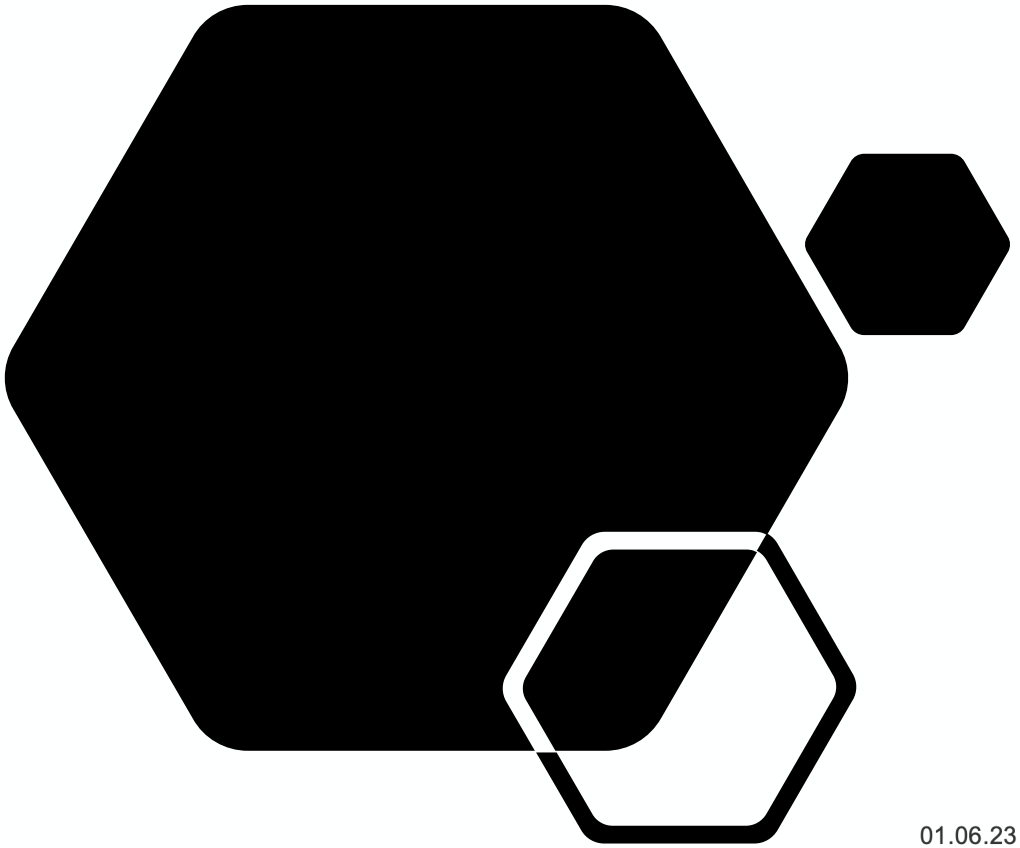
Leveraging Machine-Learning to Understand the Structure-Property Paradigm

