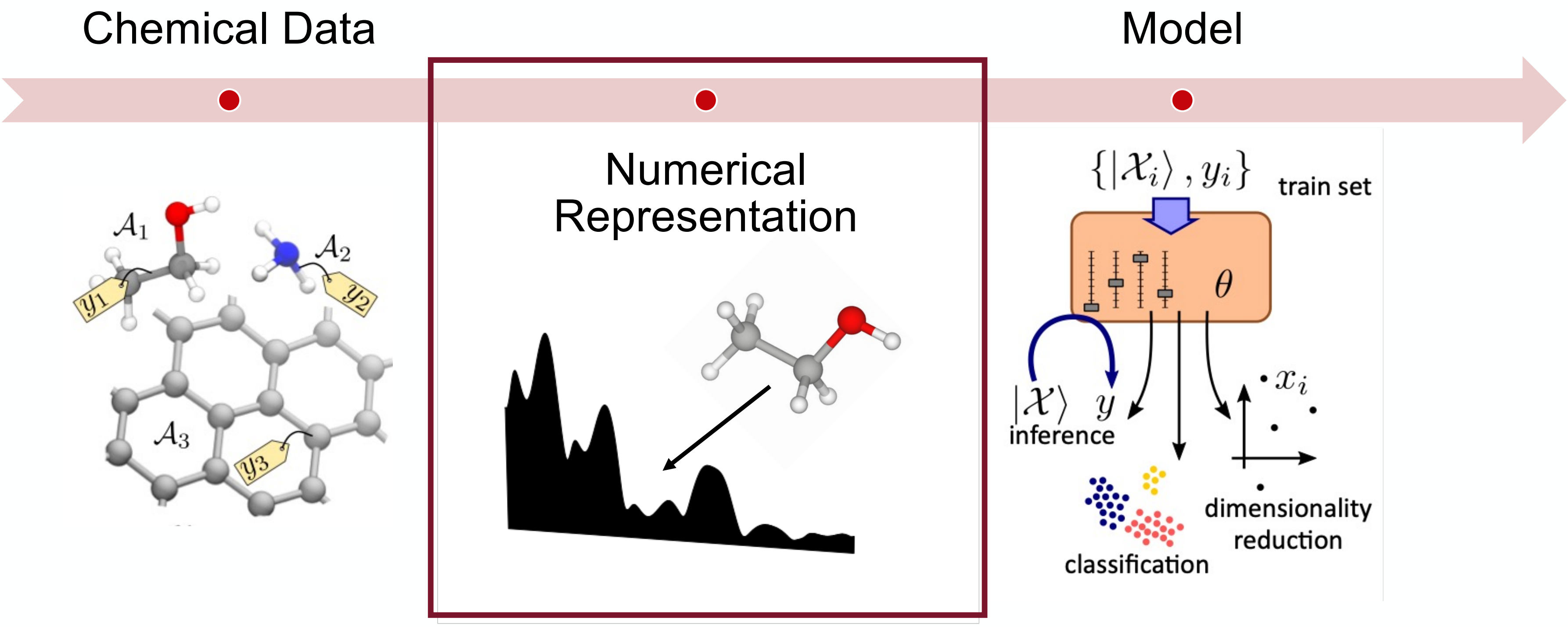
Data-driven approaches to chemical and materials science:

the impact of data selection, representation, and interpretability

Rose K. Cersonsky[1,2]

[1]Chemical and Biological Engineering, University of Wisconsin – Madison, [2] Materials Science and Engineering, University of Wisconsin – Madison

When we employ machine-learning workflows for chemical data, there are several critical steps which underlie any study.



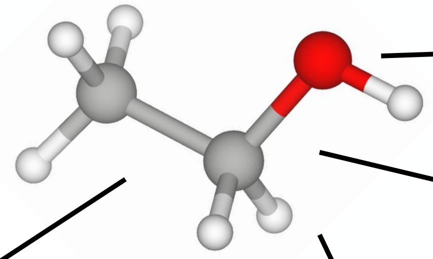
Chemical Data

Numerical Representation

Model

All of this is guided by **the scientific question at hand** and the **diversity and quality of the data**.

All machine learning models require that we translate our system into "features" that can be used to learn off of.

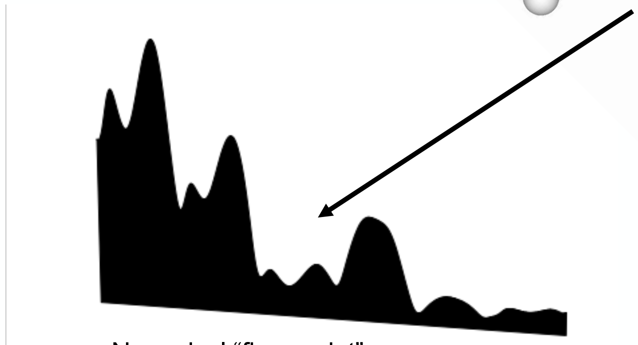


Chemical Formula

H_6C_2O

SMILES string

CCO



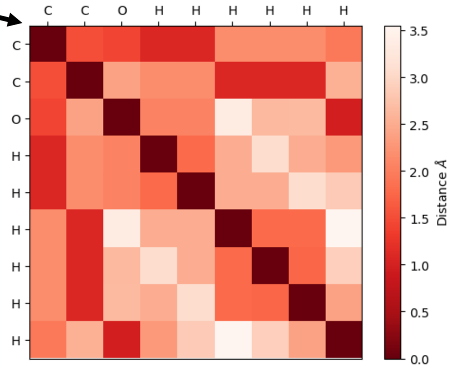
Numerical "fingerprint"

The quality of these "features" will determine how easy a target is to learn…

and how much you "learn" in the process!

	x	y	z
C	-0.47	0.514	0.007
C	0.887	-0.157	-0.005
O	-1.444	-0.404	-0.468
H	-0.746	0.833	1.016
H	-0.474	1.389	-0.649
H	1.664	0.531	0.339
H	1.14	-0.499	-1.014
H	0.889	-1.041	0.641
H	-1.447	-1.167	0.135

Internal Coordinates

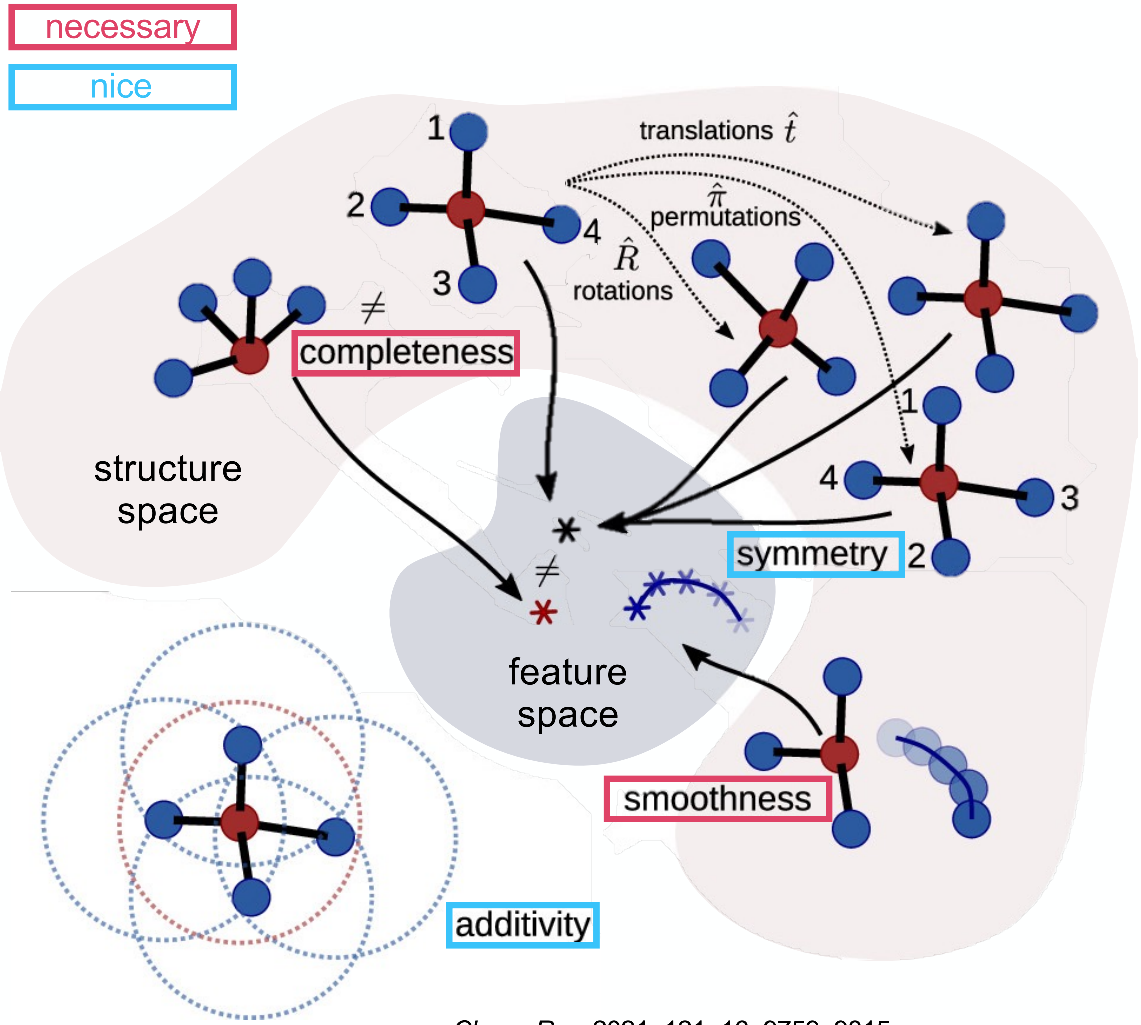


"Adjacency matrix"

In thermodynamic contexts*, what do we want from a representation?

