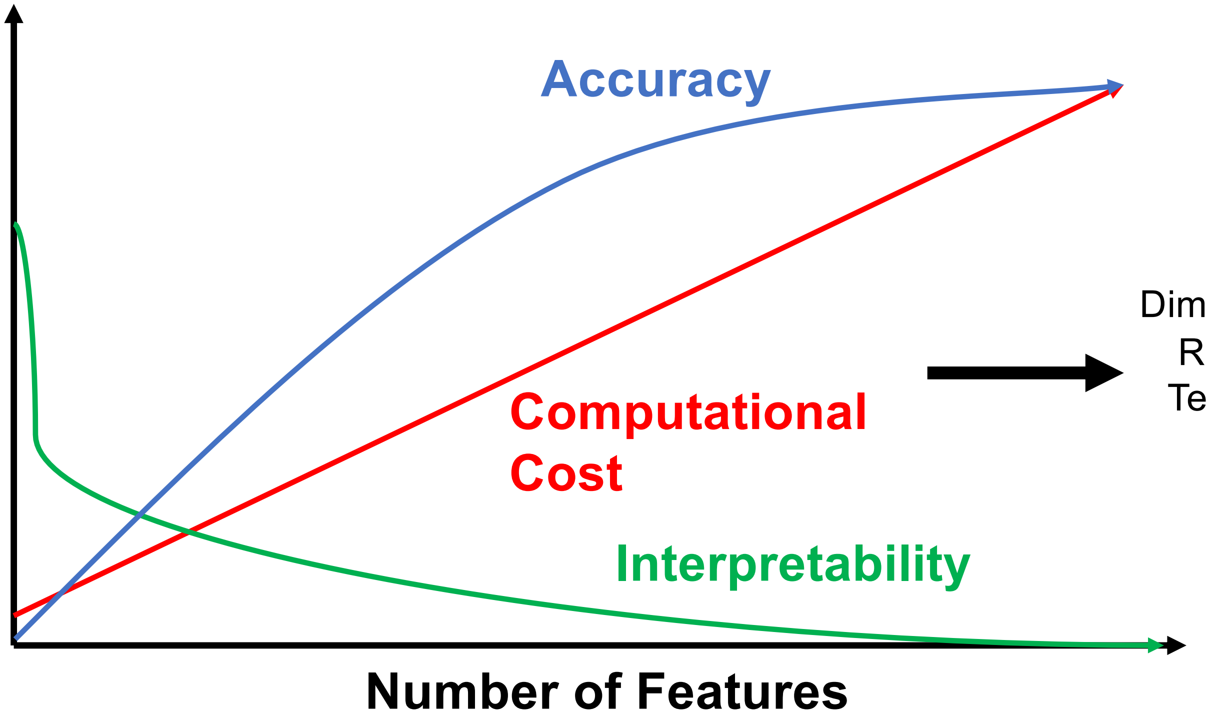


We can use feature reconstruction measures to compare features representing the same structures.



Goscinski, A., et al. (2021). The role of feature space in atomistic learning. *Machine Learning: Science and Technology*, *2*(2), 025028.

Why not just use the most extensive set of features?



→ Dimensionality Reduction Techniques

Hybrid Unsupervised-Supervised Dimensionality Reduction

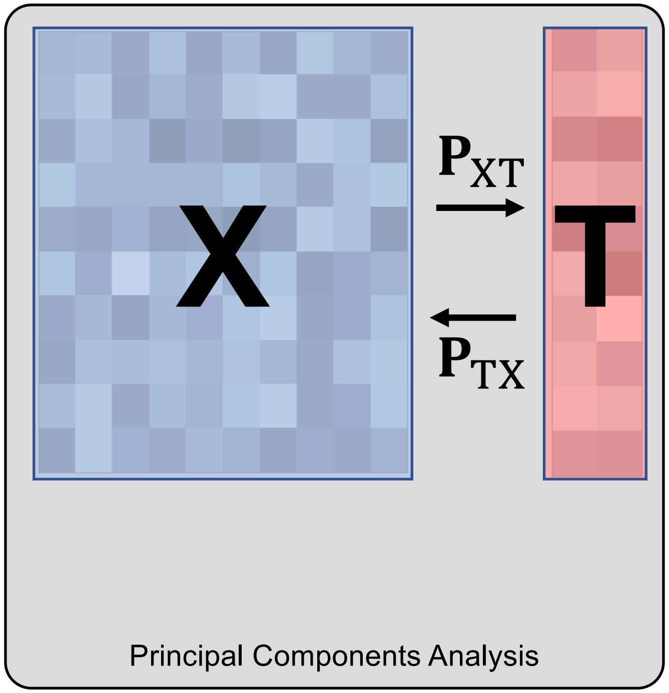
A couple words on notation…

$\mathbf{X} = \begin{bmatrix}\mathbf{x}_1 \\ \mathbf{x}_2 \\ \dots\end{bmatrix}$	A matrix containing as rows **the fingerprints of a set of structures**
$\mathbf{Y} = \begin{bmatrix}\mathbf{y}_1 \\ \mathbf{y}_2 \\ \dots\end{bmatrix}$	A matrix containing as rows **the target properties for a set of structures**

$\mathbf{P}_{\mathrm{AB}}$	A matrix that projects from space $\mathbf{A}$ to space $\mathbf{B}$
$\mathbf{T} = \mathbf{X}\mathbf{P}_{\mathrm{XT}}$	A matrix containing as rows **the latent-space projection of a set of structures**

Principal Components Analysis (PCA)

PCA determines an information-rich set of features to represent a larger set of features.



$$\ell = \quad \|\mathbf{X} - \mathbf{X}\,\mathbf{P}_{XT}\,\mathbf{P}_{TX}\|^2$$

loss

This is solved by constructing the projectors from the eigendecomposition of either the Gram matrix K or the covariance C (analogous to the SVD of X)

$$\boxed{\mathbf{K} = \mathbf{X}\mathbf{X}^T}$$

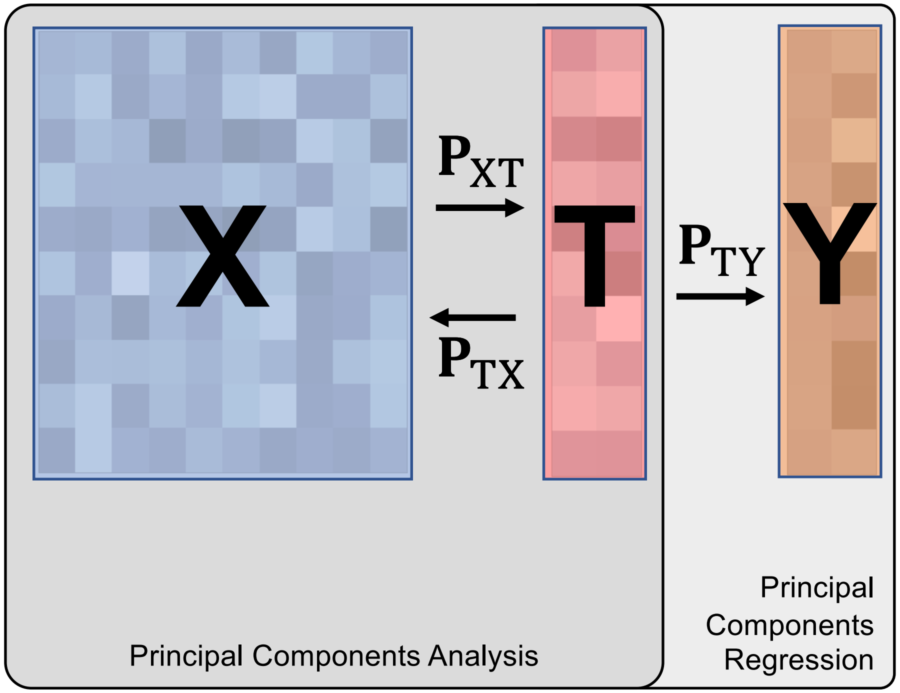
gram matrix

$$\boxed{\mathbf{C} = \mathbf{X}^T\mathbf{X}}$$

covariance matrix

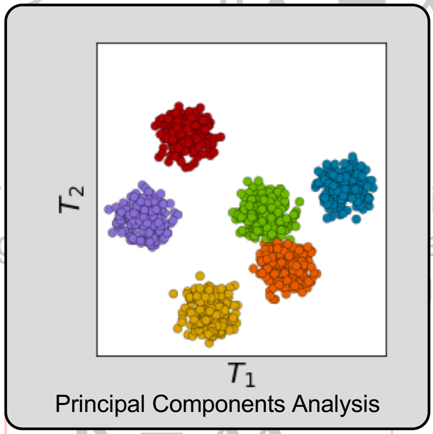
Principal Components Analysis (PCA)

PCA determines an information-rich set of features to represent a larger set of features.
PCR uses this set of features to predict a target.