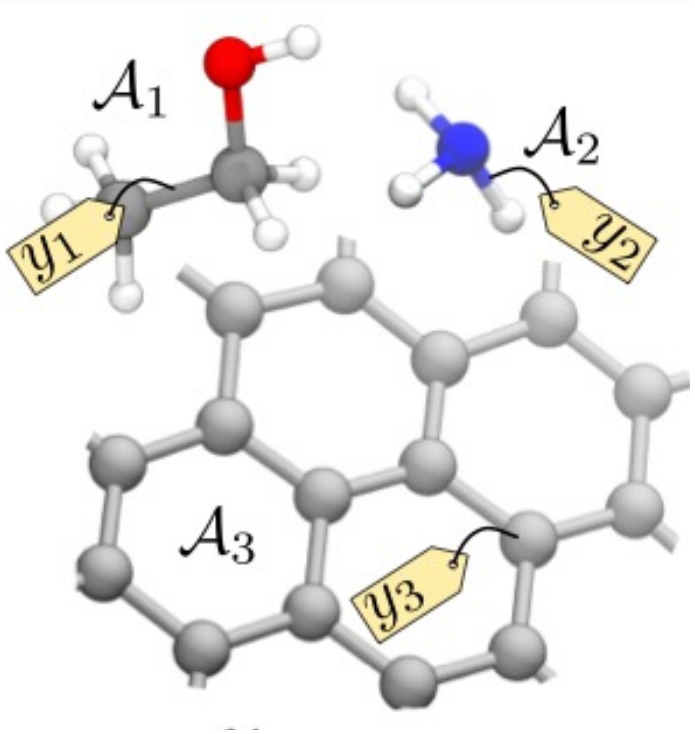
Rose K. Cersonsky

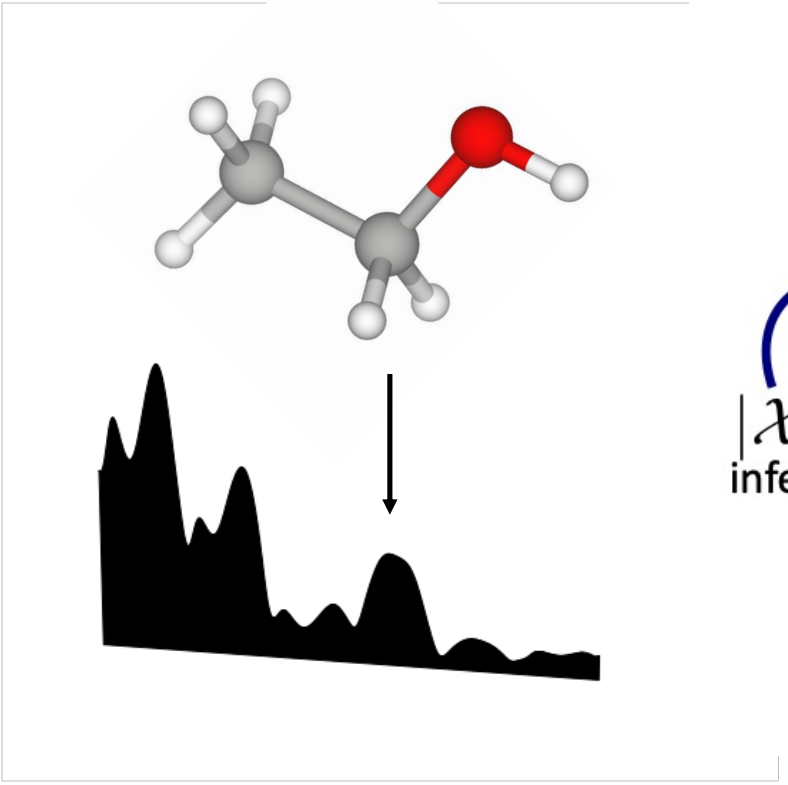
University of Wisconsin, Chemical and Biological Engineering