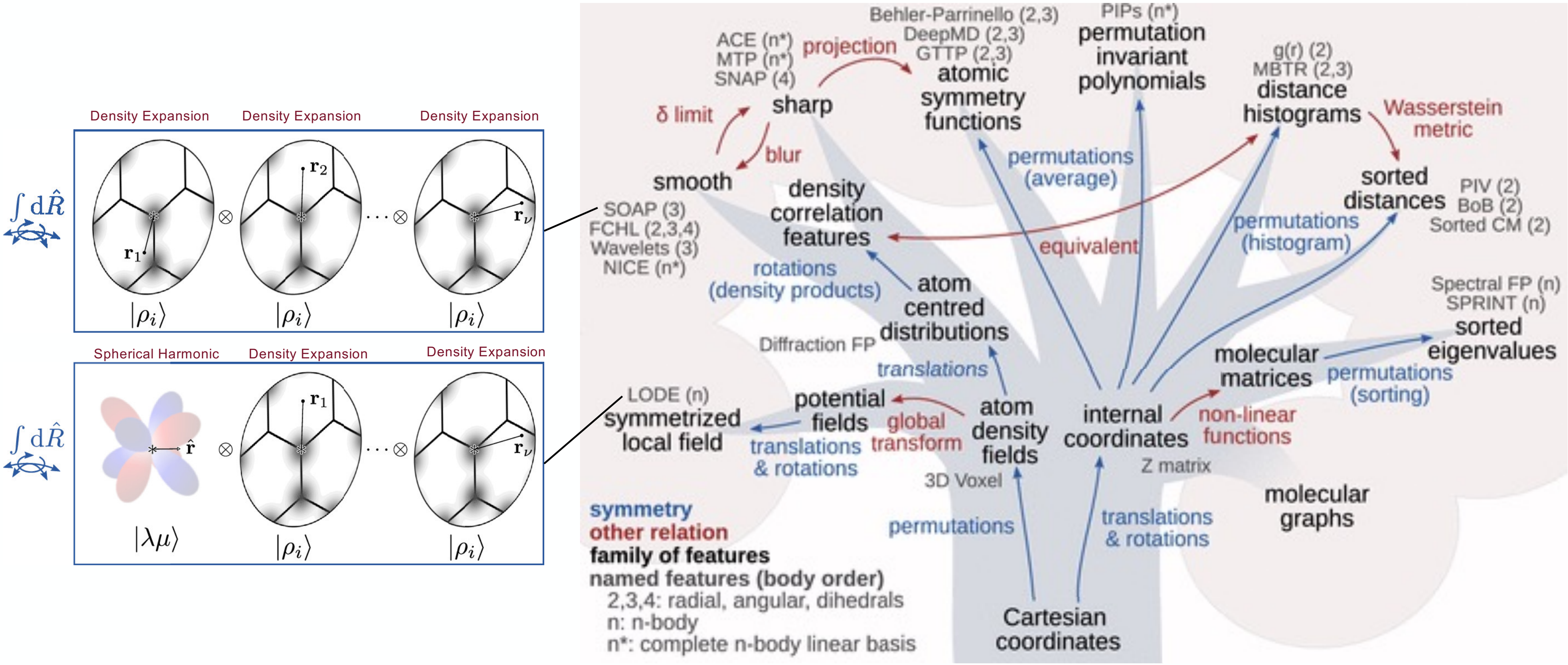
*thermodynamic contexts:

- interparticle potentials
- identifying reaction coordinates, collective variables, or axes for probability distributions
- delineating and characterizing phases
- assessing gradients, dynamical patterns, and transport processes



Chem. Rev. 2021, 121, 16, 9759–9815

Many symmetry-adapted frameworks can be expressed in terms of n-body correlations of atom positions. Only difference ***- the choice of basis.***

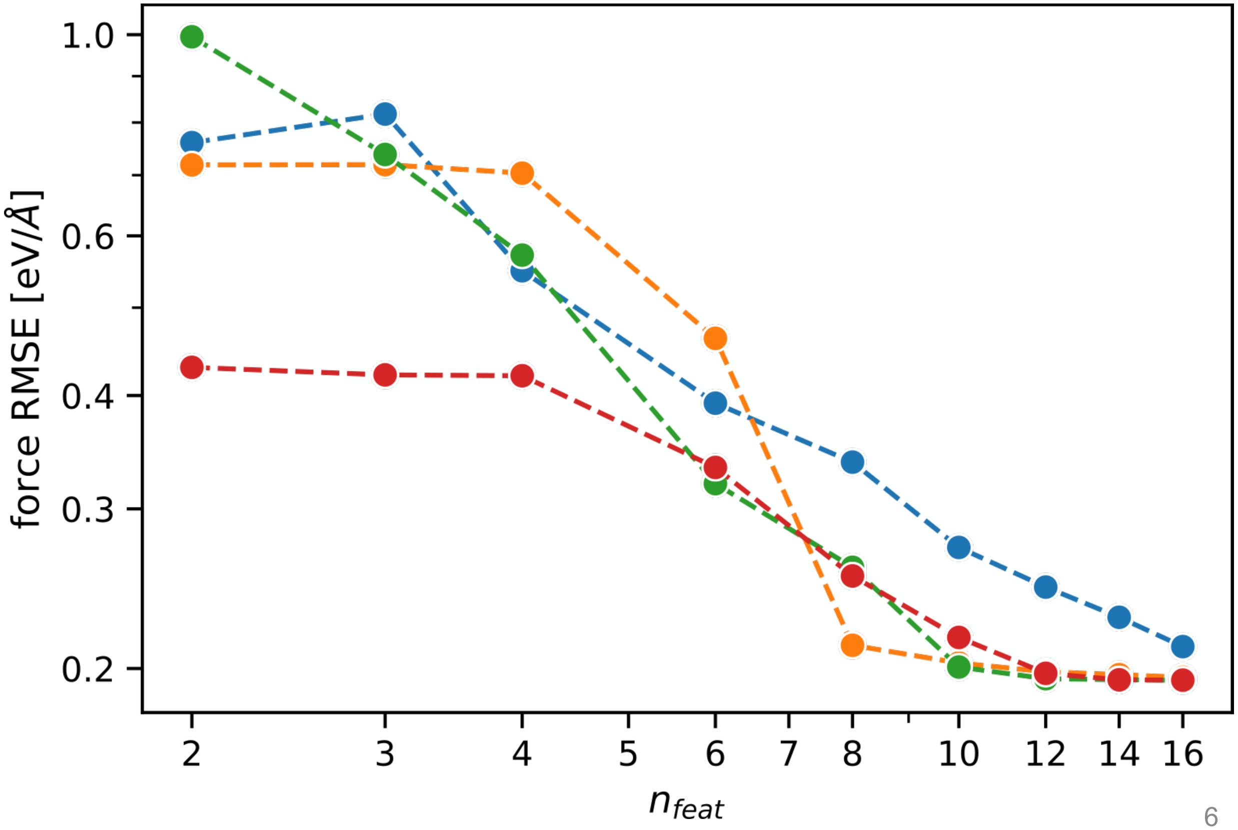


How do we know which featurization to use?

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?



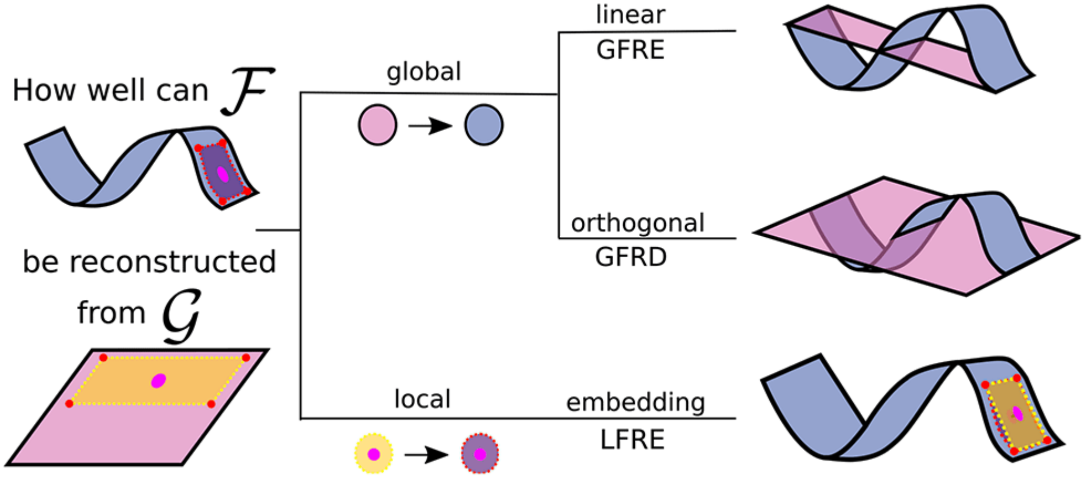
The feature reconstruction error (FRE) denotes the mutual information contained in those two feature sets.

$\mathrm{FRE}(\mathbf{X_1}, \mathbf{X_2}) = 0.0$

$\mathbf{X_1}$ can recreate 100% of the information in $\mathbf{X_2}$

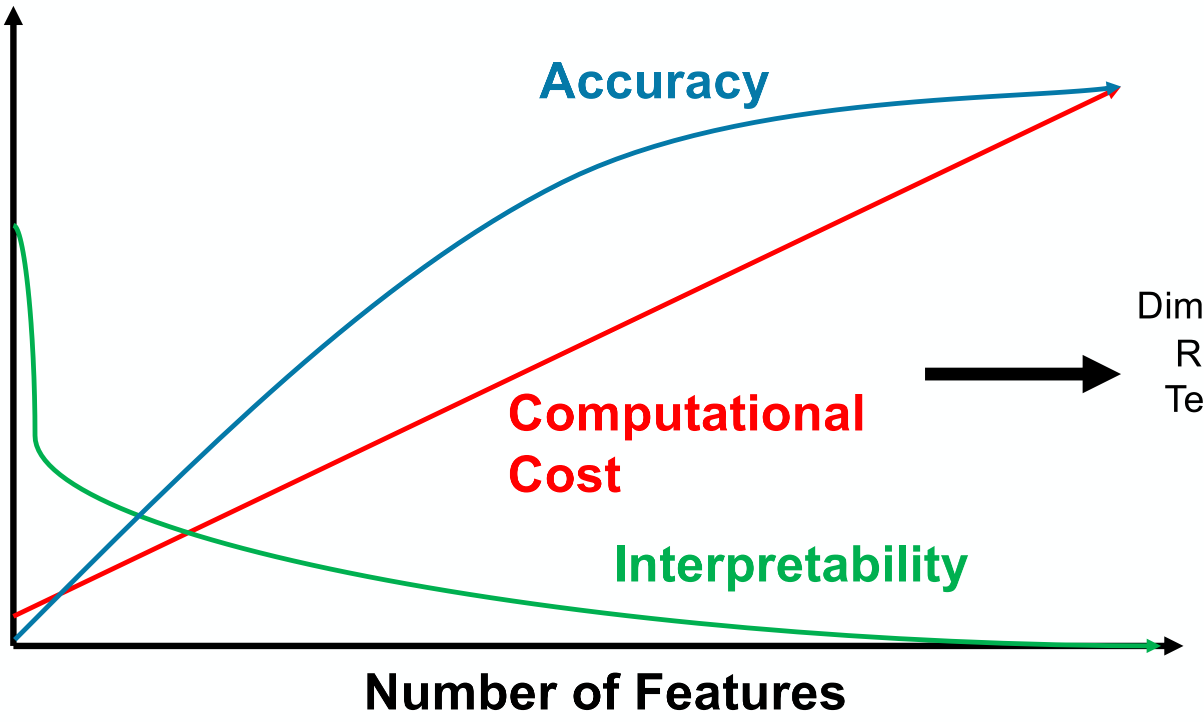
$\mathrm{FRE}(\mathbf{X_1}, \mathbf{X_2}) = 0.5$

$\mathbf{X_1}$ can only recreate 50% of the information in $\mathbf{X_2}$



A. Goscinski et al 2021 Mach. Learn.: Sci. Technol. 2 025028
A. Goscinski, …, **RKC**, 2023 Open Research Europe, 3(81).

Why not just use the most extensive set of features?



