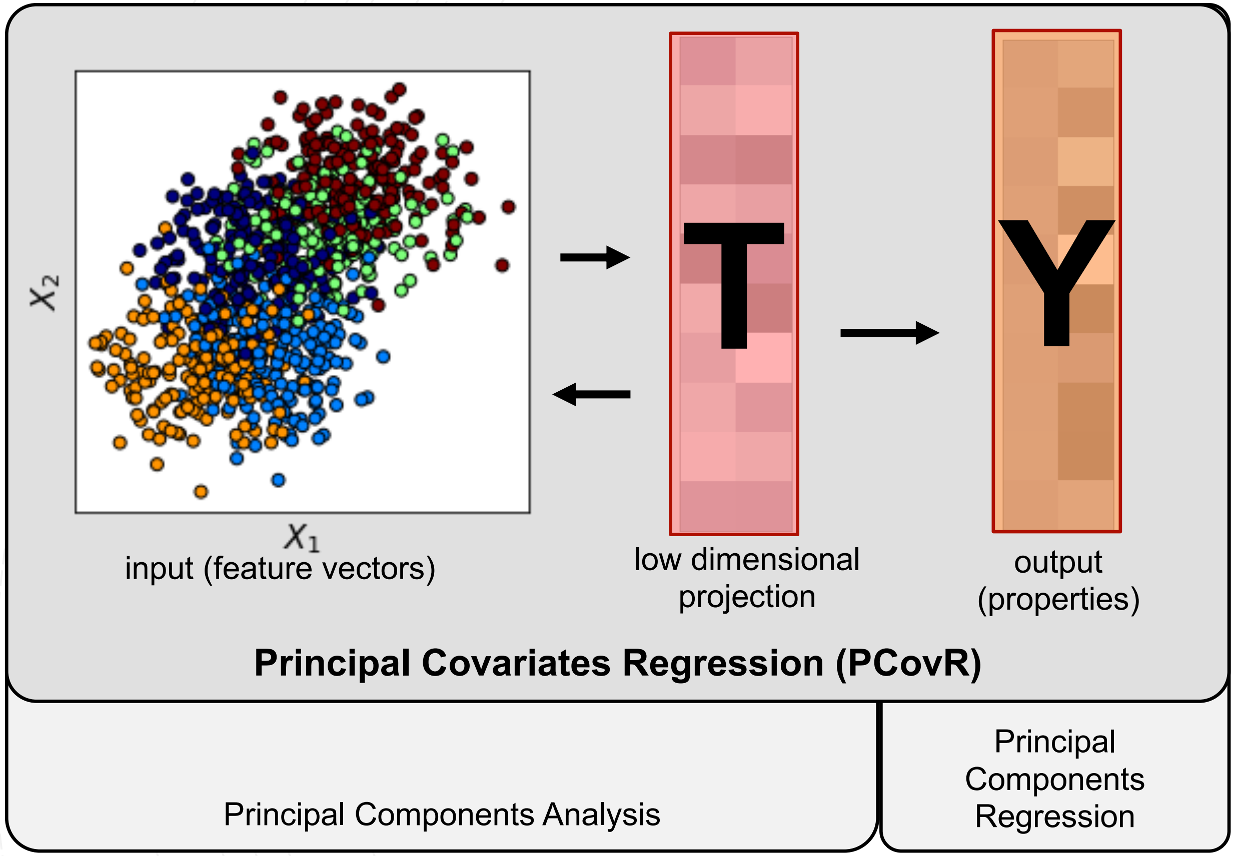
Improving Data Sub-Selection for Supervised Tasks with Principal Covariates Regression

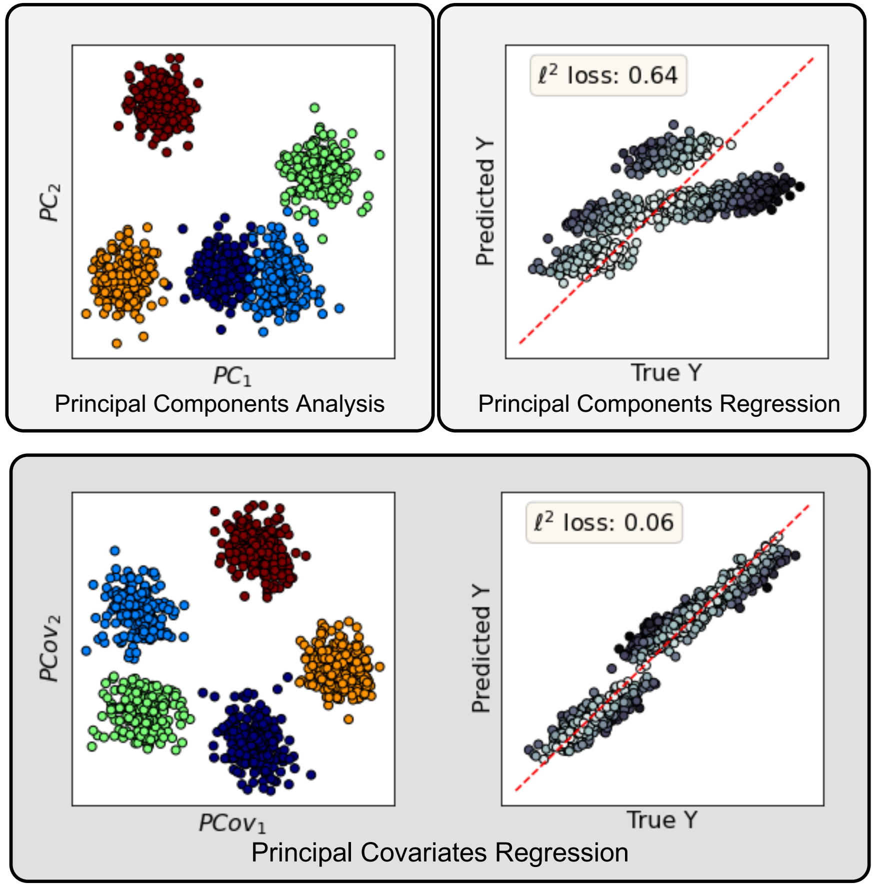
Rose K. Cersonsky, Benjamin A. Helfrecht,
Sergei Kliavinek, Edgar A. Engel, Michele Ceriotti

Laboratory of Computational Science and Modeling (COSMO), École Polytechnique Fédérale de Lausanne (EPFL)

Principal Covariates Regression (PCovR)

is a dimensionality reduction technique that determines a latent-space projection that incorporate saspects of supervised learning.