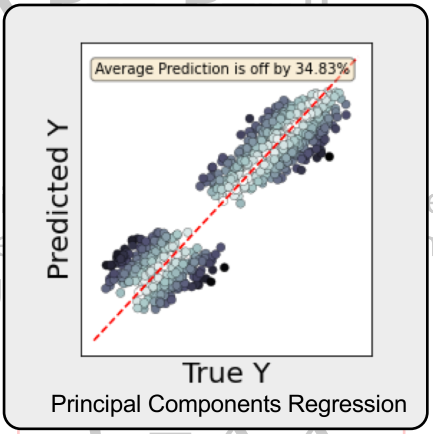


Principal Components Analysis

Principal Components Regression



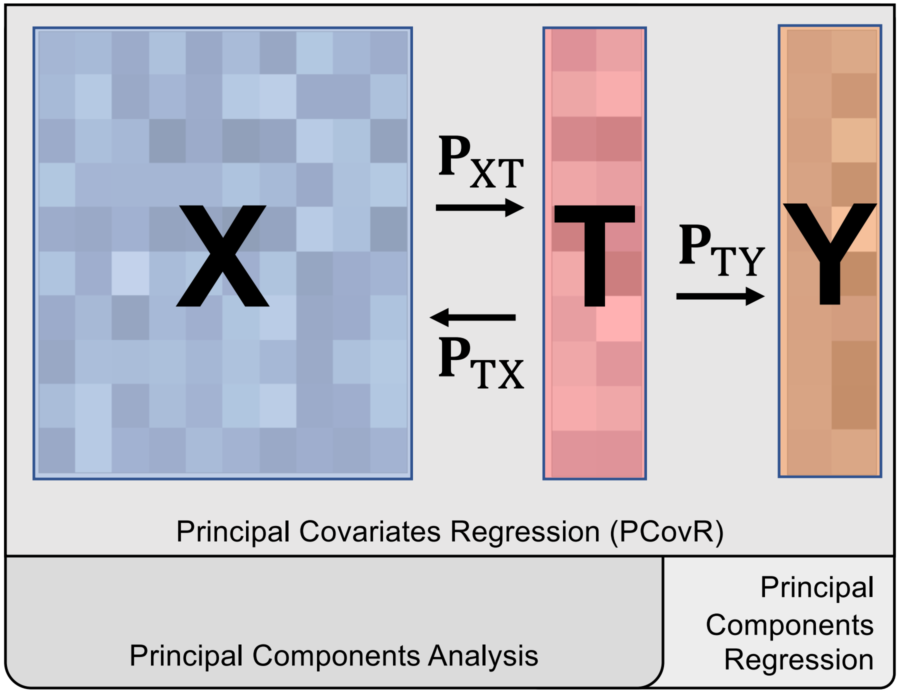
Principal Components Analysis



Principal Components Regression

Principal Covariates Regression (PCovR)

PCovR determines an information rich set of features to represent a larger set of features *and* optimally regress a set of targets.


X
$\mathbf{P}_{\mathrm{XT}}$
$\mathbf{P}_{\mathrm{TX}}$
T
$\mathbf{P}_{\mathrm{TY}}$
Y

Principal Covariates Regression (PCovR)

Principal Components Analysis

Principal Components Regression

loss in reconstructing X

$$\ell = \alpha\|\mathbf{X} - \mathbf{X}\,\mathbf{P}_{\mathrm{XT}}\,\mathbf{P}_{\mathrm{TX}}\|^2$$
$$+(1-\alpha)\|\mathbf{Y} - \mathbf{X}\,\mathbf{P}_{\mathrm{XT}}\,\mathbf{P}_{\mathrm{TY}}\|^2$$

loss in reconstructing Y

This is solved by constructing the projectors from the eigendecomposition of either a **modified** Gram matrix or a **modified** covariance

$$\boxed{\mathbf{K} \rightarrow \widetilde{\mathbf{K}}}$$

$$\boxed{\widetilde{\mathbf{K}} = \boldsymbol{\alpha}\mathbf{X}\mathbf{X}^{\mathrm{T}} + (\mathbf{1}-\boldsymbol{\alpha})\hat{\mathbf{Y}}\hat{\mathbf{Y}}^{\mathbf{T}}}$$

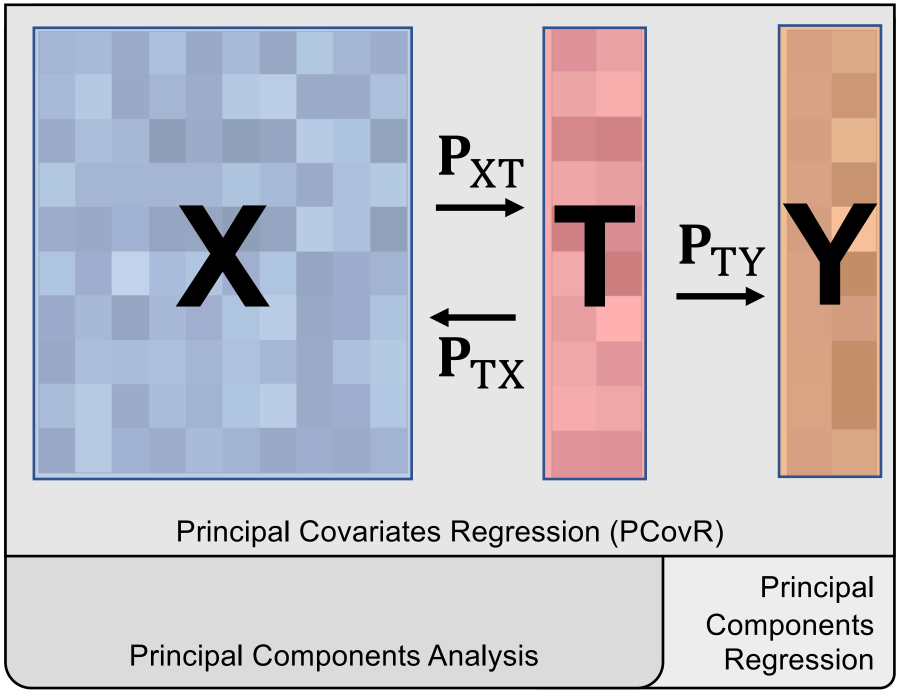
$$\boxed{\mathbf{C} \rightarrow \tilde{\mathbf{C}}}$$

$$\boxed{\tilde{\mathbf{C}} = (\mathbf{C}^{-\mathbf{1}/\mathbf{2}}\mathbf{X}^{\mathbf{T}})\widetilde{\mathbf{K}}(\mathbf{X}\mathbf{C}^{-\mathbf{1}/\mathbf{2}})}$$

B. A. Helfrecht, **RKC**, G. Fraux, and M. Ceriotti. 2020 Mach. Learn.: Sci. Technol. 1 045021
S. de Jong, H.A.L. Kiers, Chemom. intell. lab. syst. 14 (1992) 155-164.
scikit-cosmo.readthedocs.io

Principal Covariates Regression (PCovR)

PCovR determines an information rich set of features to represent a larger set of features *and* optimally regress a set of targets.