Dimensionality Reduction Techniques

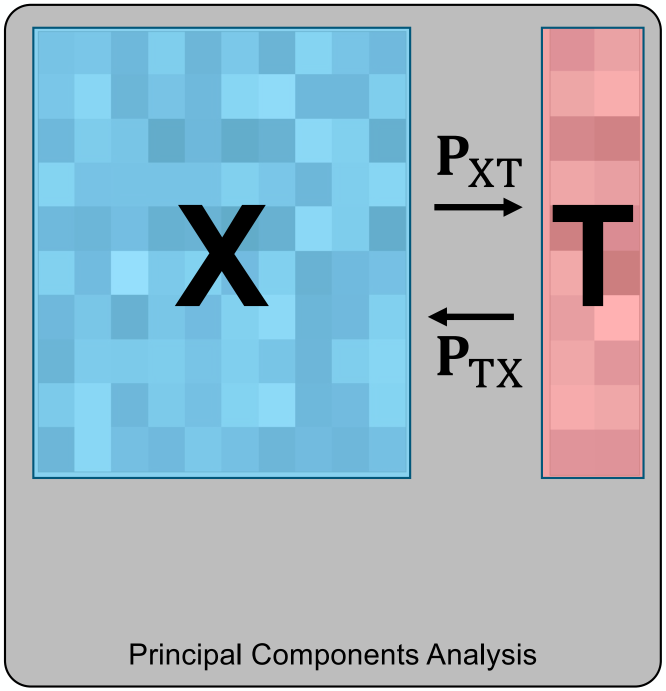
A couple words on notation…

$\mathbf{X} = \begin{bmatrix}\mathbf{x}_1 \\ \mathbf{x}_2 \\ \dots\end{bmatrix}$	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.



X
$\mathbf{P}_{\mathrm{XT}}$
$\mathbf{P}_{\mathrm{TX}}$
T
Principal Components Analysis

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

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)

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

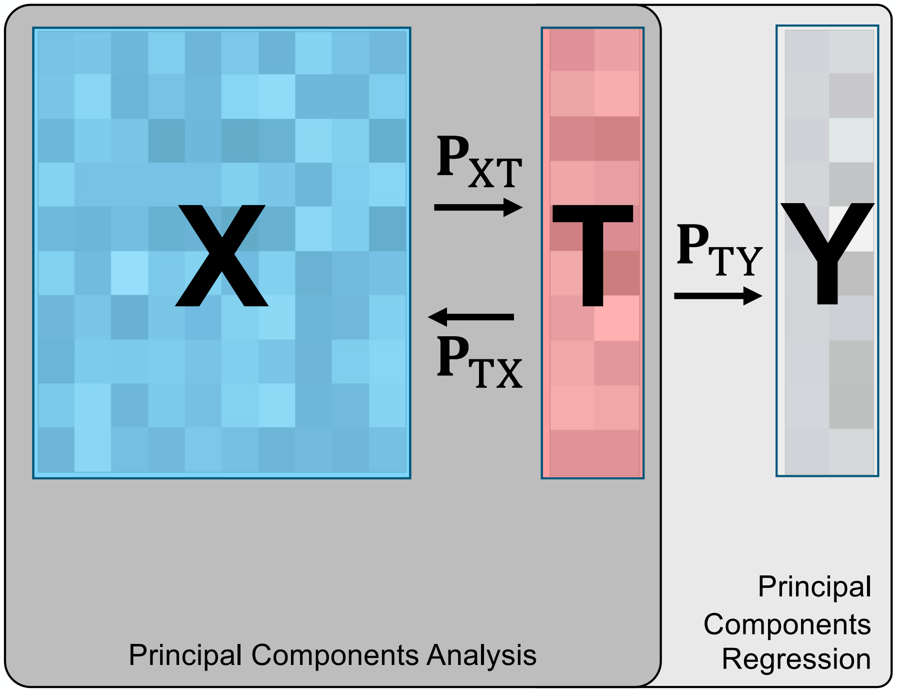
gram matrix

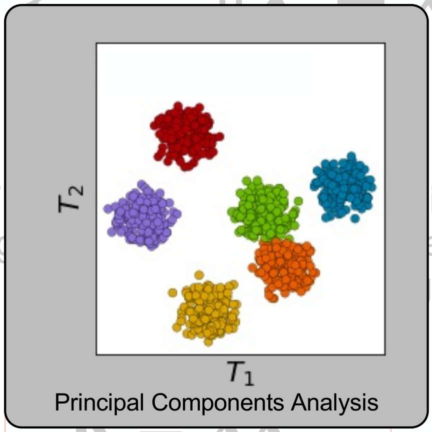
$$\boxed{\mathbf{C} = \mathbf{X}^{\mathrm{T}}\mathbf{X}}$$

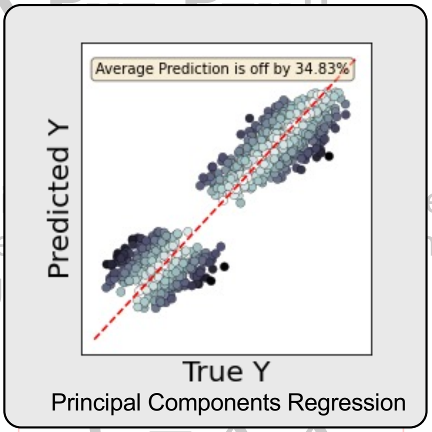
covariance matrix

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.

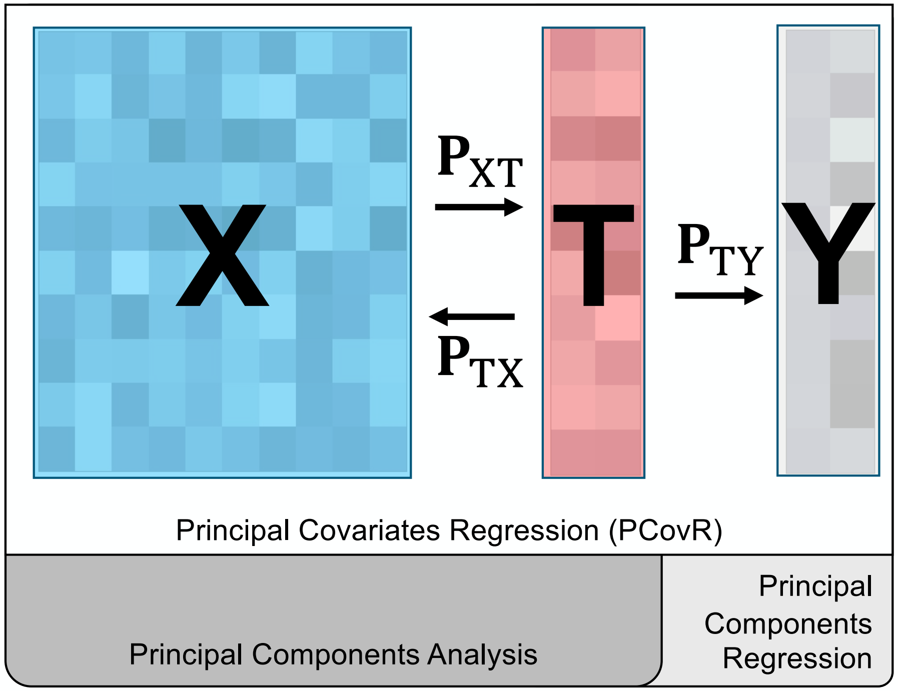

X
P_{XT}
P_{TX}
T
P_{TY}
Y
Principal Components Analysis
Principal Components Regression





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
P_XT
P_TX
T
P_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

$$\mathbf{K} \rightarrow \tilde{\mathbf{K}}$$

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

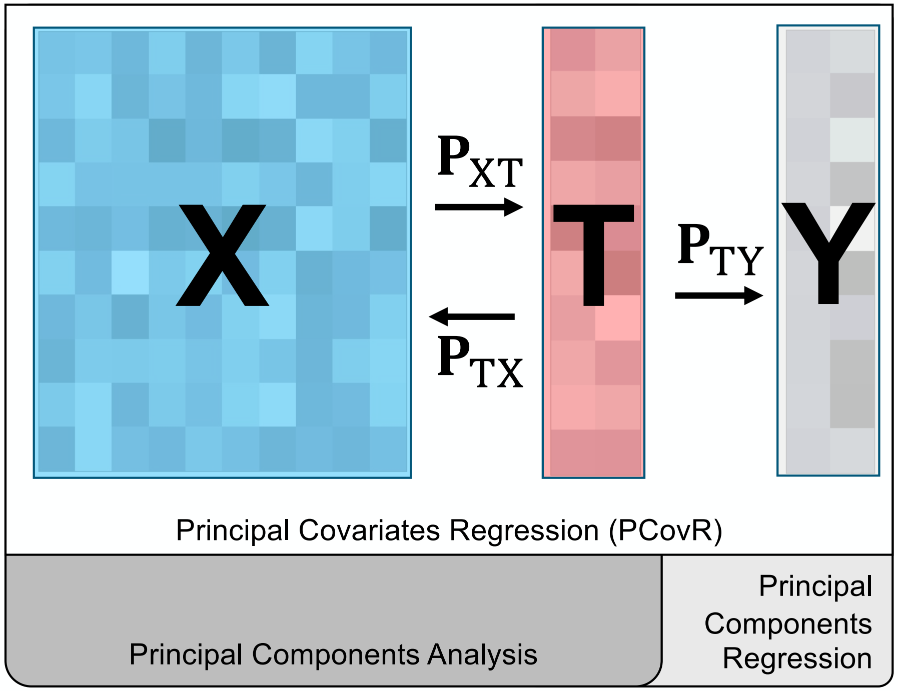
$$\mathbf{C} \rightarrow \tilde{\mathbf{C}}$$

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

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

Principal Covariates Regression (PCovR)

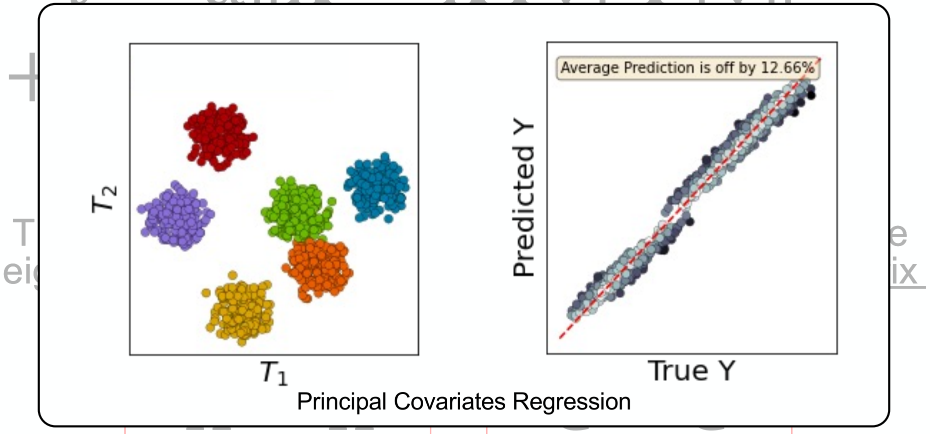
PCovR determines an information rich set of features to represent a larger set of features *and* optimally regress a set of targets.