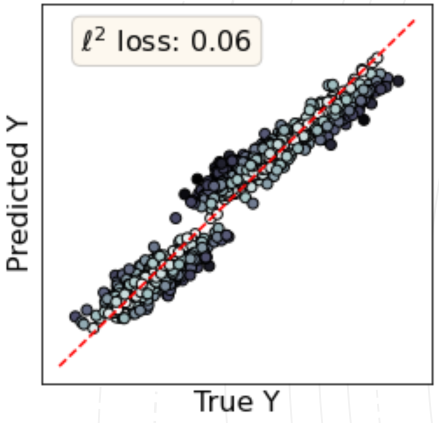


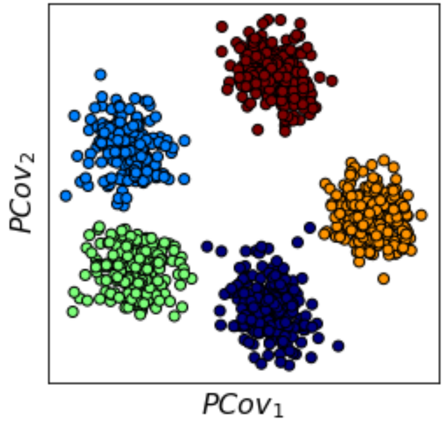
S. de Jong, H.A.L. Kiers, Chemom. intell. lab. syst. 14 (1992) 155-164.
scikit-cosmo.readthedocs.io

Inputs: sklearn.datasets.make_blobs
Regression Model: RidgeCV(cv=5)

Principal Covariates Regression (PCovR)

is controlled by a mixing parameter α that weights the regression and decomposition tasks.

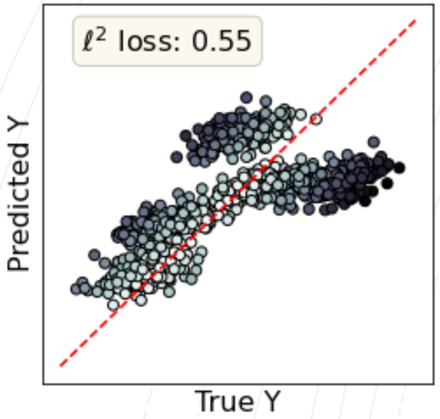


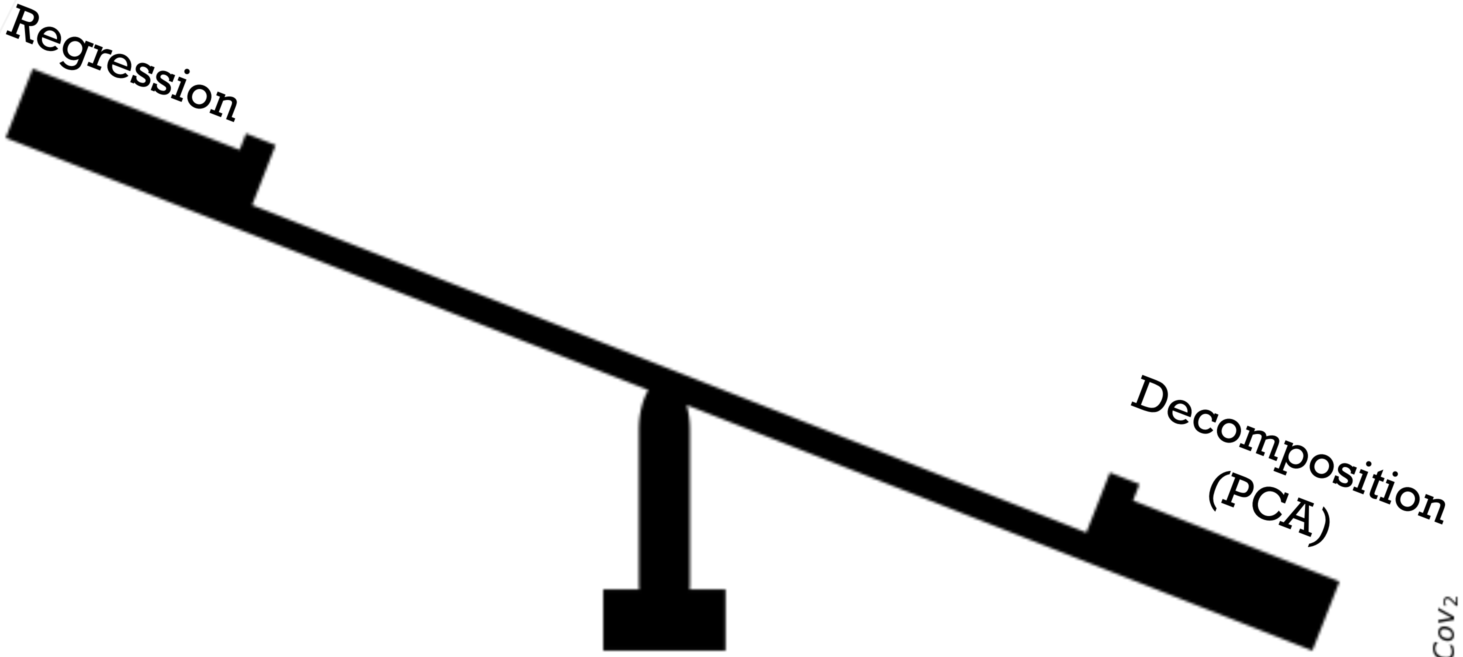


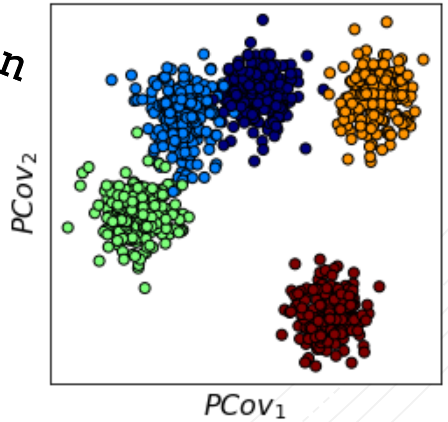
$$\alpha = 0.5$$

Principal Covariates Regression (PCovR)

is controlled by a mixing parameter α that weights the regression and decomposition tasks.