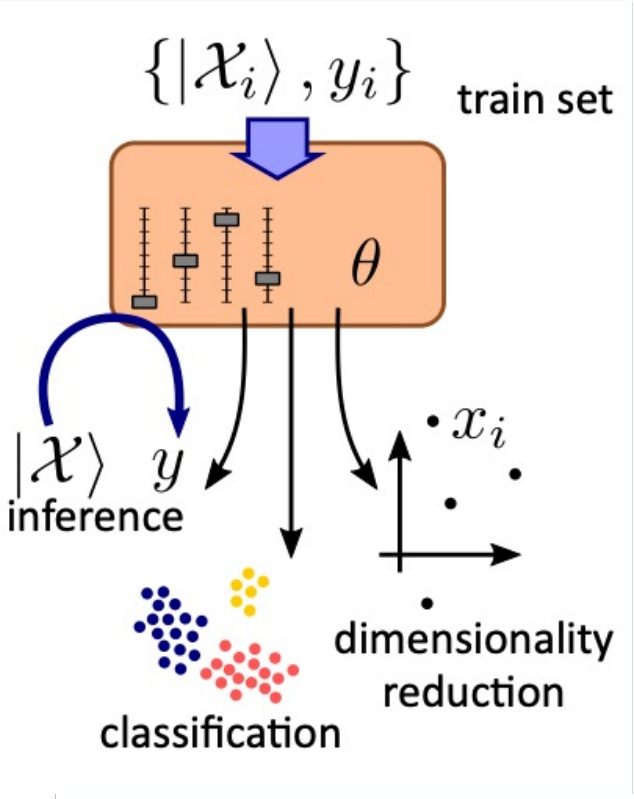


01.06.23

Chemical Data	Numerical Representation	Machine Learning Model



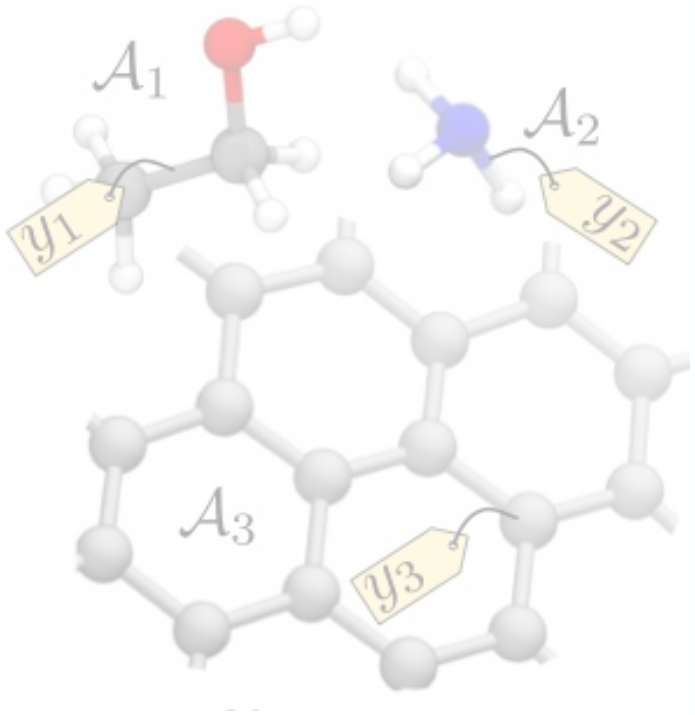


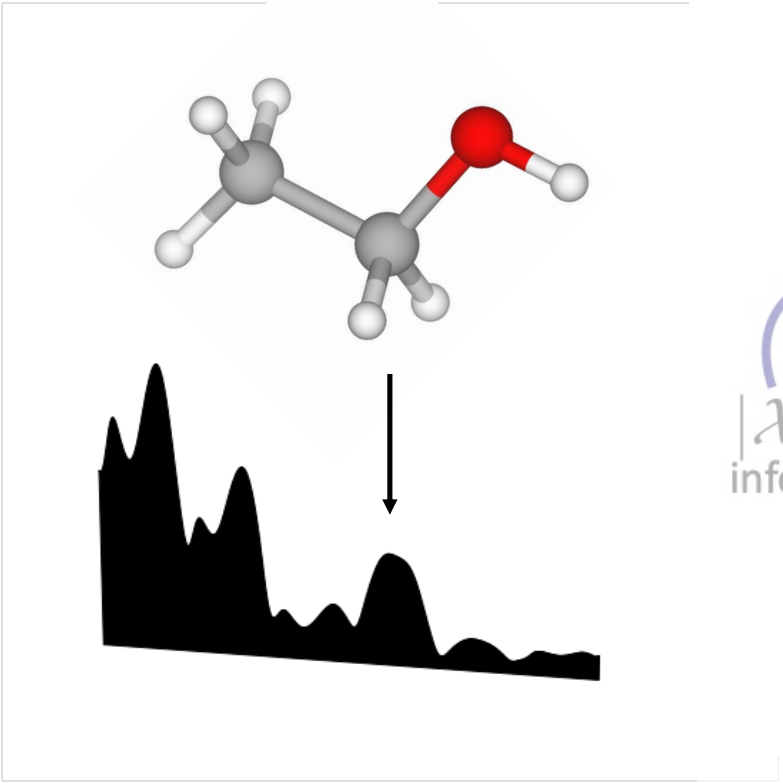


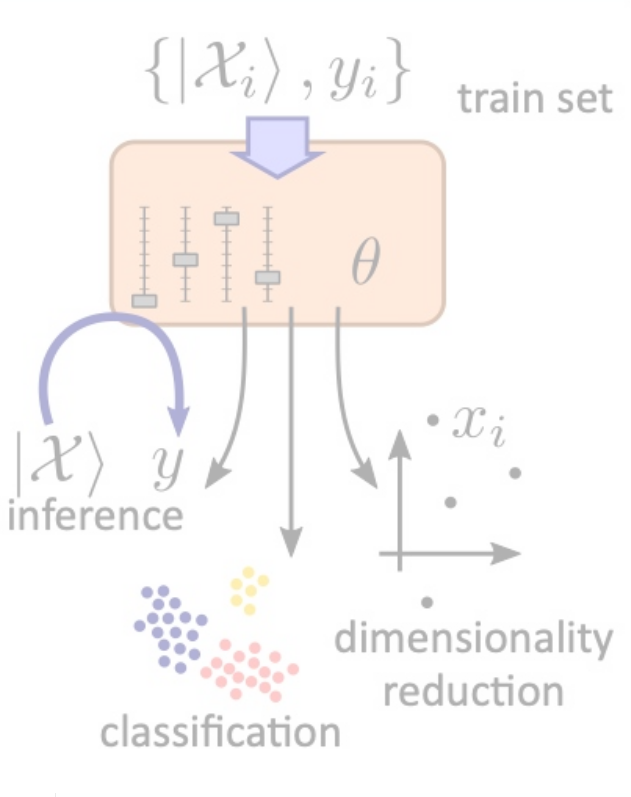
Chemical Data