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

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

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

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

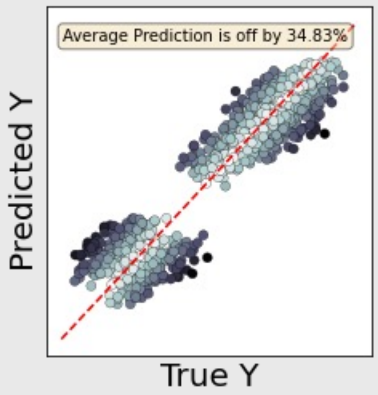
Core to PCA / PCovR is the gram kernel, which is equivalent to the linear kernel.
We can replace this with any number of non-linear kernels to better represent non-linear structure-property relations.

gram matrix,
a.k.a. “linear kernel”

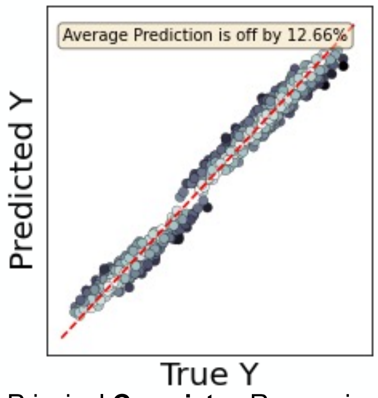
$$\mathbf{K} = \mathbf{X}\mathbf{X}^{\mathrm{T}}$$

$$\mathrm{K}_{\mathrm{ij}} = \mathrm{k}(\mathbf{x}_{\mathrm{i}}, \mathbf{x}_{\mathrm{j}}) = \mathbf{x}_{\mathrm{i}} \cdot \mathbf{x}_{\mathrm{j}} \quad \mathbf{e}^{-\gamma\|\mathbf{x}_{\mathrm{i}}-\mathbf{x}_{\mathrm{j}}\|^{2}}$$

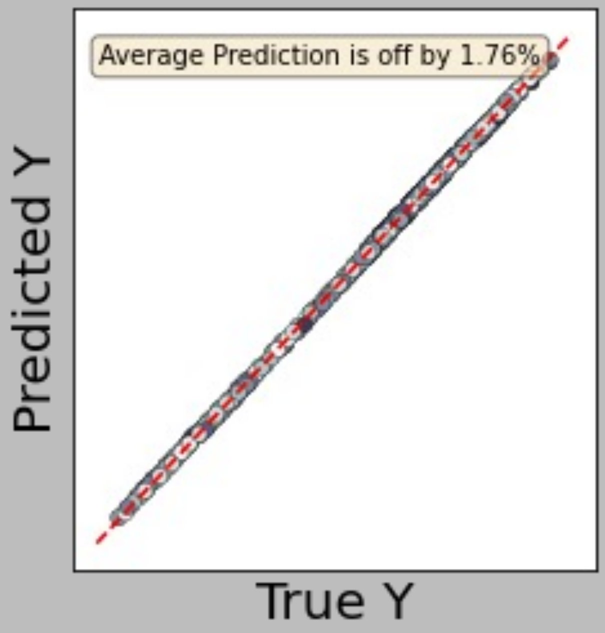
non-linear kernel



Principal **Components** Regression



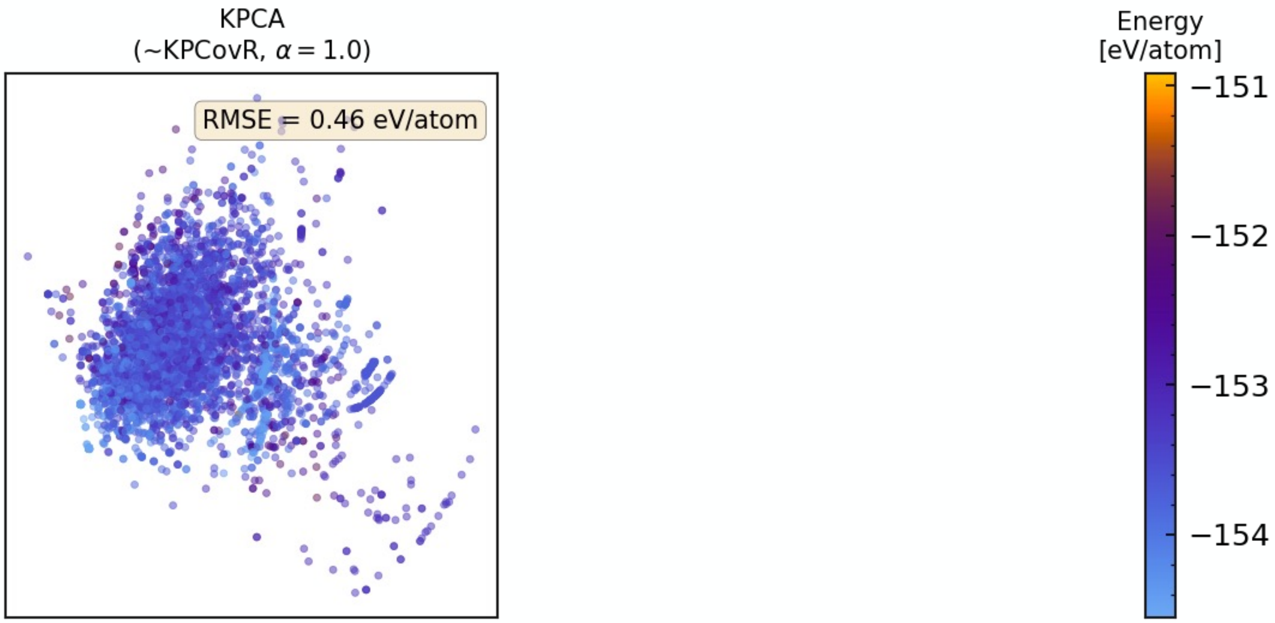
Principal **Covariates** Regression

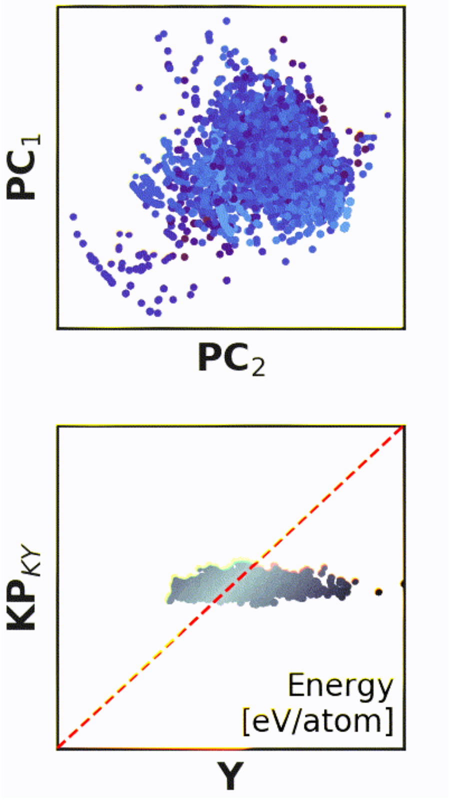


Kernel Principal Covariates Regression

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

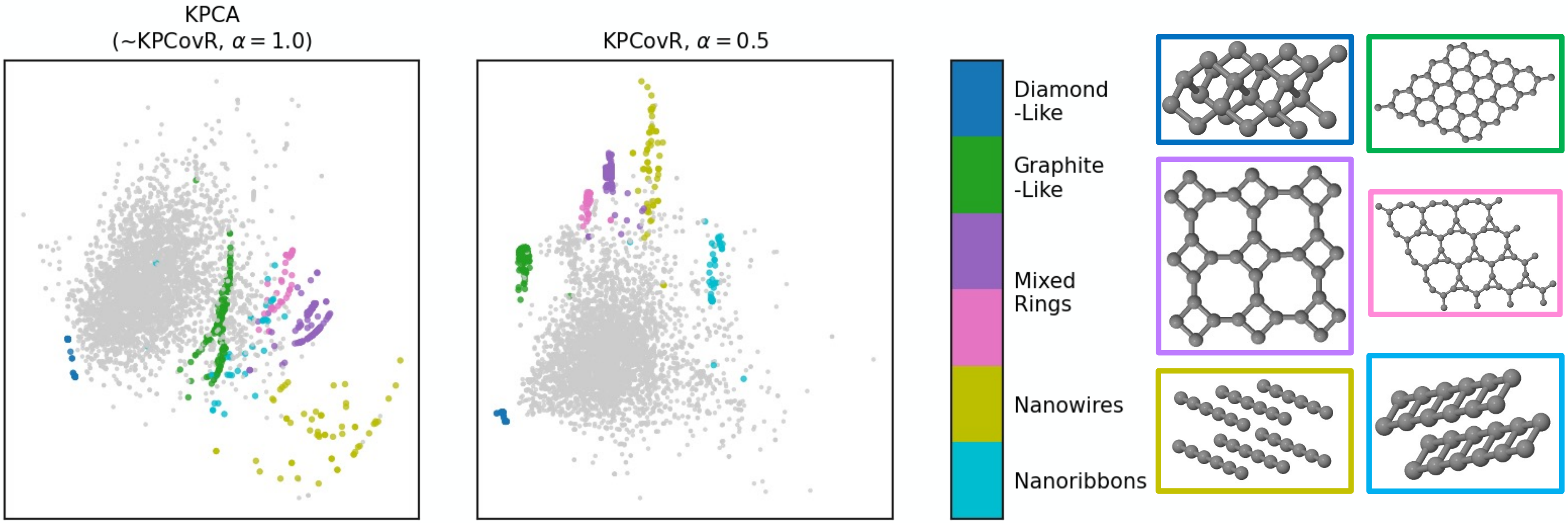
Analysis of SOAP features of Ab-Initio Random Structure Search (AIRSS) carbon crystals and their energies in eV/atom