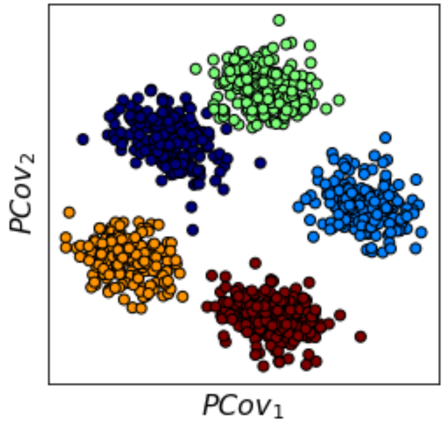


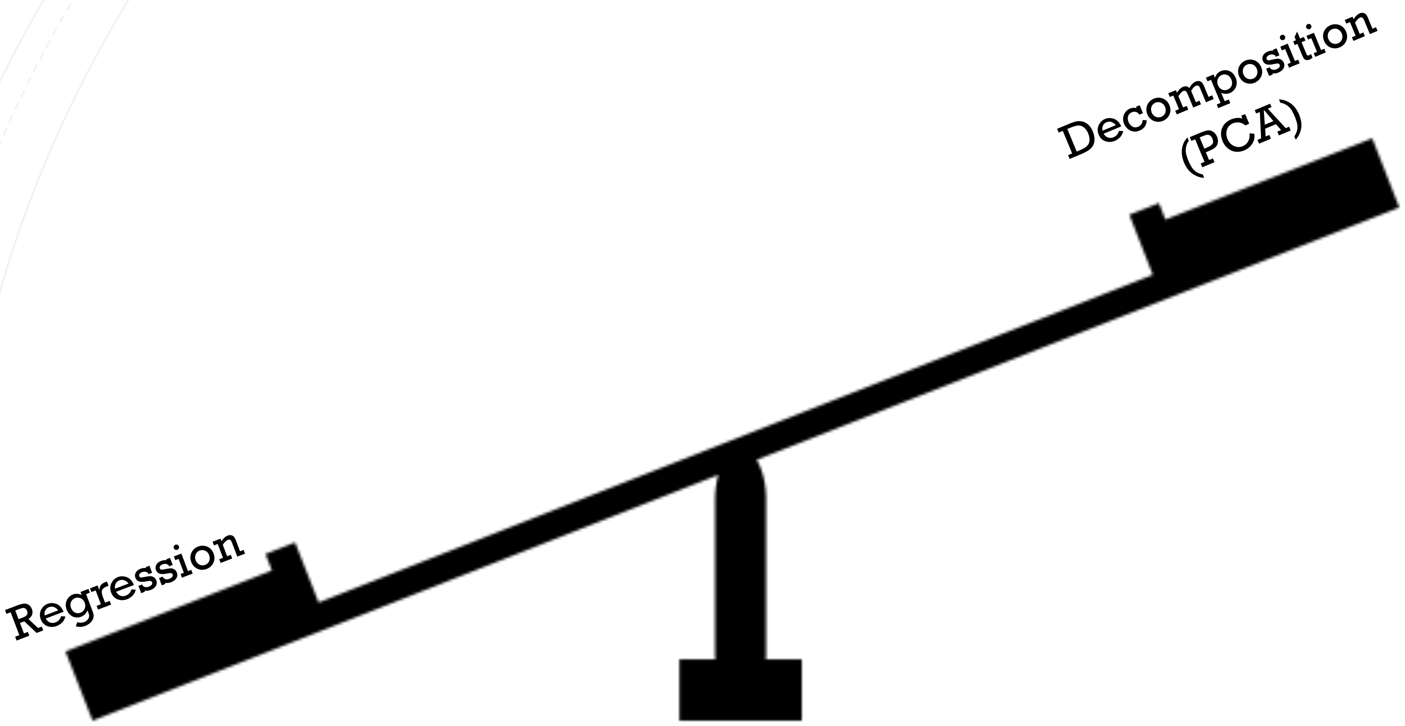


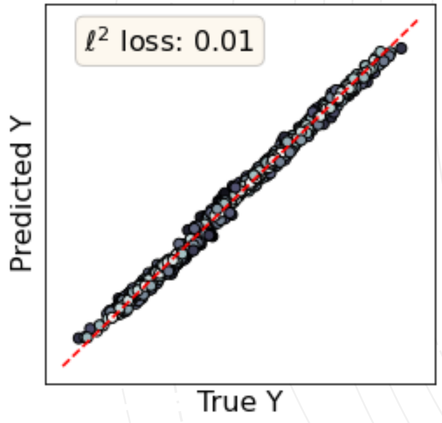
$\alpha = 0.9$

Principal Covariates Regression (PCovR)

is controlled by a mixing parameter α that weights the regression and decomposition tasks.