Principal Covariates Regression (PCovR)

Principal Components Analysis

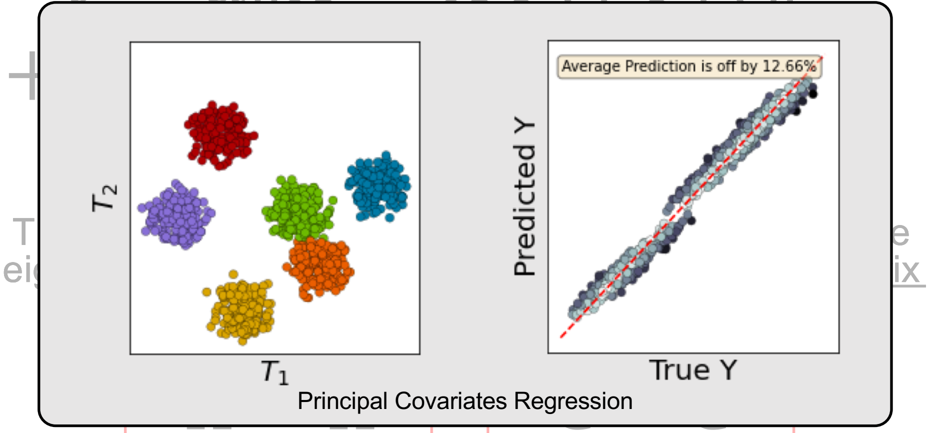
Principal Components Regression

$$\ell = \alpha\|\mathbf{X} - \mathbf{X}\,\mathbf{P}_{XT}\,\mathbf{P}_{TX}\|^2$$

+

T ei

e ix



Principal Covariates Regression

$\tilde{\mathbf{K}} = \alpha\mathbf{X}\mathbf{X}^T + (1-\alpha)\hat{\mathbf{Y}}\hat{\mathbf{Y}}^T$

$\tilde{\mathbf{C}} = (\mathbf{C}^{-1/2}\mathbf{X}^T)\tilde{\mathbf{K}}(\mathbf{X}\mathbf{C}^{-1/2})$

B. A. Helfrecht, **RKC**, G. Fraux, and M. Ceriotti. 2020 Mach. Learn.: Sci. Technol. 1 045021
S. de Jong, H.A.L. Kiers, Chemom. intell. lab. syst. 14 (1992) 155-164.
scikit-cosmo.readthedocs.io

Kernel Principal Covariates Regression

Core to PCA / PCovR is the gram kernel, which is equivalent to the linear kernel.
We can replace this with any number of non-linear kernels to better represent non-linear structure-property relations.

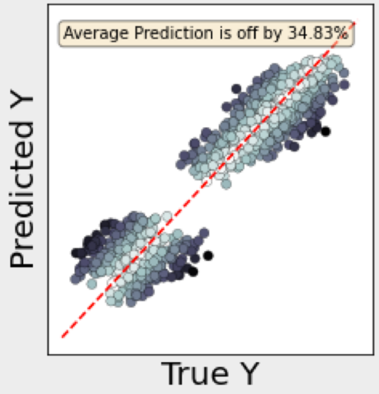
gram matrix,
a.k.a. “linear kernel”

$$\mathbf{K} = \mathbf{X}\mathbf{X}^{\mathrm{T}}$$

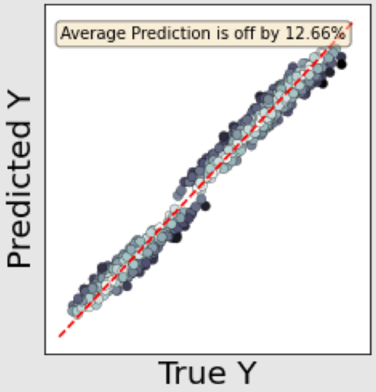
$$\mathrm{K}_{\mathrm{ij}} = \mathrm{k}\left(\mathbf{x}_{\mathrm{i}}, \mathbf{x}_{\mathrm{j}}\right) = \mathbf{x}_{\mathrm{i}} \cdot \mathbf{x}_{\mathrm{j}}$$

$$\mathbf{e}^{-\gamma\left\|\mathbf{x}_{\mathrm{i}}-\mathbf{x}_{\mathrm{j}}\right\|^{2}}$$