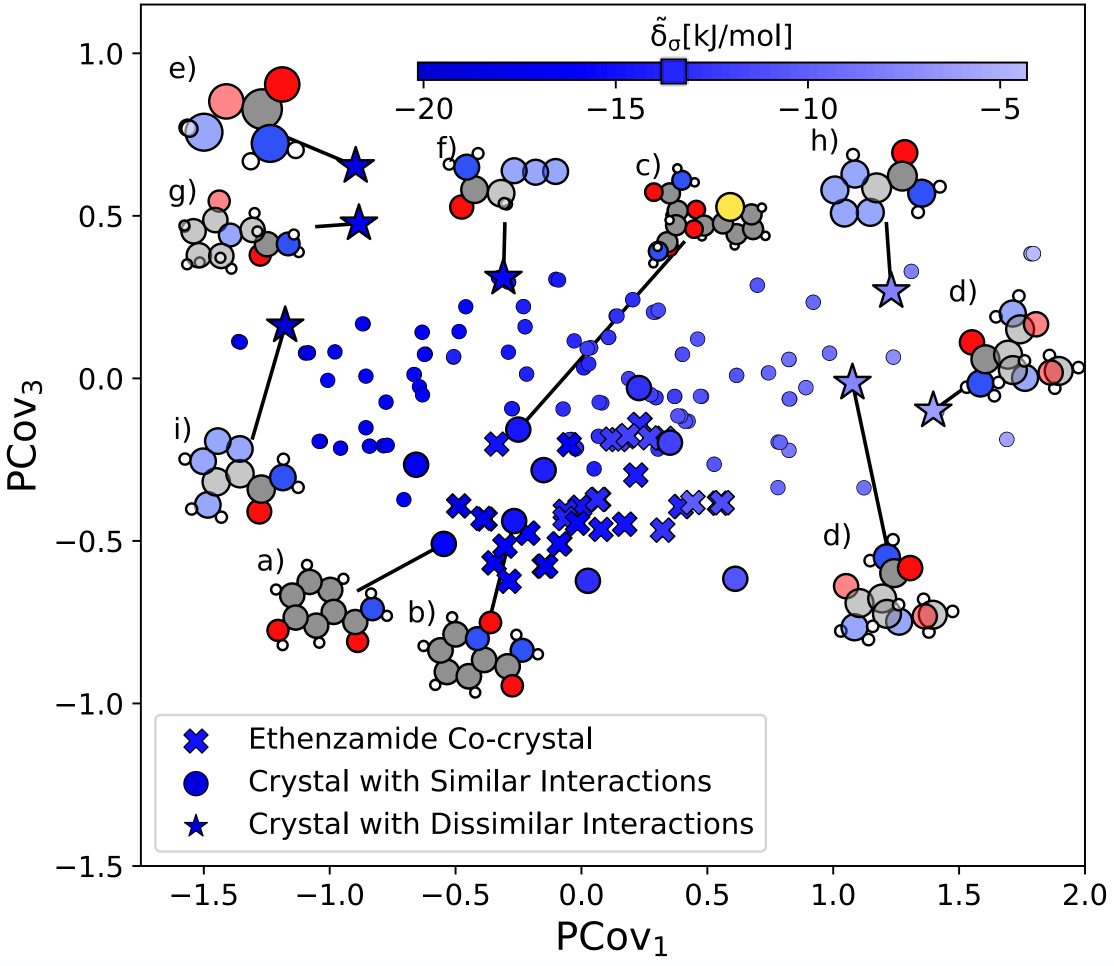


B. A. Helfrecht, **RKC**, G. Fraux, and M. Ceriotti. 2020 Mach. Learn.: Sci. Technol. 1 045021
C. J. Pickard. AIRSS Data for Carbon at 10GPa and the C+N+H+O System at 1GPa (2020).

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

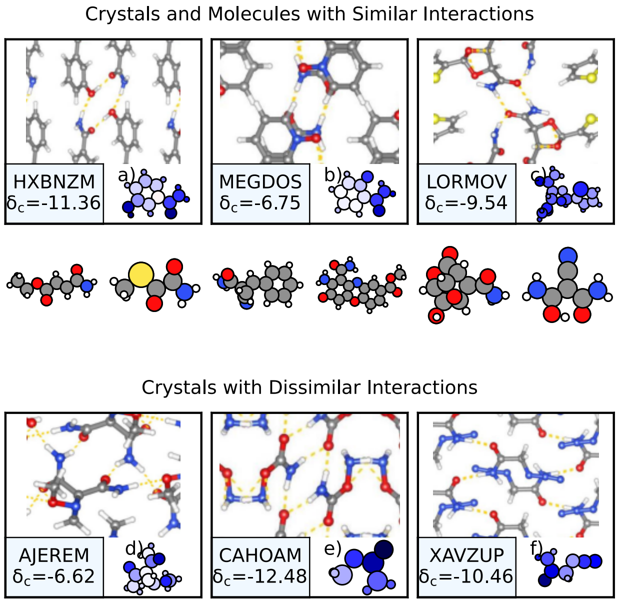


B. A. Helfrecht, **RKC**, G. Fraux, and M. Ceriotti. 2020 Mach. Learn.: Sci. Technol. 1 045021
C. J. Pickard. AIRSS Data for Carbon at 10GPa and the C+N+H+O System at 1GPa (2020).



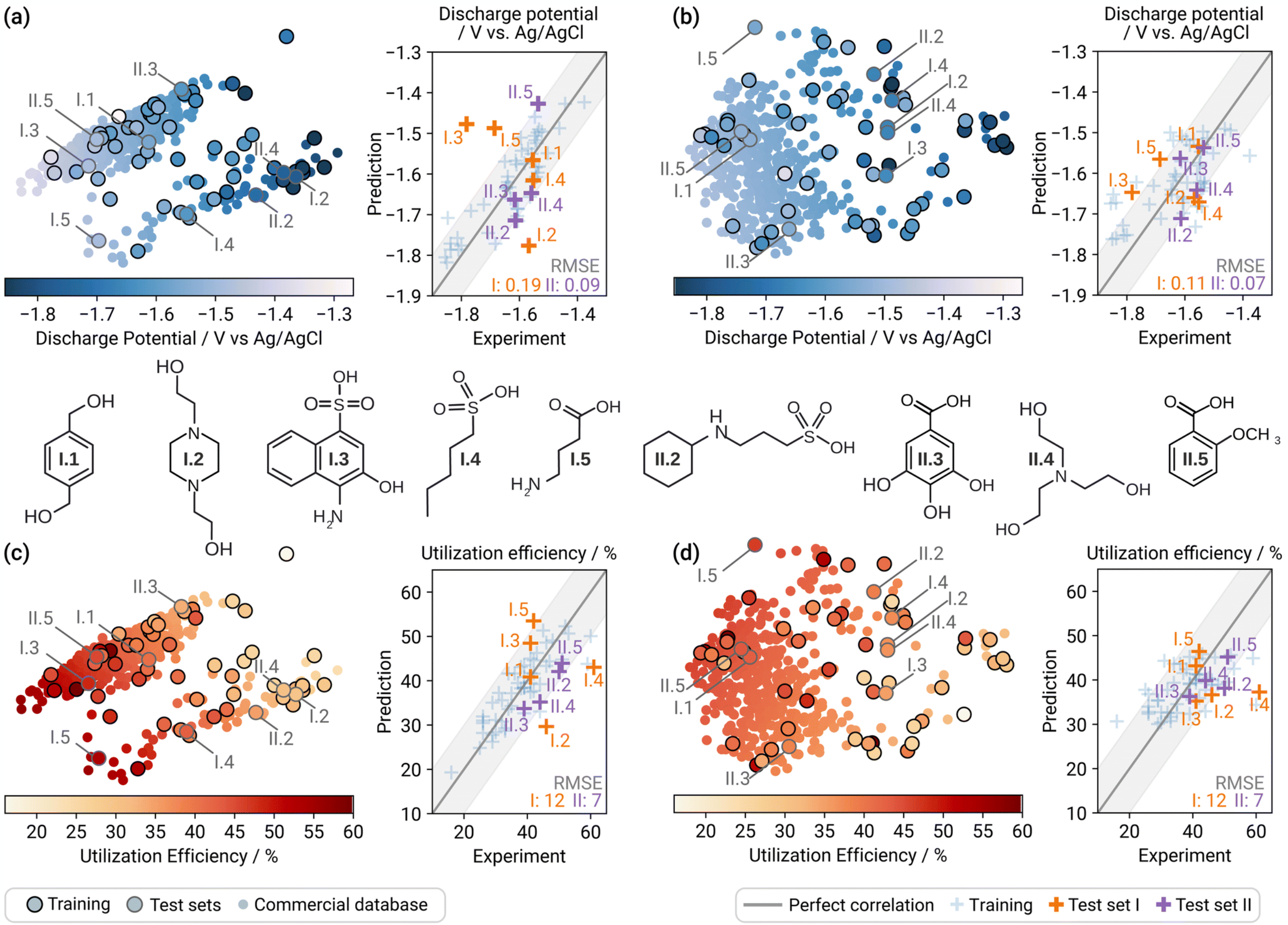
New co-former candidates

for the analgesic ethenzamide


Crystals and Molecules with Similar Interactions

Crystals with Dissimilar Interactions

RKC, et al., A data-driven interpretation of the stability of organic molecular crystals. 2023 Chem. Sci. 14, 1272–1285.