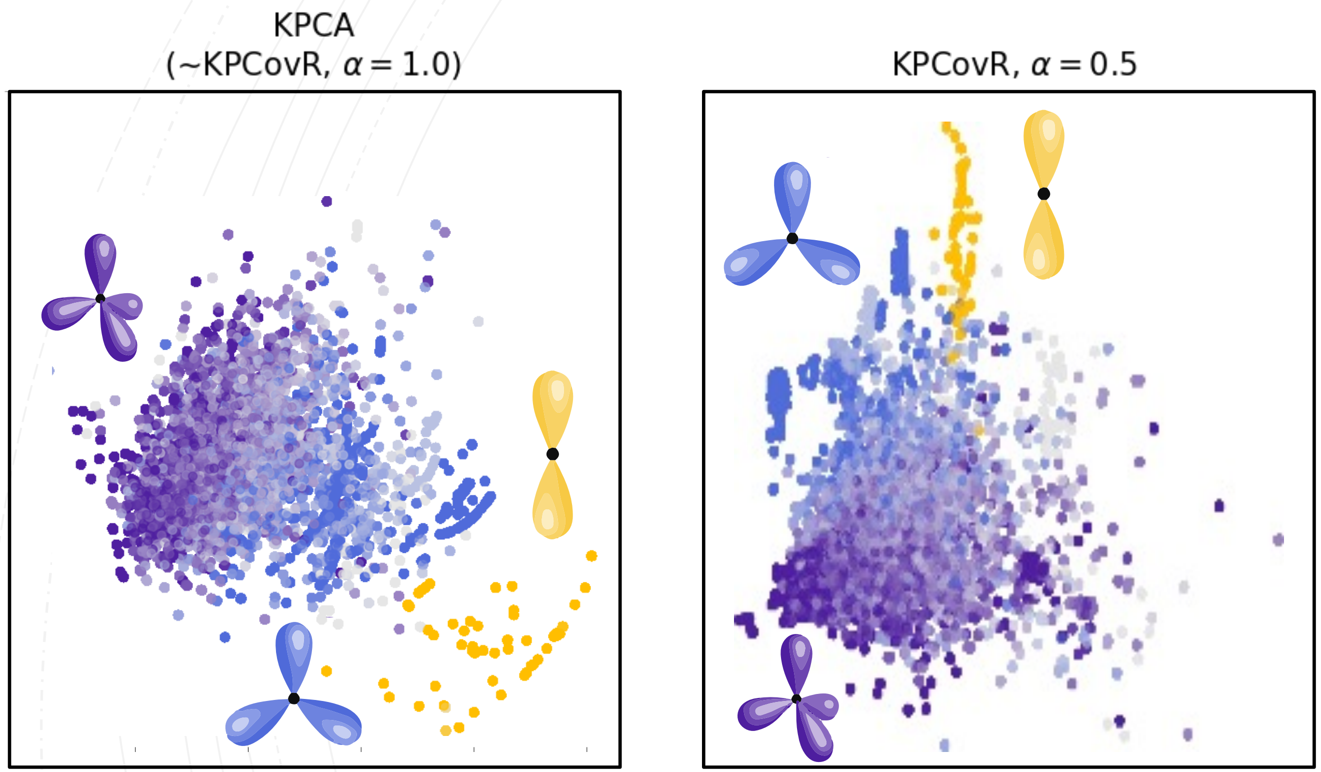




$\alpha = 0.1$

Kernel Principal Covariates Regression

Determines a low-dimension projection from a similarity kernel, considering target data when constructing the projection.

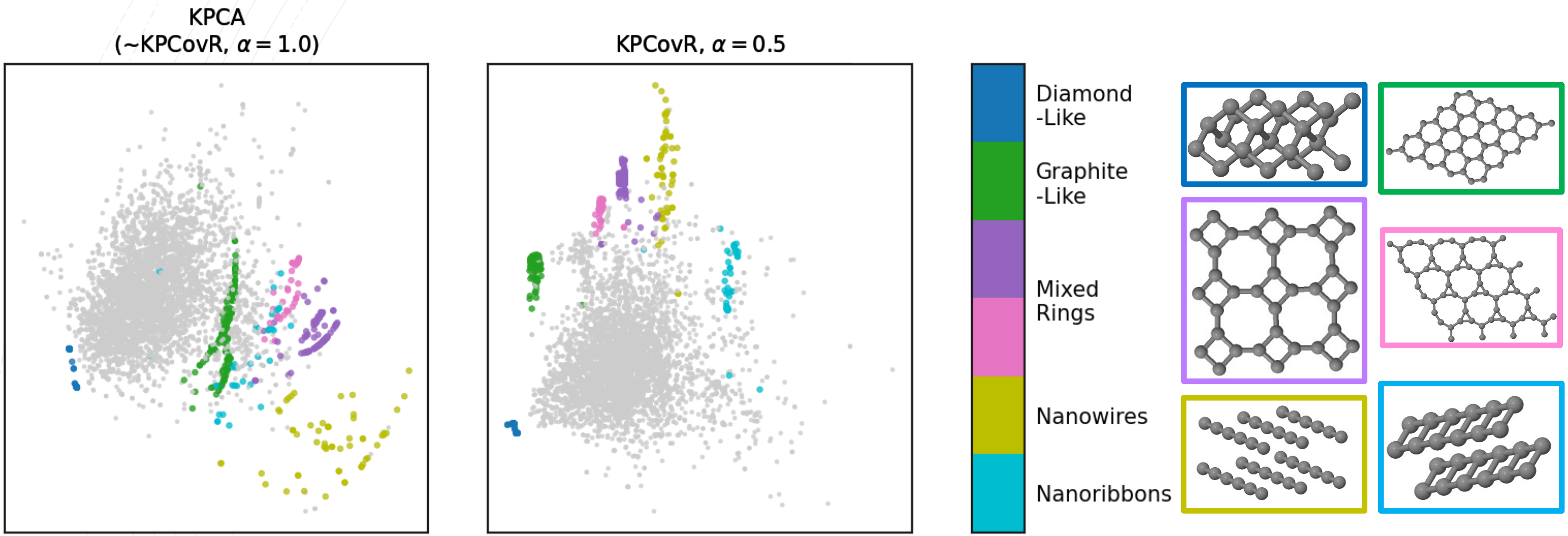


Inputs: SOAP features of 10,000 AIRSS carbon crystals
Target: energies in [eV / atom]
Kernel Parameters: RBF kernel, γ=$10^{-3.8}$
(1/1) train / test split

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).
scikit-cosmo.readthedocs.io

Kernel Principal Covariates Regression

Determines a low-dimension projection from a similarity kernel, considering target data when constructing the projection.