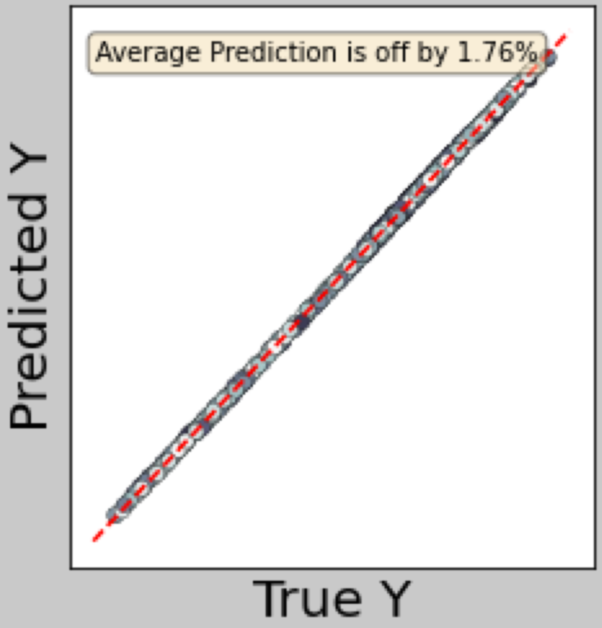
non-linear kernel



Principal **Components** Regression



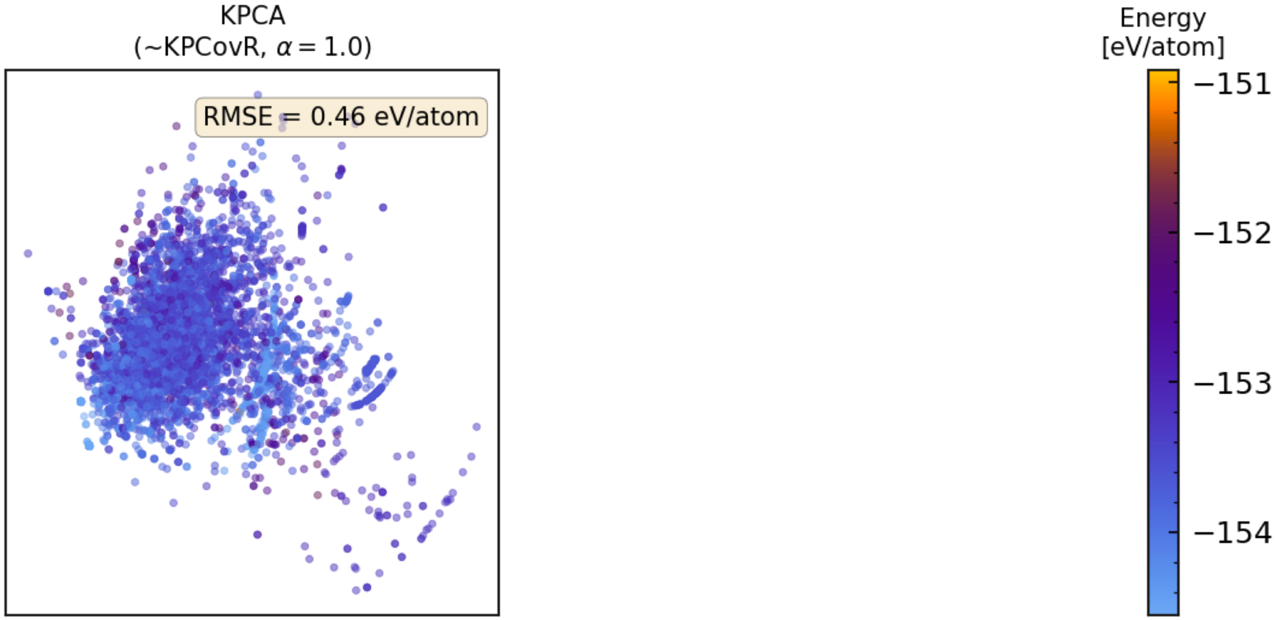
Principal **Covariates** Regression



Kernel Principal Covariates Regression

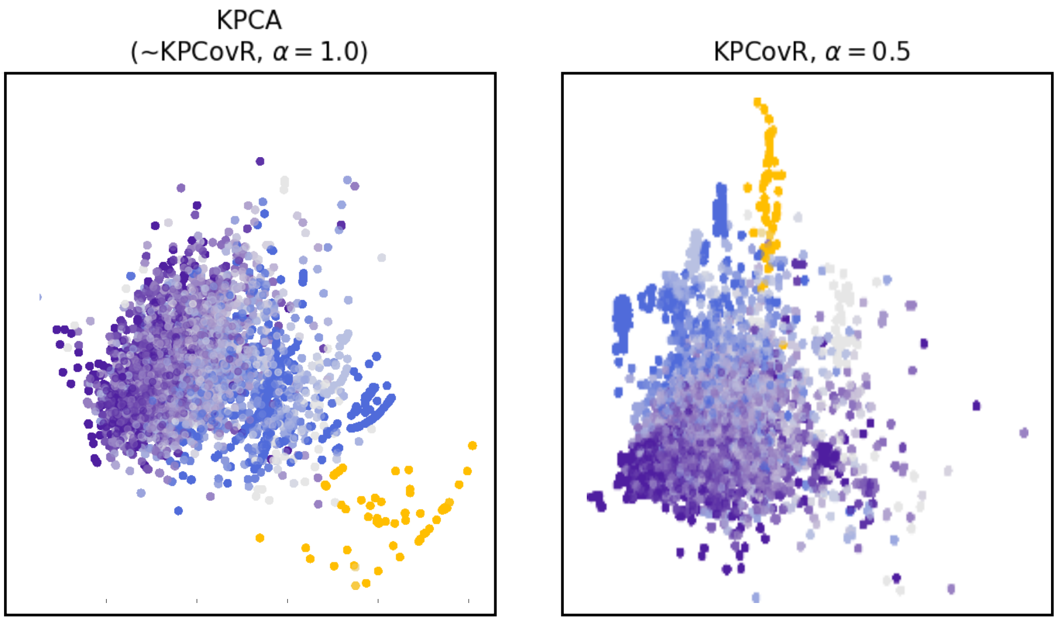
B. A. Helfrecht, **RKC**, G. Fraux, and M. Ceriotti. 2020 Mach. Learn.: Sci. Technol. 1 045021
S. de Jong, H.A.L. Kiers, Chemom. intell. lab. syst. 14 (1992) 155-164.
scikit-cosmo.readthedocs.io

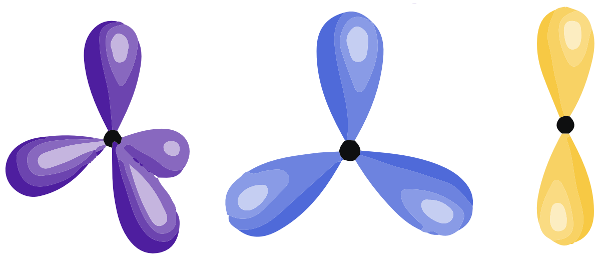
Analysis of SOAP features of Ab-Initio Random Structure Search (AIRSS) carbon crystals and their energies in eV/atom



B. A. Helfrecht, **RKC**, G. Fraux, and M. Ceriotti. 2020 Mach. Learn.: Sci. Technol. 1 045021
C. J. Pickard. AIRSS Data for Carbon at 10GPa and the C+N+H+O System at 1GPa (2020).

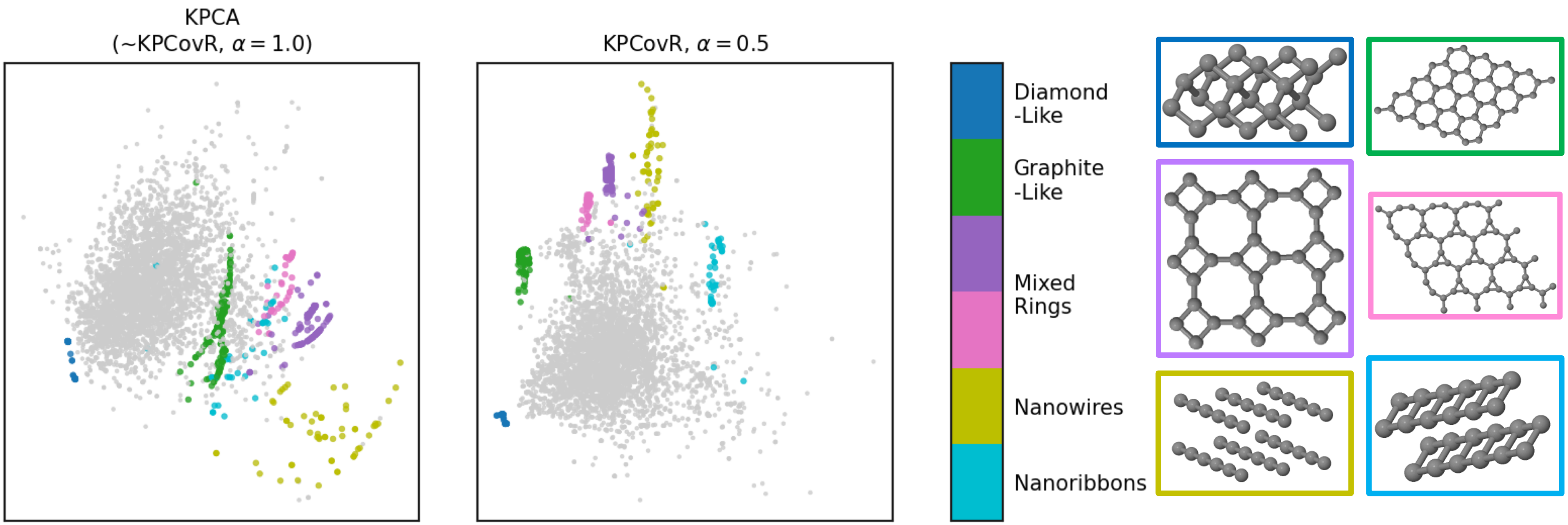
Analysis of SOAP features of Ab-Initio Random Structure Search (AIRSS) carbon crystals and their energies in eV/atom