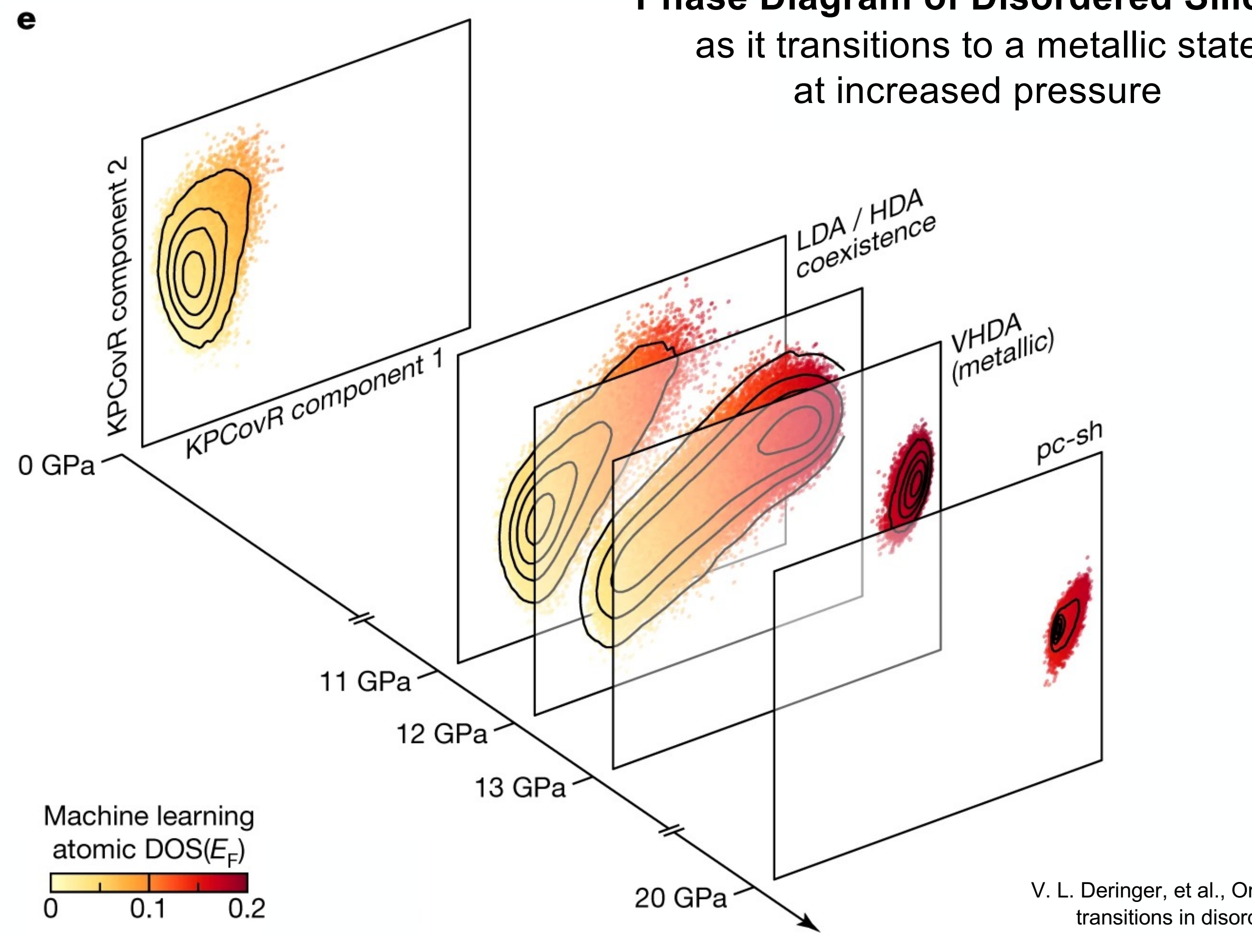
Discharge potential and utilization efficiency

for different electrolyte additives

Tim Würger, et al. Data-driven selection of electrolyte additives for aqueous magnesium batteries. J. Mater. Chem. A, 2022,10, 21672-21682

Phase Diagram of Disordered Silicon

as it transitions to a metallic state
at increased pressure

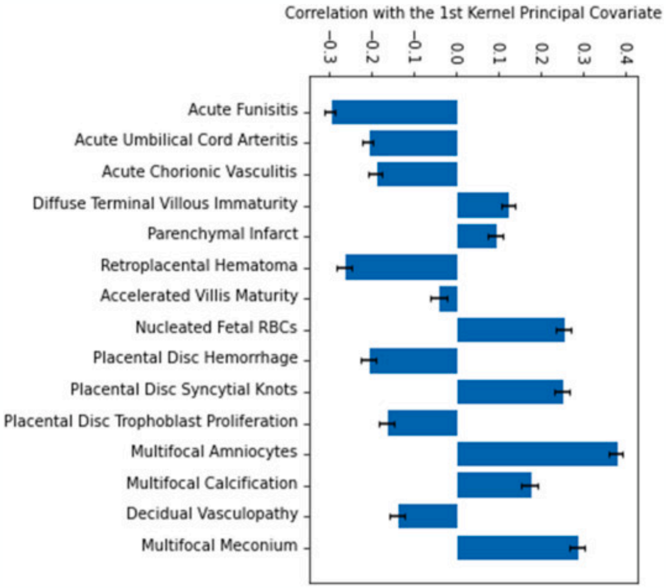


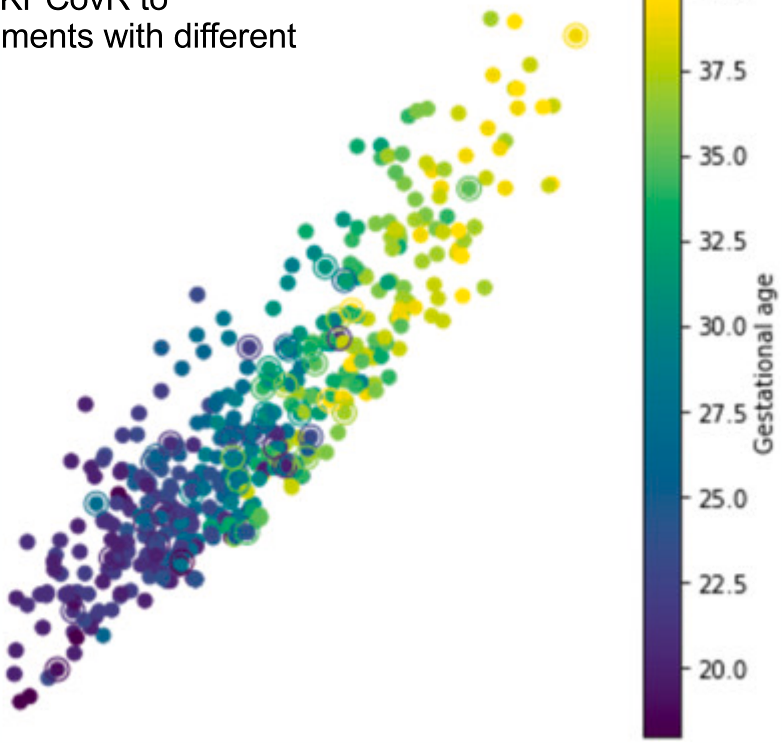
V. L. Deringer, et al., Origins of structural and electronic transitions in disordered silicon. Nature 589, 59–64 (2021). , pages 59–64 (2021).

Kernel Principal Covariates Regression can be useful beyond chemical contexts.

THE WARREN ALPERT
Medical School

In Cersonsky, Cersonsky, et al (2023), we used KPCovR to understand the correlation of placental measurements with different stillbirth outcomes.

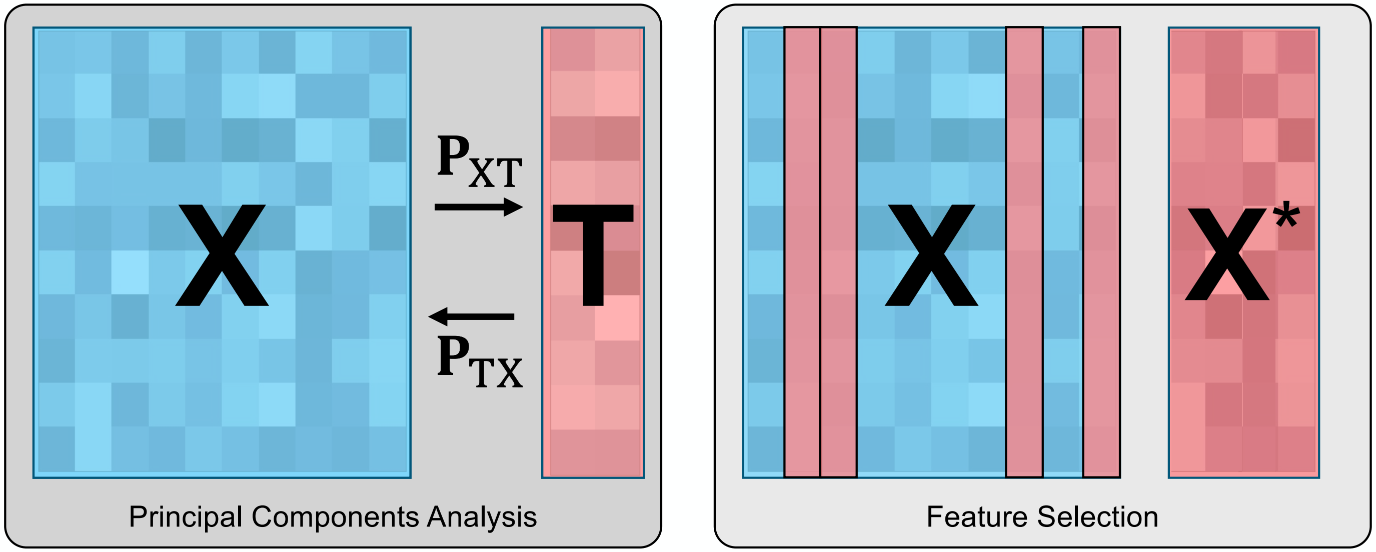




TEKC, RKC, et al., Placental lesions associated with stillbirth by gestational age, according to feature importance: Results from the stillbirth collaborative research network, Placenta, Volume 137, 2023, Pages 59-64,ISSN 0143-4004; "Placental Lesions Associated With Stillbirth by Gestational Age, as Related to Cause of Death: Follow-Up Results From the Stillbirth Collaborative Research Network." *Pediatric and Developmental Pathology* (2023): 10935266231197349.

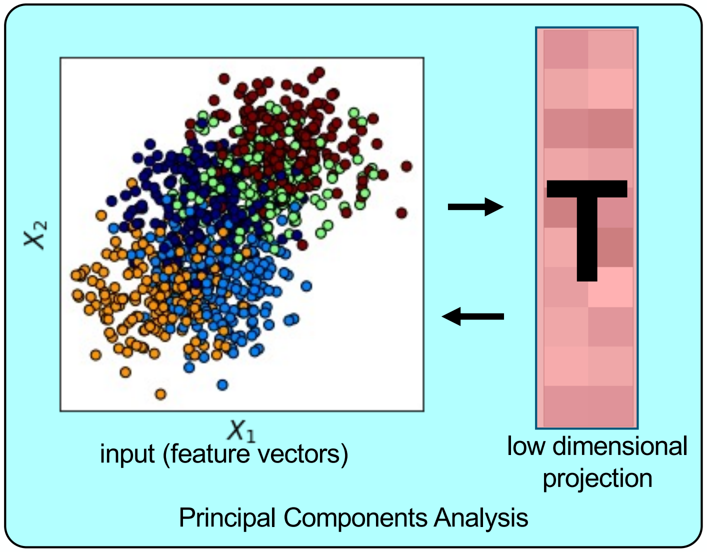
What if the features carry inherent meaning?