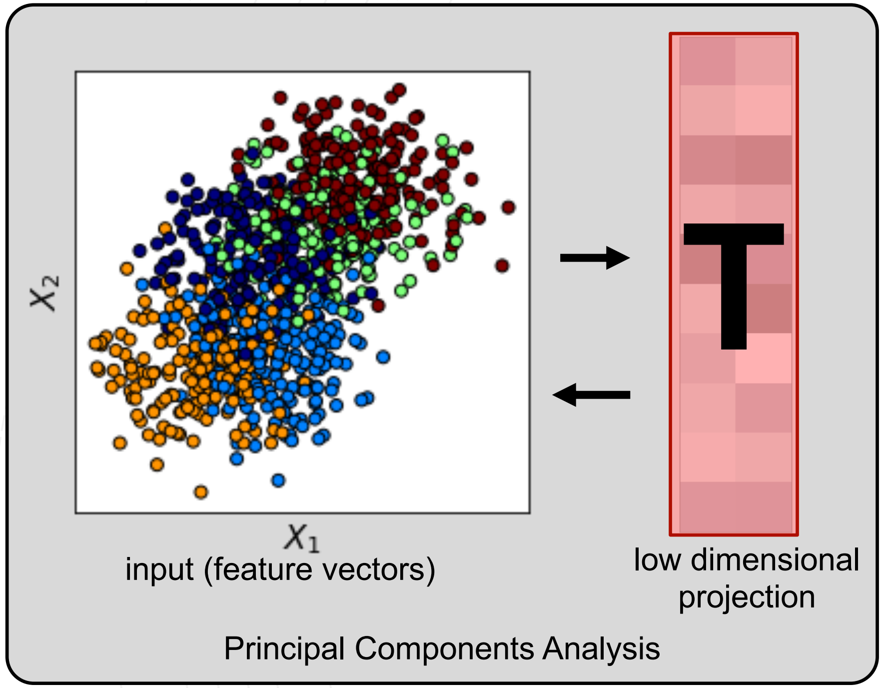


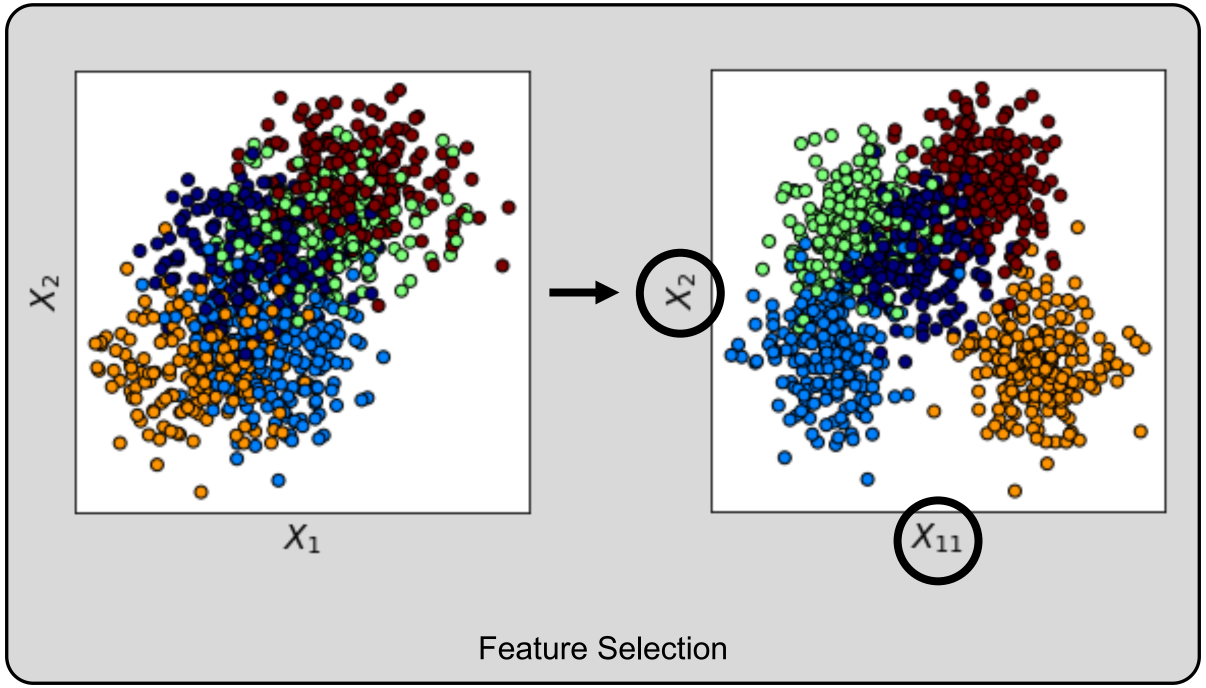
Inputs: SOAP features of 10,000 AIRSS carbon crystals
Target: energies in [eV / atom]
Kernel Parameters: RBF kernel, $\gamma=10^{-3.8}$
(1/1) train / test split

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).
scikit-cosmo.readthedocs.io

What if the features carry inherent meaning?

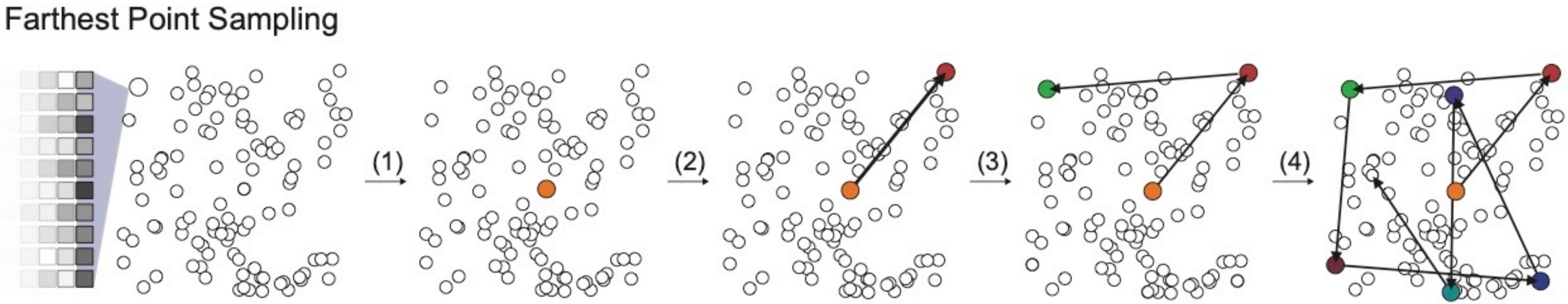
Many dimensionality reduction techniques construct a *new* set of features, but what if you want to just work with a subset of the old set?