Numerical Representation

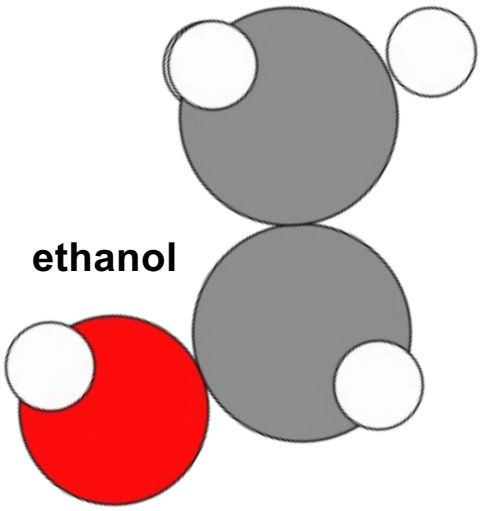
Machine Learning Model



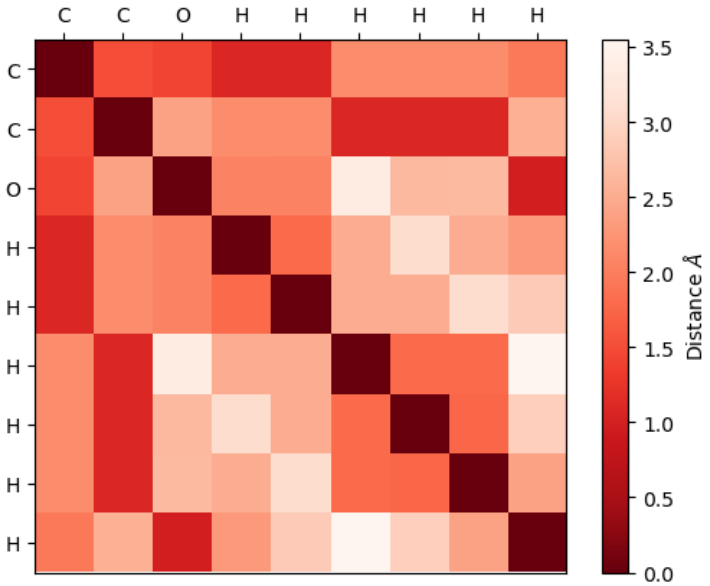




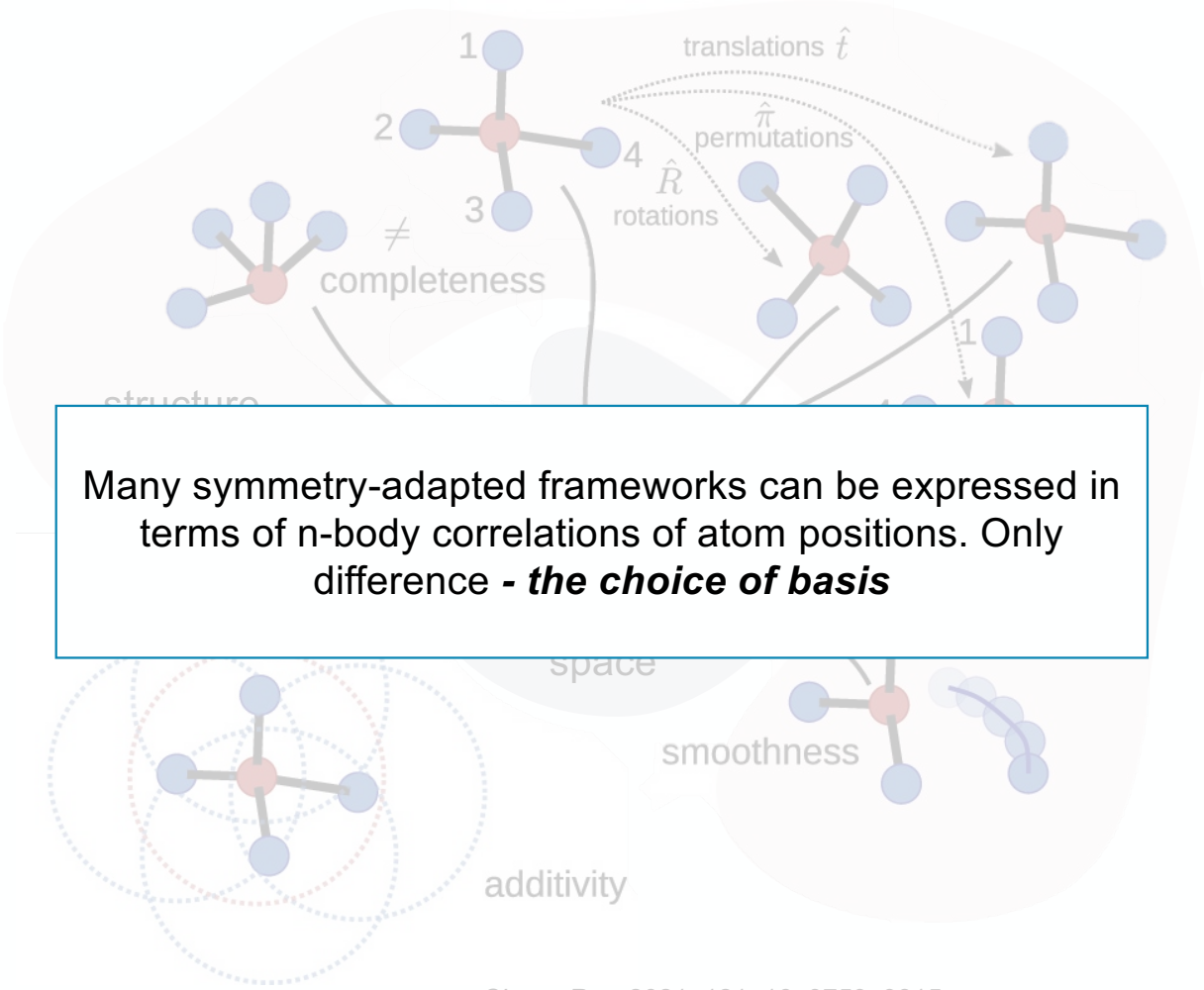
There are many ways to numerically encode configurations in chemistry.