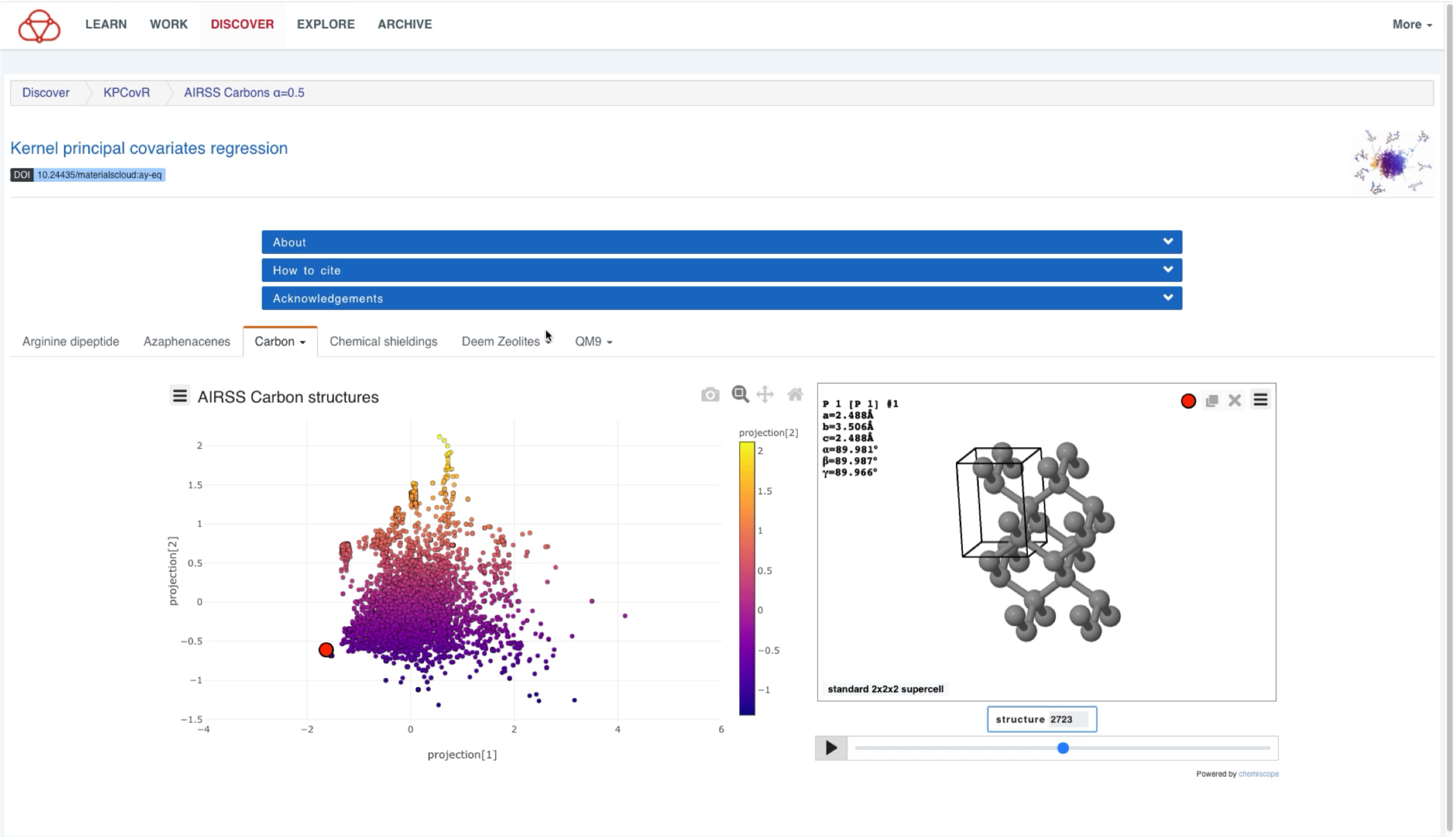
B. A. Helfrecht, **RKC**, G. Fraux, and M. Ceriotti. 2020 Mach. Learn.: Sci. Technol. 1 045021
C. J. Pickard. AIRSS Data for Carbon at 10GPa and the C+N+H+O System at 1GPa (2020).

Analysis of SOAP features of Ab-Initio Random Structure Search (AIRSS) carbon crystals and their energies in eV/atom



B. A. Helfrecht, **RKC**, G. Fraux, and M. Ceriotti. 2020 Mach. Learn.: Sci. Technol. 1 045021
C. J. Pickard. AIRSS Data for Carbon at 10GPa and the C+N+H+O System at 1GPa (2020).

https://www.materialscloud.org/discover/kpcovr/carbons-05



Kernel principal covariates regression

DOI 10.24435/materialscloud:ay-eq

P 1 [P 1] #1
a=2.488Å
b=3.506Å
c=2.488Å
α=89.981°
β=89.987°
γ=89.966°

standard 2x2x2 supercell

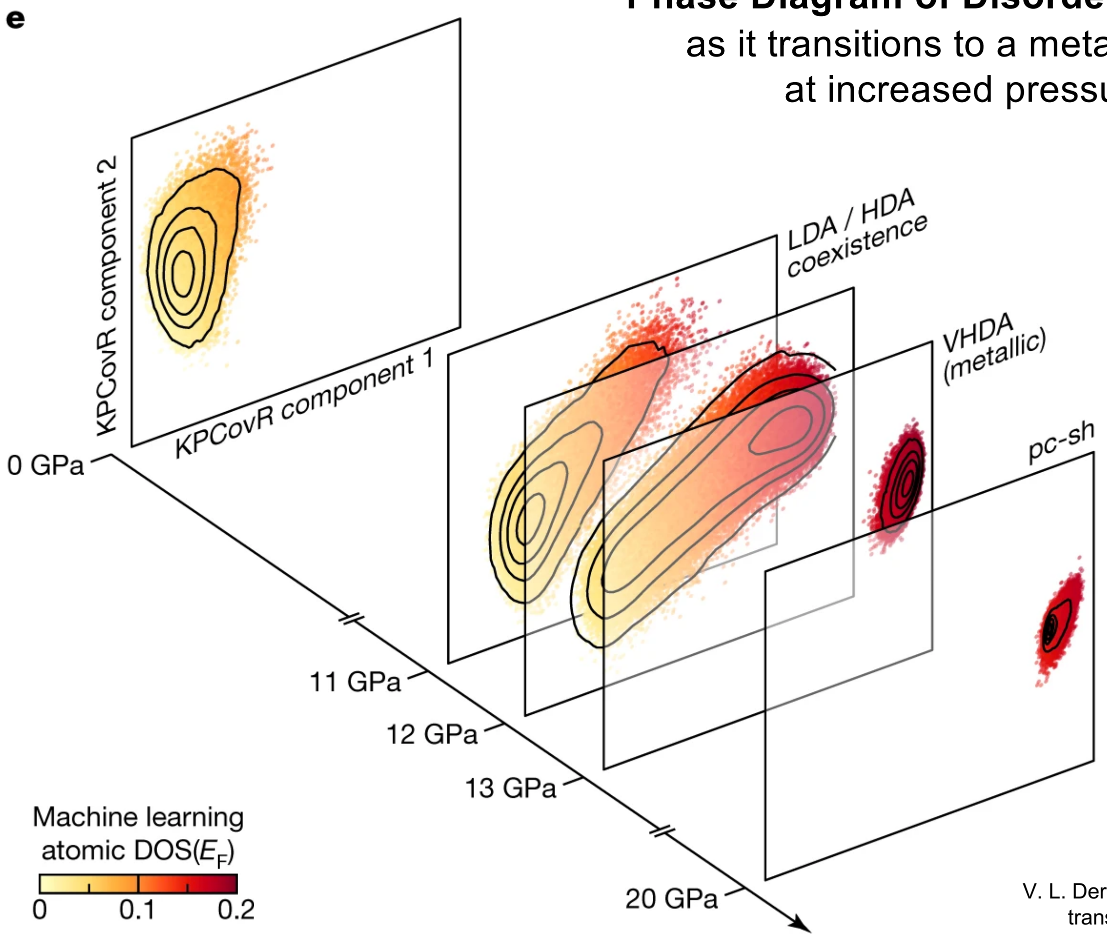
structure 2723

B. A. Helfrecht, **RKC**, G. Fraux, and M. Ceriotti. 2020 Mach. Learn.: Sci. Technol. 1 045021
G. Fraux, **RKC**, M. Ceriotti. 2020. Journal of Open Source Software, 5(51), 2117.
B. A. Helfrecht, **RKC**, G. Fraux, and M. Ceriotti. Materials Cloud Archive 2020.80 (2020).