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
4. Repeat 1-3 until you have enough features!

CUR Decomposition

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

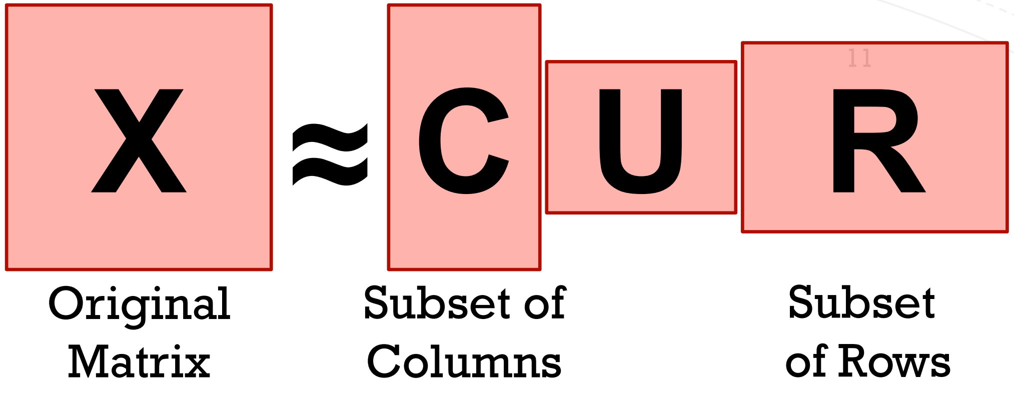
Head Canon:
"Columns und Rows" decomposition

$$X \approx C\,U\,R$$

Original Matrix

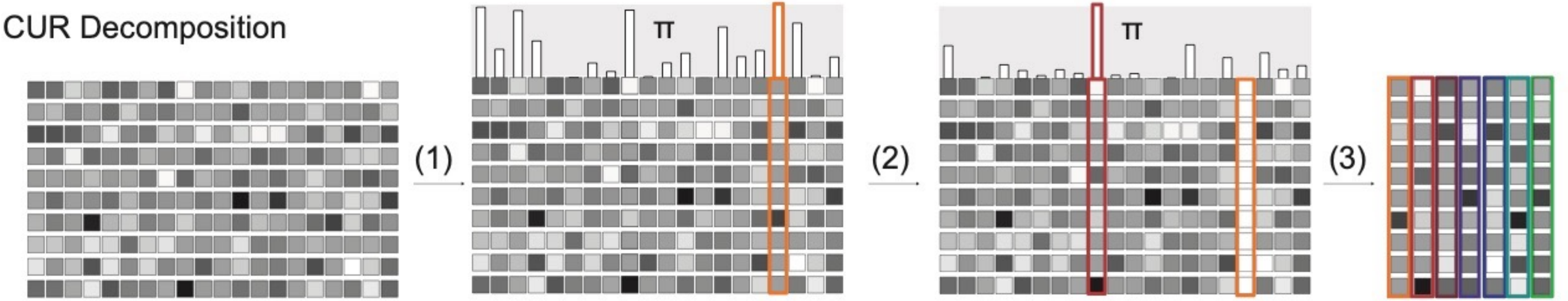
Subset of Columns

Subset of Rows



CUR Decomposition

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

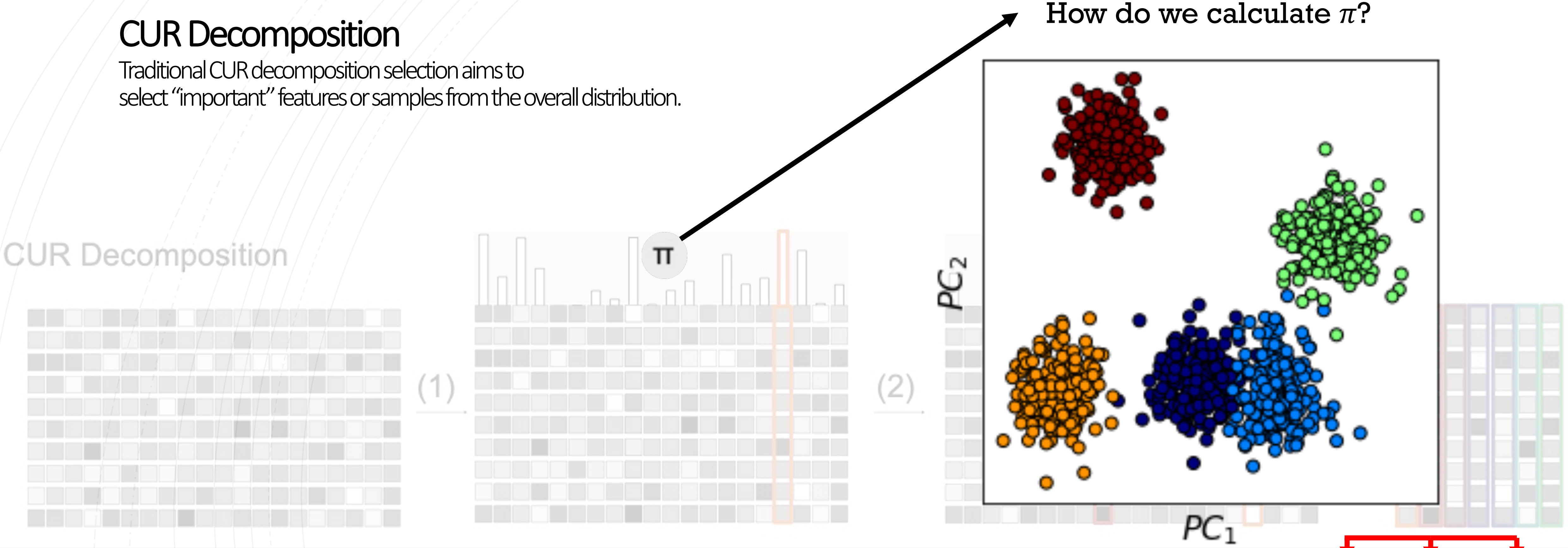


1. Compute importance score π
2. Choose column with highest π
3. Orthogonalize with respect to last chosen column.
4. Repeat 1-3 until you have enough features!