	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



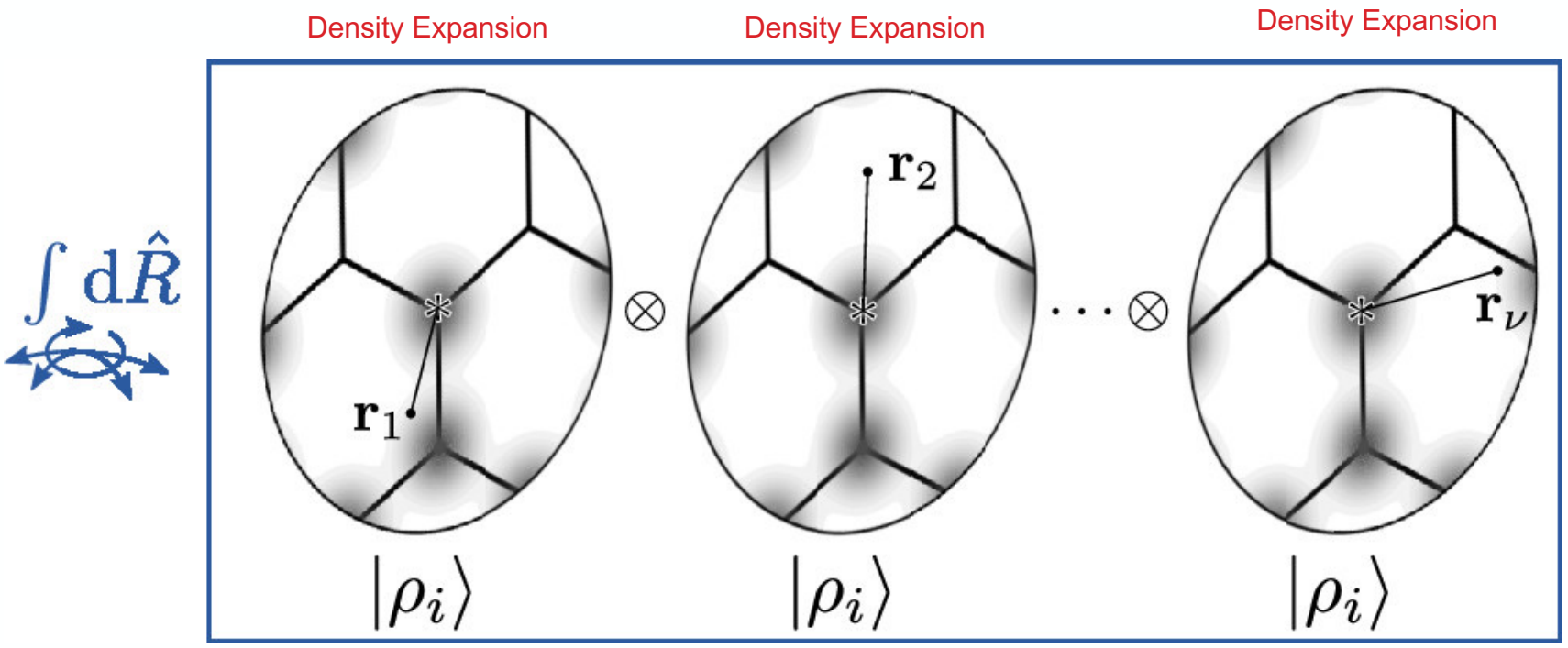
In thermodynamic contexts, what do we want from a representation?