Powered by chemiscope.org

Phase Diagram of Disordered Silicon

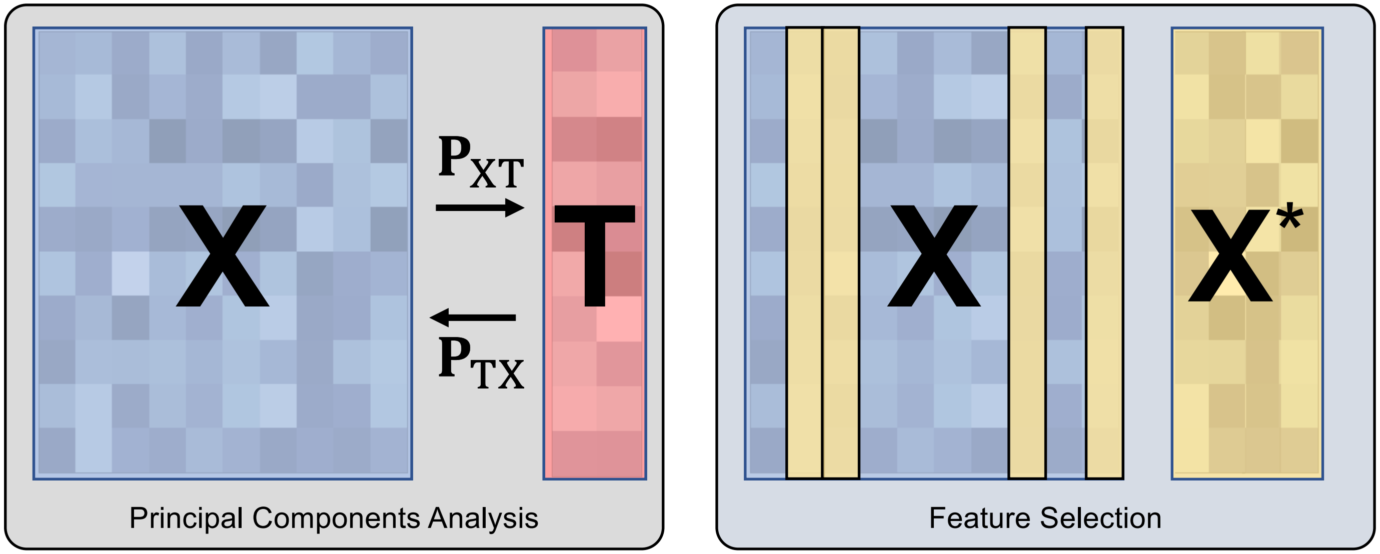
as it transitions to a metallic state
at increased pressure



V. L. Deringer, et al., Origins of structural and electronic transitions in disordered silicon. Nature 589, 59–64 (2021). , pages 59–64 (2021).

What if the features carry inherent meaning?

Many dimensionality reduction techniques construct a *new* set of features, but what if you want to just work with a subset of the old set?



A few more words on notation…

$\mathbf{X_c}$	A selection of columns from $\mathbf{X}$
$\mathbf{X_r}$	A selection of rows from $\mathbf{X}$

$\mathbf{A^-}$	The pseudo-inverse of $\mathbf{A}$
$\mathbf{U_A}$	A matrix containing the eigenvectors of $\mathbf{A}$

Data Sub-selection carries two components: the metric and the wrapper

This is true of both feature and sample sub-selection

Metric

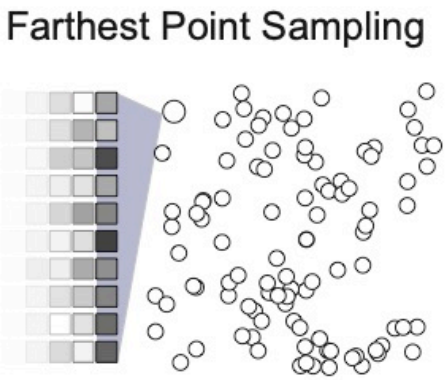
Area
Population
Median Income
Date Established
# of hipster coffee shops
# of lakes that they touch

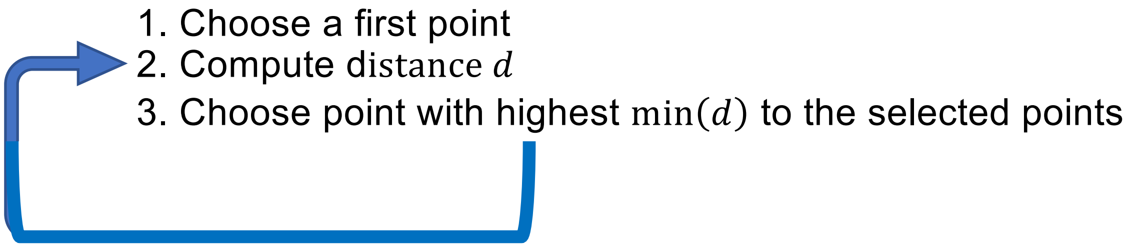
Wrapper

Smallest → Largest
Largest → Smallest
Representative of the Overall Distribution

Farthest Point Sampling (FPS)

FPS aims to select a diverse subset of features or samples that cover the greatest portion of sample or feature space.

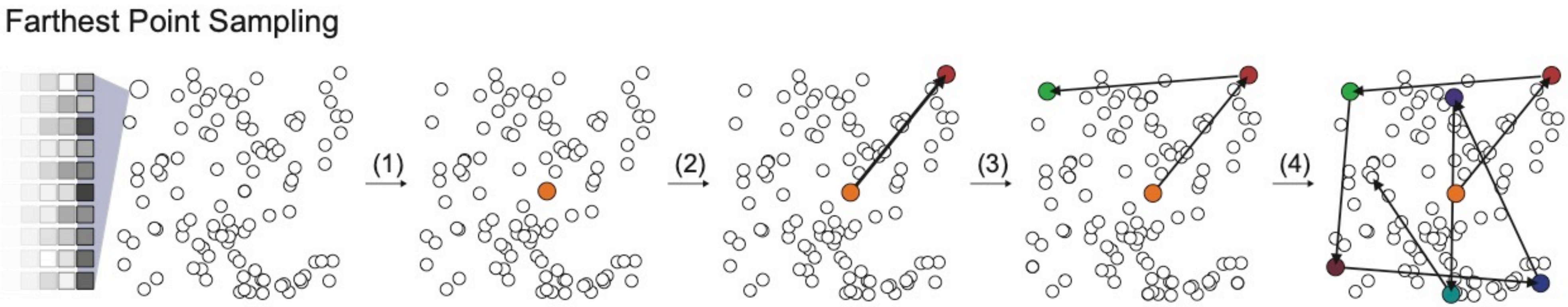




1. Choose a first point
2. Compute distance d
3. Choose point with highest $\min(d)$ to the selected points

Farthest Point Sampling (FPS)

FPS aims to select a diverse subset of features or samples that cover the greatest portion of sample or feature space.



Feature Selection

$$d_{ij} = \left\| \mathbf{X}_i - \mathbf{X}_j \right\|^2$$

$$d_{ij} = C_{ii} - 2C_{ij} + C_{jj}$$

covariance matrix

$$\mathbf{C} = \mathbf{X}^{\mathrm{T}}\mathbf{X}$$

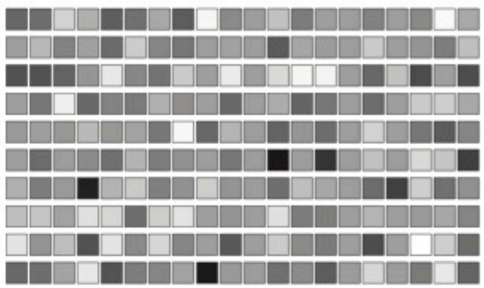
CUR Decomposition

Traditional CUR decomposition selection aims to select “important” features or samples from the overall distribution.