CUR Decomposition

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

How do we calculate π?



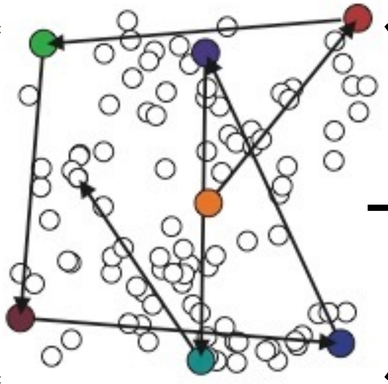
$$PC_1 = AX_1 + BX_2 + CX_3 \ldots$$

PCov-FPS and Pcov-CUR

Both FPS and CUR can be translated to PCovR space for both feature (and sample) selection.

Farthest Point Sampling (FPS)

$$d = f\Big($$



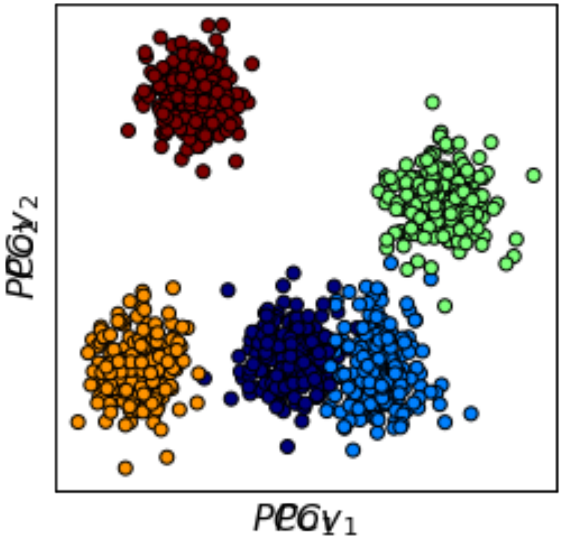
$$+$$

Contribution
to Regression
Weights

$$\Big)$$

CUR Decomposition

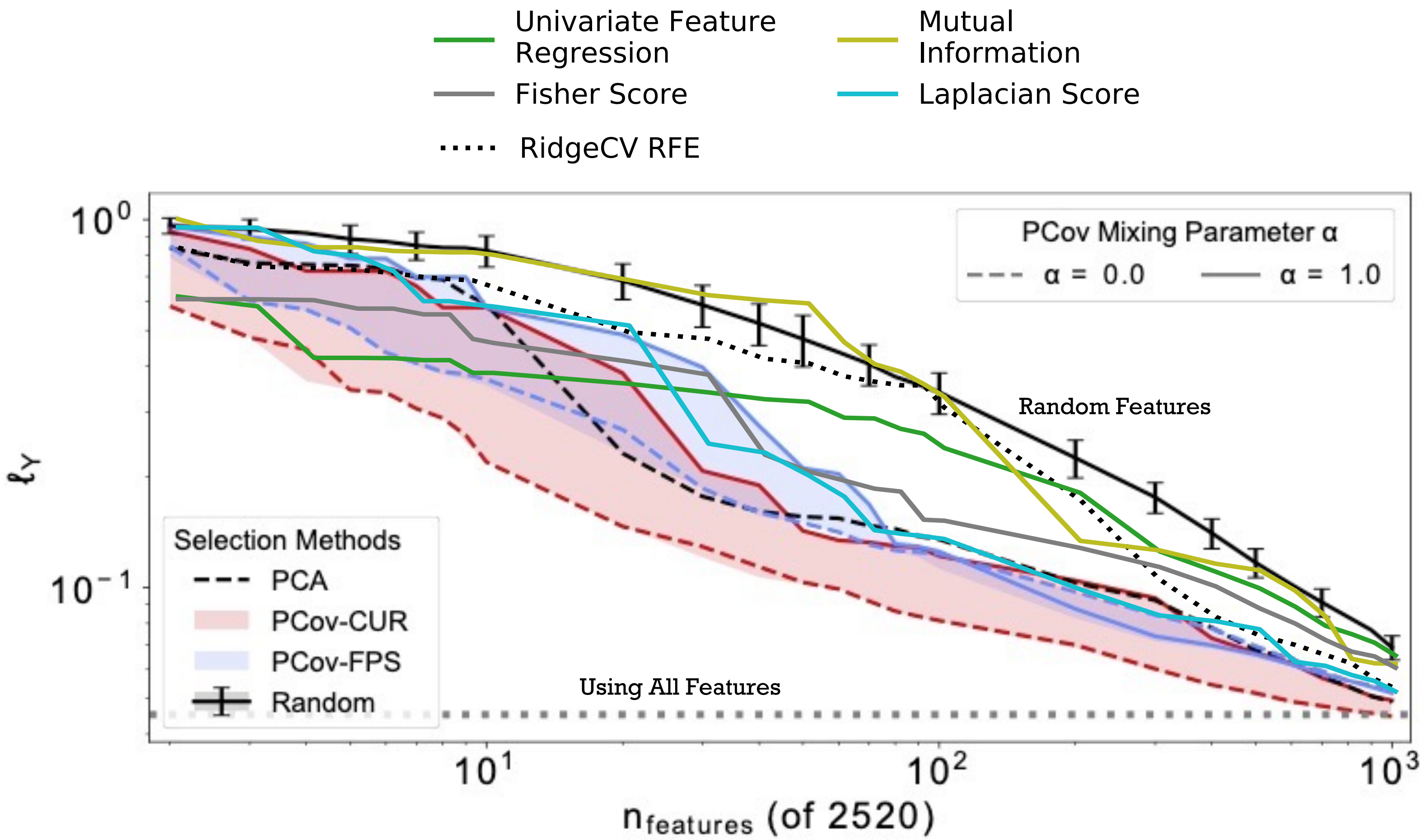
$$\pi = f\Big($$



$$\Big)$$

Linear Regression

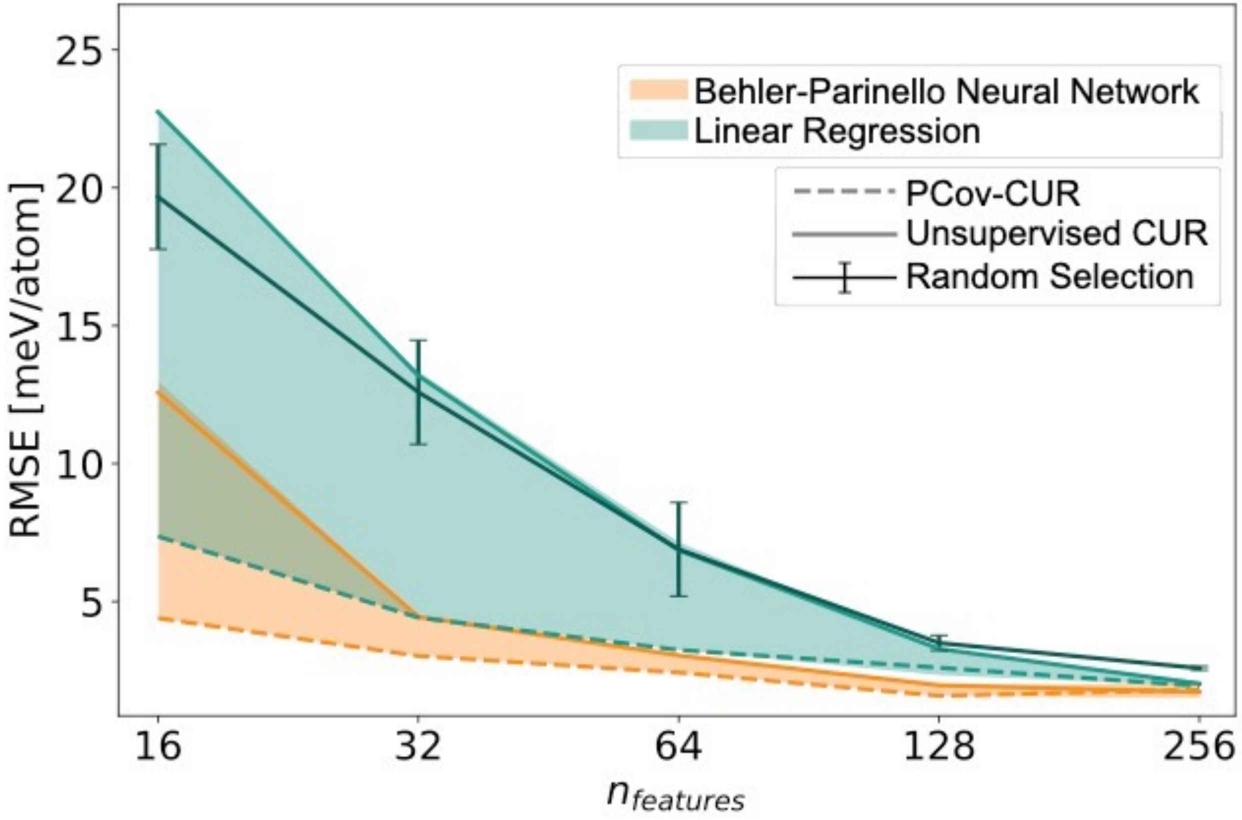
Using PCov-style feature selection will universally out-perform common feature selection metrics available via popular packages.