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

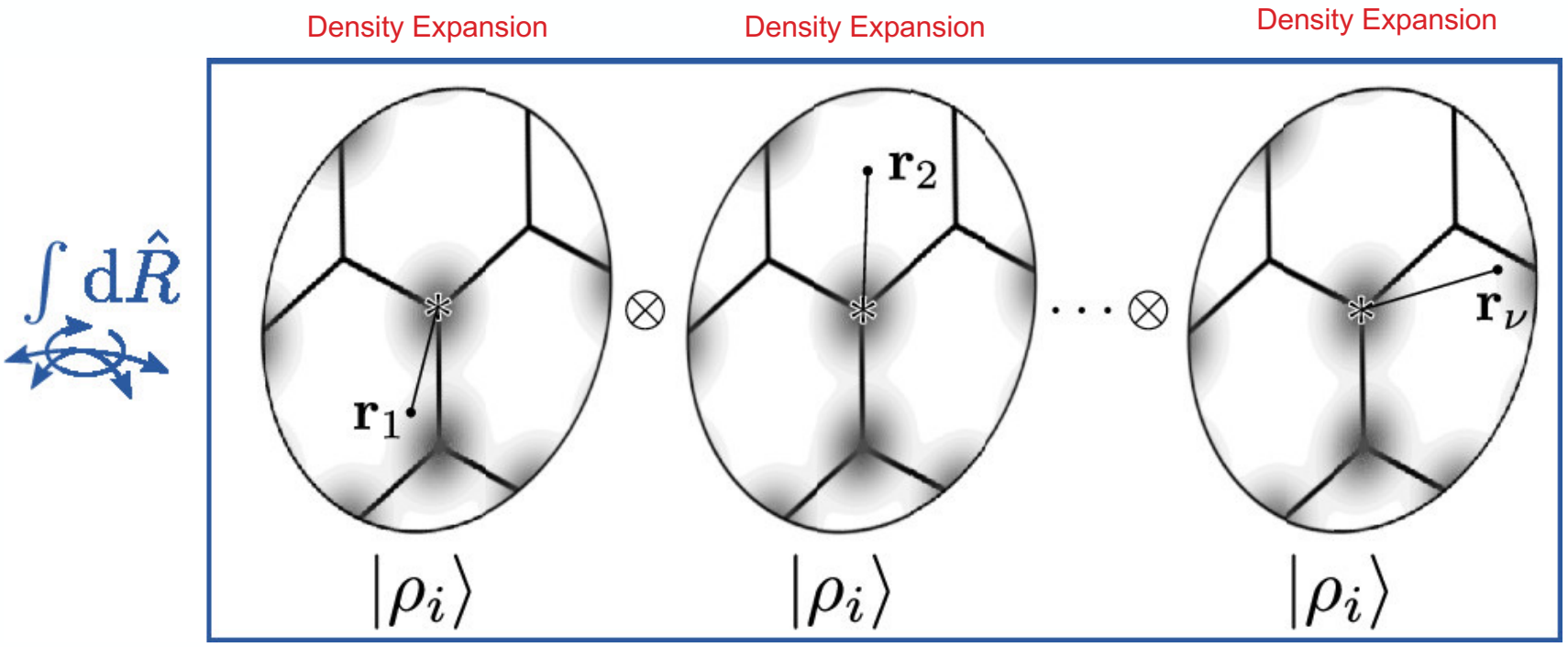
Chem. Rev. 2021, 121, 16, 9759–9815

SOAP vectors are the n-body correlations of atomic densities.



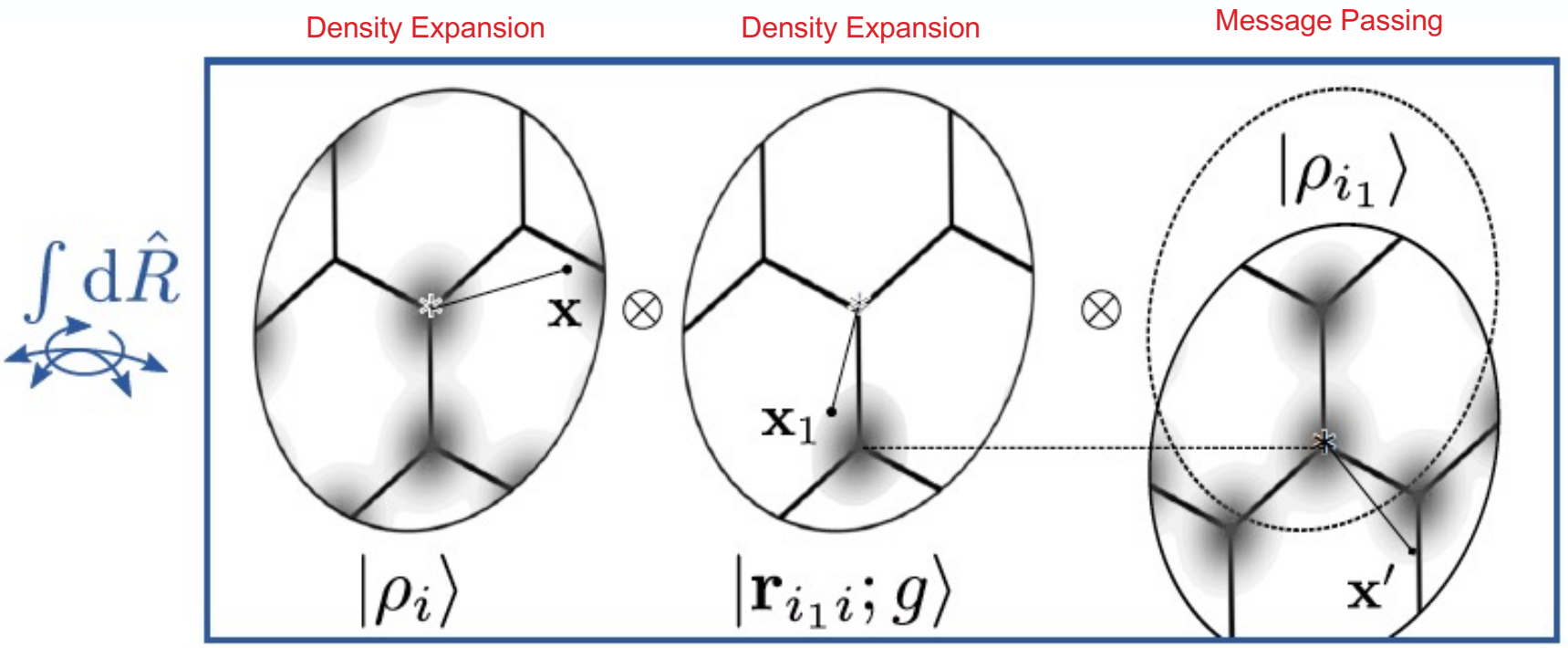
Chem. Rev. 2021, **121**, 16, 9759–9815
Phys. Rev. B 2013, **87**, 184115.

NICE vectors (and the similar MACE framework) are the n-body correlations of atomic densities, contracted efficiently include higher body-order correlations.