Head Canon:
“Columns und Rows” decomposition

$$\hat{\mathbf{X}} = \mathbf{X}_{\mathrm{c}}(\mathbf{X}_{\mathrm{c}}^{-}\mathbf{X}\,\mathbf{X}_{\mathrm{r}}^{-})\mathbf{X}_{\mathrm{r}}$$

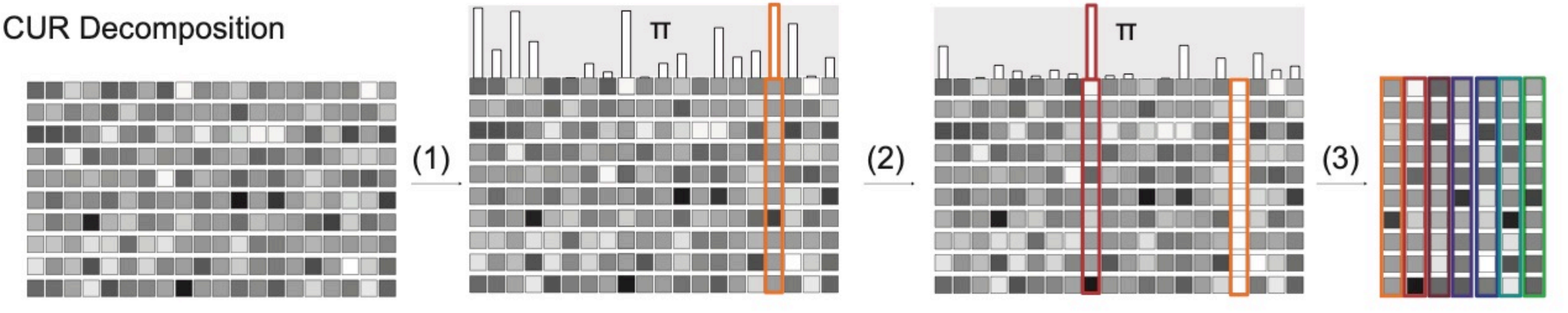
CUR Decomposition



1. Compute importance score π
2. Choose column with highest π
3. Orthogonalize with respect to last chosen column.

CUR Decomposition

Traditional CUR decomposition selection aims to
select “important” features or samples from the overall distribution.



Feature Selection

$$\pi_j = \sum_i^k (\mathbf{U_C})^2_{ij} \,.$$

covariance matrix

$\mathbf{C} = \mathbf{X}^{\mathbf{T}}\mathbf{X}$

Both FPS and CUR use feature metrics that can be written in terms of feature covariances (C).

Farthest Point Sampling (FPS)

FPS aims to select a diverse subset of features or samples that cover the greatest portion of sample or feature space.

Feature Selection

$$\mathrm{d}_{ij} = \mathrm{C}_{ii} - 2\mathrm{C}_{ij} + \mathrm{C}_{jj}$$

CUR Decomposition

Traditional CUR decomposition selection aims to select “important” features or samples from the overall distribution.

Feature Selection

$$\pi_j = \sum_i^k \left(\mathbf{U}_\mathbf{C}\right)_{ij}^2 .$$

Both FPS and CUR can be adapted to use PCovR-style covariances.

Farthest Point Sampling (FPS)

FPS aims to select a diverse subset of features or samples that cover the greatest portion of sample or feature space.

Feature Selection

$$\tilde{\mathrm{d}}_{ij} = \tilde{\mathrm{C}}_{ii} - 2\tilde{\mathrm{C}}_{ij} + \tilde{\mathrm{C}}_{jj}$$

CUR Decomposition

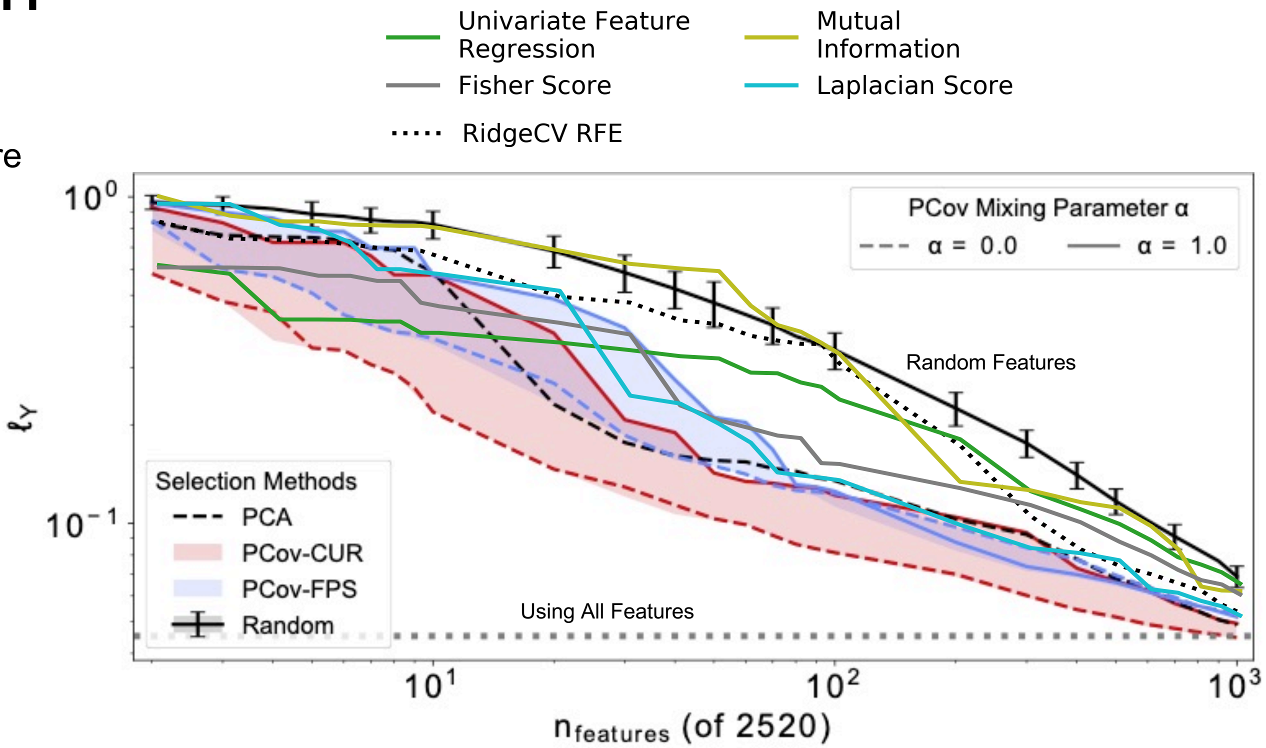
Traditional CUR decomposition selection aims to select “important” features or samples from the overall distribution.

Feature Selection

$$\pi_j = \sum_i^k \left(\mathbf{U}_{\tilde{\mathbf{C}}}\right)_{ij}^2 .$$

Linear Regression

Using PCov-style feature selection will universally out-perform common feature selection metrics available via popular packages.