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?



What if the features carry inherent meaning?

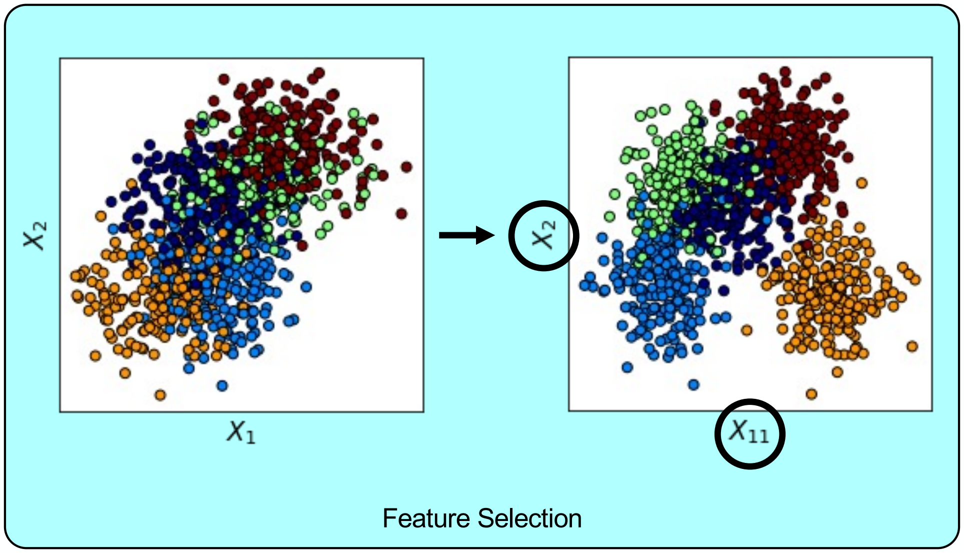
Many dimensionality reduction techniques construct a *new* set of features, but what if you want to just work with a subset of the old set?



input (feature vectors)

low dimensional projection

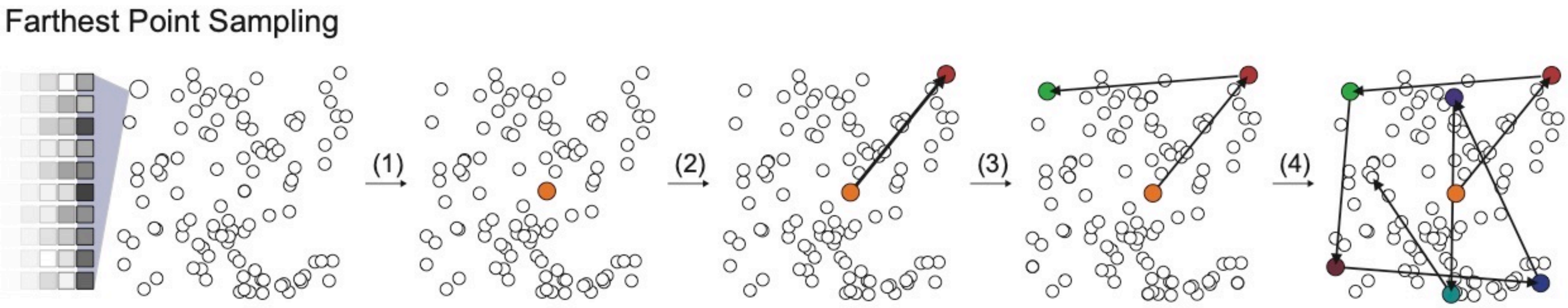
Principal Components Analysis



Feature Selection

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.

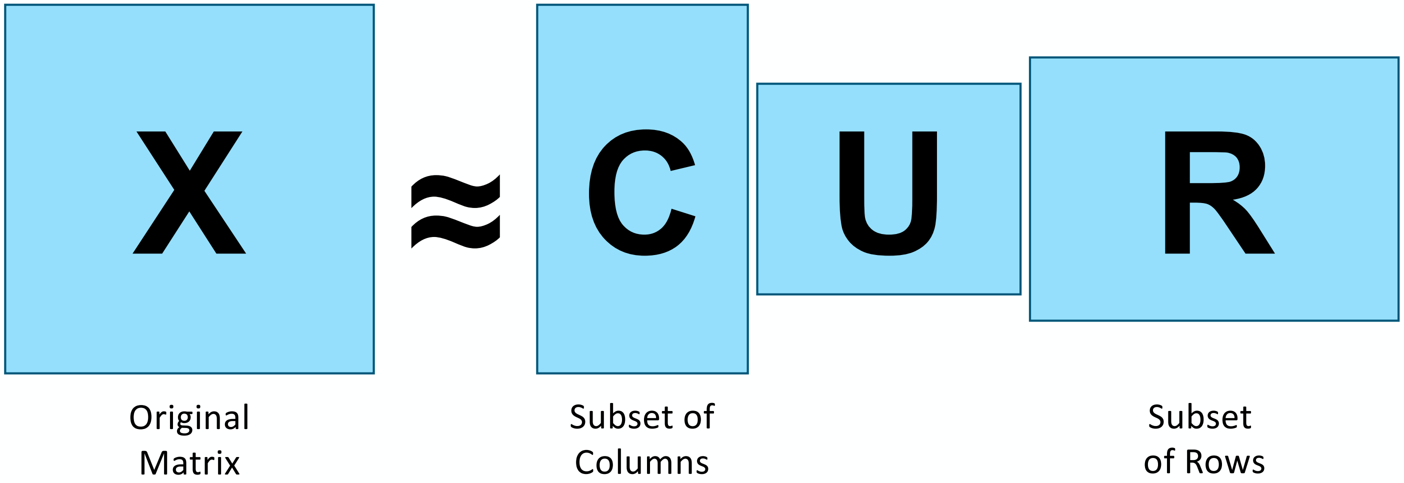


1. Choose a first point
2. Compute distance d
3. Choose point with highest $\min(d)$ to the selected points
4. Repeat 1-3 until you have enough features!

CUR Decomposition

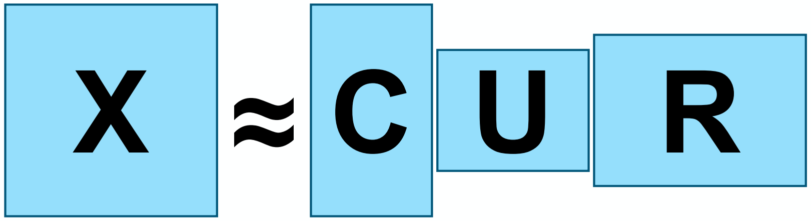
Traditional CUR decomposition selection aims to select “important” features or samples from the overall distribution.

Head Canon:
“Columns und Rows” decomposition



CUR Decomposition

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

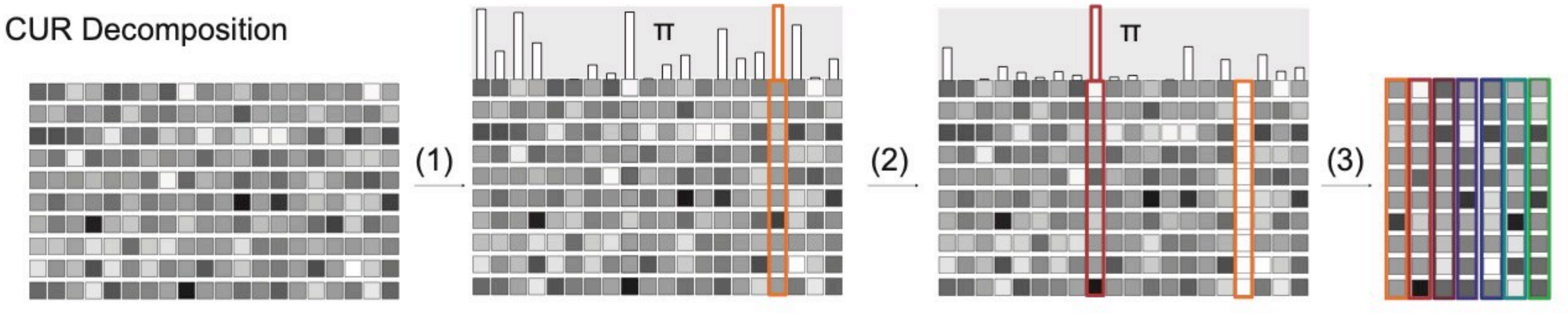


X ≈ C U R

Original Matrix

Subset of Columns

Subset of Rows

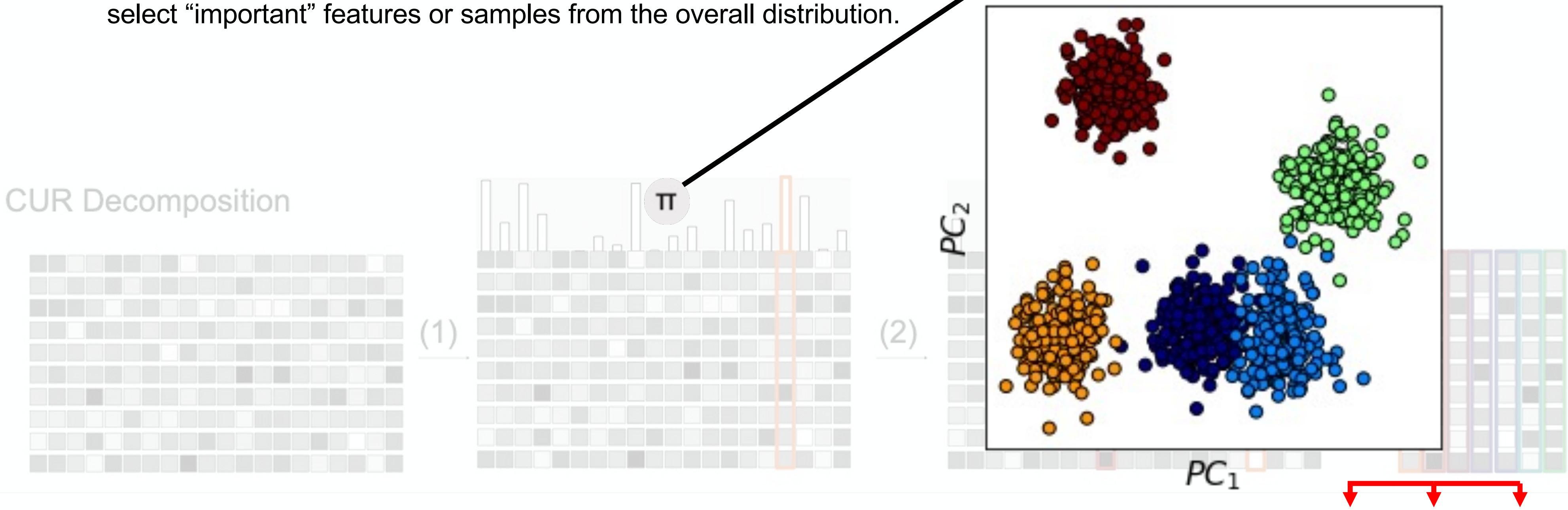


1. Compute importance score π
2. Choose column with highest π
3. Orthogonalize with respect to last chosen column.
4. Repeat 1-3 until you have enough features!