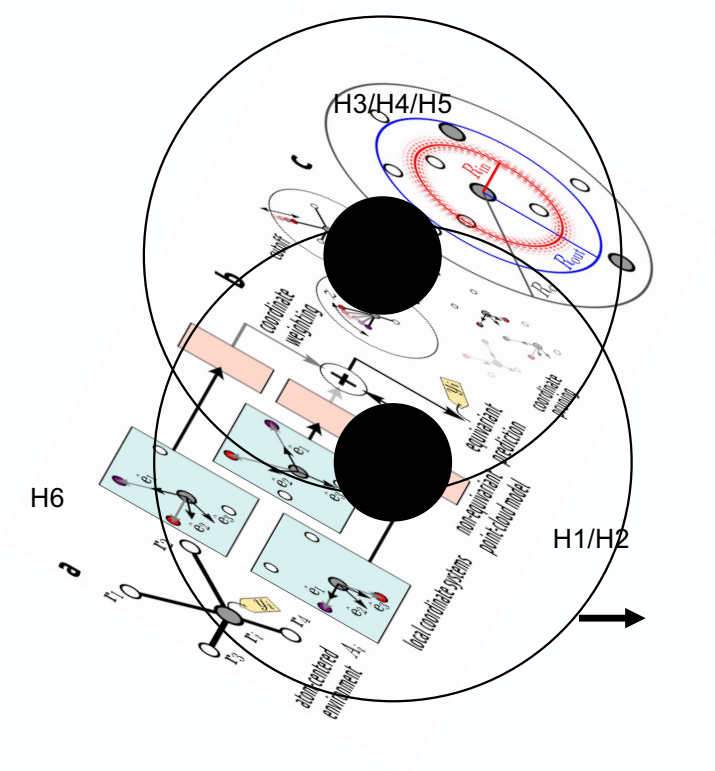
Chem. Rev. 2021, 121, 16, 9759–9815
JCP 2020, 153, 12, 121101.

Atom-centered density correlations (ACDCs) can also be formulated to be consistent with typical message-passing frameworks.

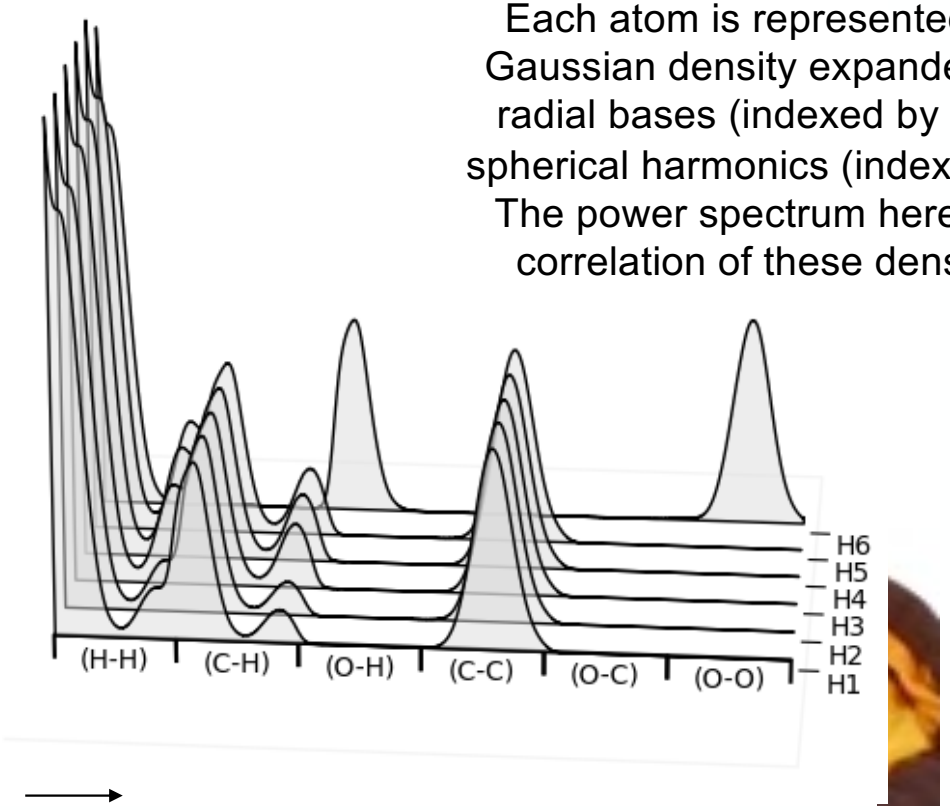


J. Chem. Phys. 156, 204115 (2022)

A collection of atoms can be represented by the combination of the atomic fingerprints.



Each atom is represented as a Gaussian density expanded over radial bases (indexed by n) and spherical harmonics (indexed by l). The power spectrum here is the correlation of these densities.



Power Spectrum SOAP Vectors
Computed with n=12, l=9, rcut=2.0, σ=0.3
Filtered through a Gaussian filter for visual clarity

With these representation spaces,

(PROS) we gain…

- an “agnostic” way of describing molecular configurations
- an increased accuracy in predicting thermodynamic quantities within shallower model infrastructures