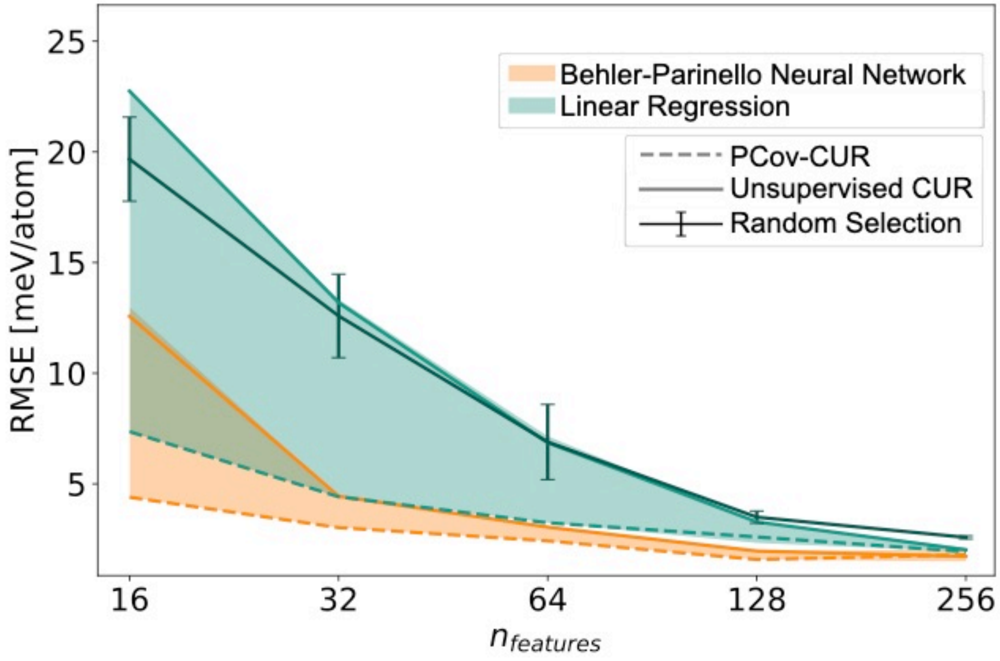


Inputs: SOAP vectors for small molecules containing C + H + N + O, (9 / 1) train / test split
Target: NMR chemical shieldings in ppm
Model used: 5-fold cross-validated linear ridge regression

RKC, et al 2021 Mach. Learn.: Sci. Technol. 2 035038
scikit-cosmo.readthedocs.io

Behler-Parinello Neural Networks

Introducing supervised aspects to feature selection invariably improves regression performance – even in non-linear models -- such as determining energies and forces using a neural network.

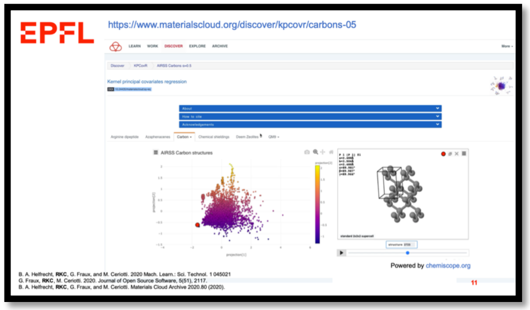


Inputs: symmetry functions of benzene rings from a simulation trajectory, (7/2/1) train / validation / test split

Target: energies in [meV / atom]

Models used: 5-fold cross-validated linear ridge regression, Behler-Parinello Neural Network

RKC, et al 2021 Mach. Learn.: Sci. Technol. 2 035038
scikit-cosmo.readthedocs.io



https://www.materialscloud.org/discover/kpcovr/carbons-05

Kernel principal covariates regression

Powered by chemiscope.org

B. A. Helfrecht, **RKC**, G. Fraux, and M. Ceriotti. 2020 Mach. Learn.: Sci. Technol. 1 045021
G. Fraux, **RKC**, M. Ceriotti. 2020. Journal of Open Source Software, 5(51), 2117.
B. A. Helfrecht, **RKC**, G. Fraux, and M. Ceriotti. Materials Cloud Archive 2020.80 (2020).

kernel-tutorials

A set of utilities and pedagogic notebooks for the use of linear and kernel methods in atomistic modeling

https://www.github.com/cosmo-epfl/kernel-tutorials/

librascal

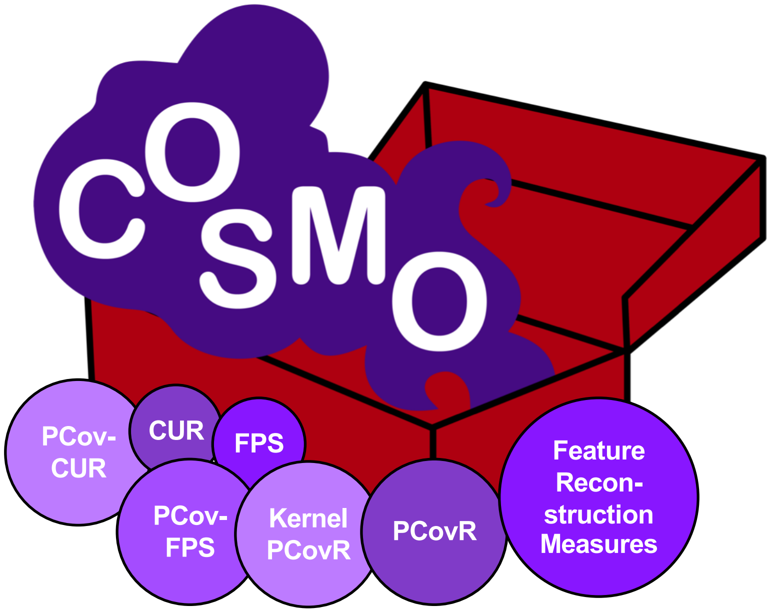
A scalable and versatile library to generate representations for atomic-scale learning

https://www.github.com/cosmo-epfl/librascal/

chemiscope

chemiscope is an interactive structure/property explorer for materials and molecules. The goal of chemiscope is to provide interactive exploration of large databases of materials and molecules and help researchers to find structure-properties correlations inside such databases.

chemiscope.org



scikit-COSMO

scikit-COSMO is a collection of scikit-learn compatible utilities that implement methods developed at COSMO.

scikit-cosmo.readthedocs.io
https://www.github.com/cosmo-epfl/scikit-cosmo/

Hybrid Unsupervised-Supervised Machine Learning Models for Materials Science

Rose K. Cersonsky

B. A. Helfrecht, **RKC**, G. Fraux, and M. Ceriotti "Structure-property maps with Kernel principal covariates regression."
2020 Mach. Learn.: Sci. Technol. 1045021.
https://iopscience.iop.org/article/10.1088/2632-2153/aba9ef

G. Fraux, **RKC**, M. Ceriotti. "Chemiscope"
2020 Journal of Open Source Software, 5(51), 2117.
https://doi.org/10.21105/joss.02117

Goscinski, A., et al. "The role of feature space in atomistic learning."
2021 Mach. Learn.: Sci. Technol. 025028.
https://doi.org/10.1088/2632-2153/abdaf7

F. Musil, et al. "Physics-Inspired Structural Representations for Molecules and Materials."
Chem. Rev. 2021.
https://doi.org/10.1021/acs.chemrev.1c00021

RKC, B. A Helfrecht, E. A. Engel, and M. Ceriotti . "Improving Sample and Feature Selection with Principal Covariates Regression"
2021 Mach. Learn.: Sci. Technol. 2 035038
https://doi.org/10.1088/2632-2153/abfe7c.

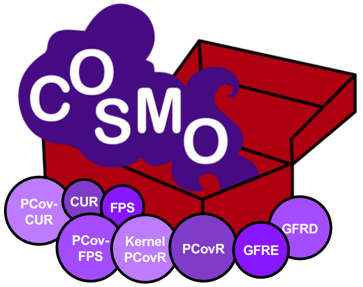
B. A. Helfrecht, **RKC**, G. Fraux, and M. Ceriotti.
Materials Cloud Archive 2020.80 (2020).
https://archive.materialscloud.org/record/2020.80

V. L. Deringer, et al.,
Origins of structural and electronic transitions in disordered silicon.
Nature 589, 59–64 (2021).
https://doi.org/10.1038/s41586-020-03072-z

S. de Jong, H.A.L. Kiers
Principal Covariates Regression: Part 1.
Chemom. intell. lab. syst. 14 (1992) 155-164.
https://doi.org/10.1016/0169-7439(92)80100-I

scikit-COSMO

scikit-COSMO is a collection of scikit-learn compatible utilities that implement methods developed at COSMO.



scikit-cosmo.readthedocs.io
https://www.github.com/cosmo-epfl/scikit-cosmo/