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

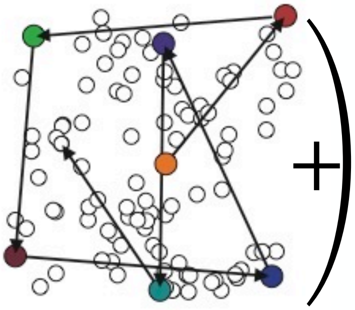
Both FPS and CUR can be translated to PCovR space
for both feature (and sample) selection.

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

feature selection

Farthest Point Sampling (FPS)

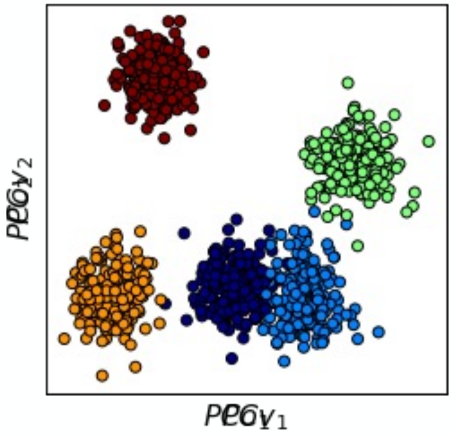
$d = f\Bigg($



$+$ Contribution to Regression Weights $\Bigg)$

CUR Decomposition

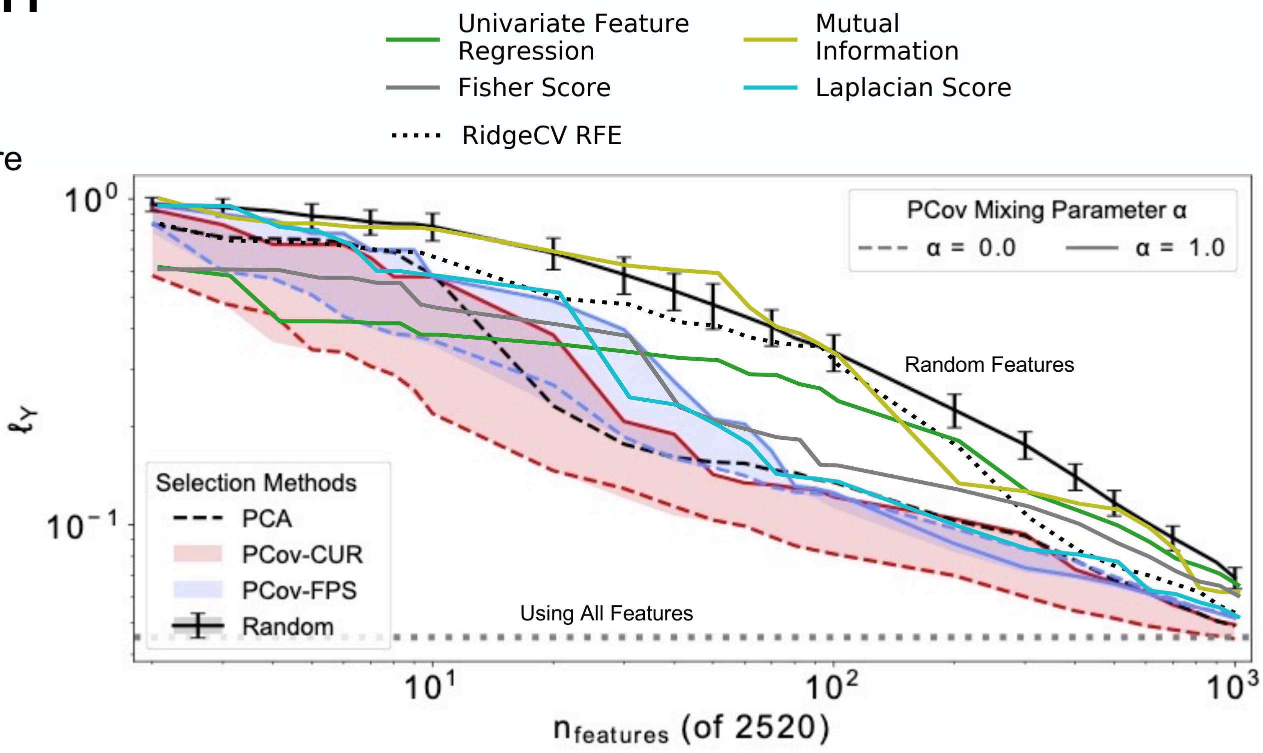
$\pi = f\Bigg($



$\Bigg)$

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

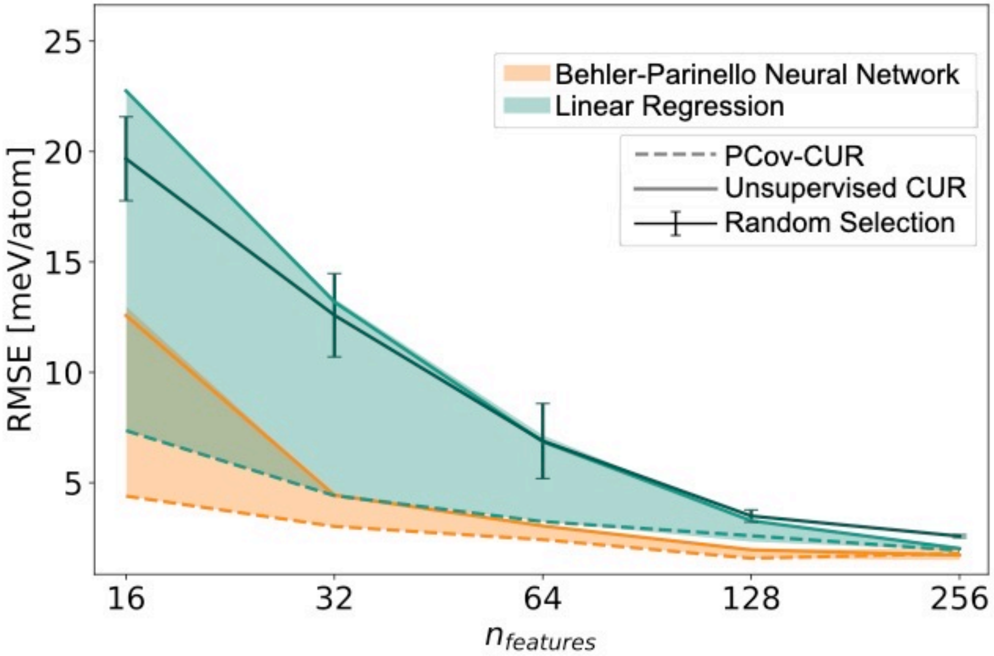
Target: NMR chemical shieldings in ppm

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

All of these functionalities are implemented in `scikit-matter` in the `scikit-learn` API.

```
X_scaled # some input matrix whose variance has been scaled to 1
y_scaled # some target matrix whose variance has been scaled to 1

# PCovR
from skmatter.decomposition import PCovR
pcovr = PCovR(mixing=0.5, n_components=2)
pcovr.fit(X_scaled, y_scaled)
T = pcovr.transform(X_scaled)

# KPCovR with RBF kernel
from skmatter.decomposition import KernelPCovR
kpcovr = KPCovR(mixing=0.5, kernel=‘rbf’, gamma=0.1, n_components=2)
kpcovr.fit(X_scaled, y_scaled)
T = pcovr.transform(X_scaled)

# PCov-CUR
from skmatter.feature_selection import PCovCUR
cur = PCovCUR(mixing=0.5, n_to_select=10)
cur.fit(X_scaled, y_scaled)
X_select = cur.transform(X_scaled)
```

scikit-matter is a collection of scikit-learn compatible utilities that implement methods born out of the materials science and chemistry communities.

scikit-matter.readthedocs.io

A. Goscinski, …, **RKC**, 2023 Open Research Europe, 3(81).
https://doi.org/10.12688/openreseurope.15789.1

Density-based machine learning representations provide a way to characterize systems at an atomistic or molecular level and enable *interpretable* machine learning.

Methods & Software:

Feature Reconstruction Errors: A. Goscinski et al 2021 MLST 2 025028

Structure-property mappings: B. A. Helfrecht, **RKC**, et al. 2020 MLST1 045021.

Feature subselection: **RKC**, et al. 2021 MLST 2 035038.

Unsupervised Learning for Quantum Chemistry: **RKC**, S. De. 2022, *Elsevier*.

`pip install chemiscope`: G. Fraux, **RKC**, et al. 2020 JOSS 5(51), 2117.

`pip install skmatter`*:* A. Goscinski, …, **RKC**, 2023 Open Research Europe, 3(81).

Applications:

RKC, et al., 2023 *Chem. Sci.* **14**, 1272–1285.

T.E.K. Cersonsky, **RKC,** et al., 2023 *Placenta.* Volume 137.

T.E.K. Cersonsky, **RKC**, et al., 2023 *Ped. and Developmental Pathology* 10935266231197349.

Tim Würger, et al. *J. Mater. Chem. A,* 2022,10, 21672-21682

V. L. Deringer, et al.,. *Nature* 589, 59–64 (2021). , pages 59–64.

If current trends do not change, fields such as chemical engineering and materials science will not reach gender parity in our lifetimes. Why is this? What can we do?

Not Yet Defect Free: The Currently Landscape for Women in Computational Materials Research. L. B. Pàrtay, E. Teich, R.K. Cersonsky. https://www.nature.com/articles/s41524-023-01054-z

Our work is generously supported by:

WARF
Wisconsin Alumni Research Foundation

University of Wisconsin
MRSEC