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



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

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

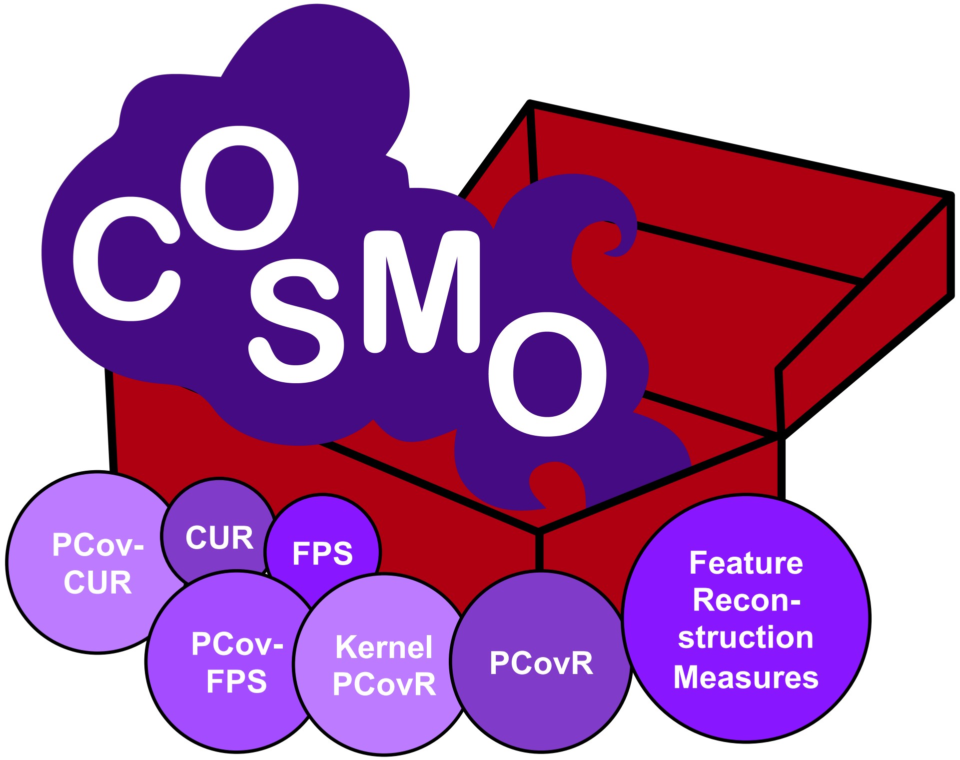
A scalable and versatile library to generate representations for atomic-scale learning

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/

Improving Data Sub-Selection for Supervised Tasks with Principal Covariates Regression

Rose K. Cersonsky, Benjamin A. Helfrecht, Sergei Kliavinek, Edgar A. Engel, Michele Ceriotti

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

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.

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

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/