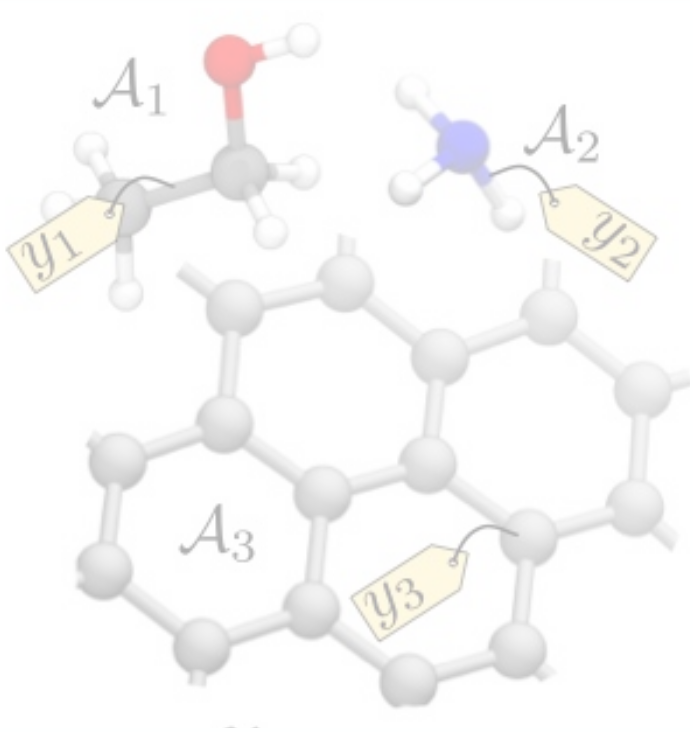
(CONS) we lose…

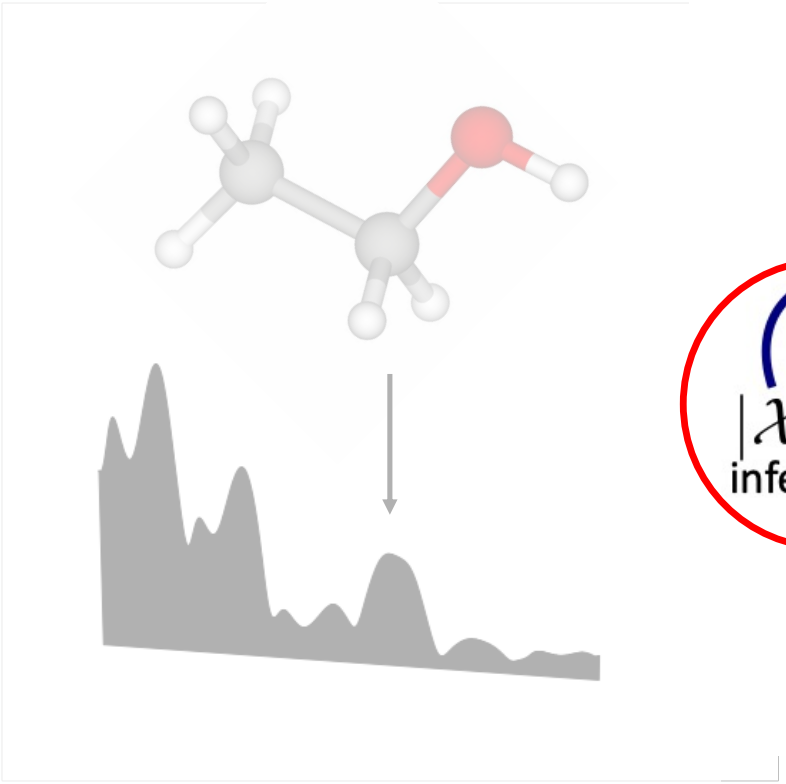
- compactness
- human readability when that human doesn’t spend all day looking at these fingerprints

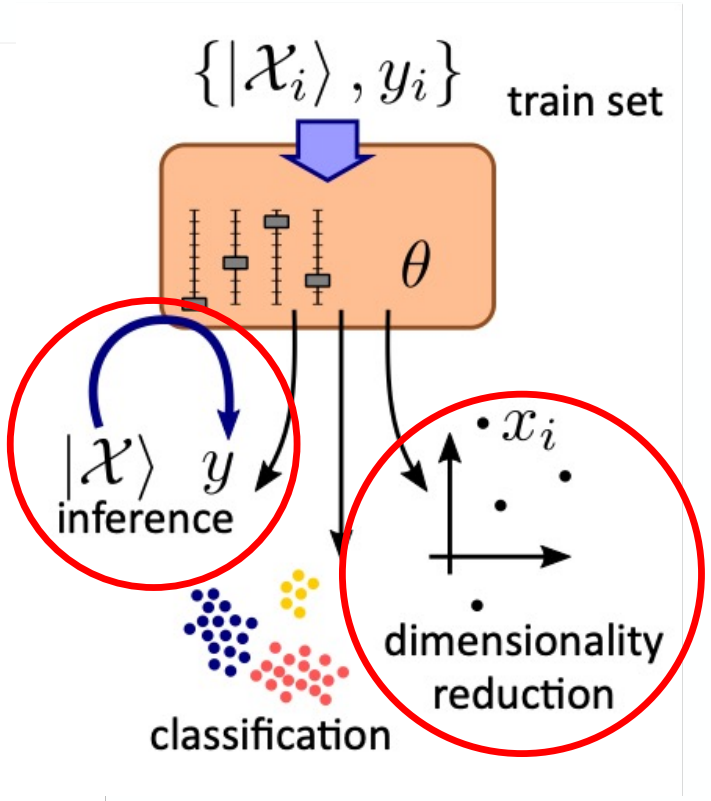
Chemical Data

Numerical Representation

Machine Learning Model







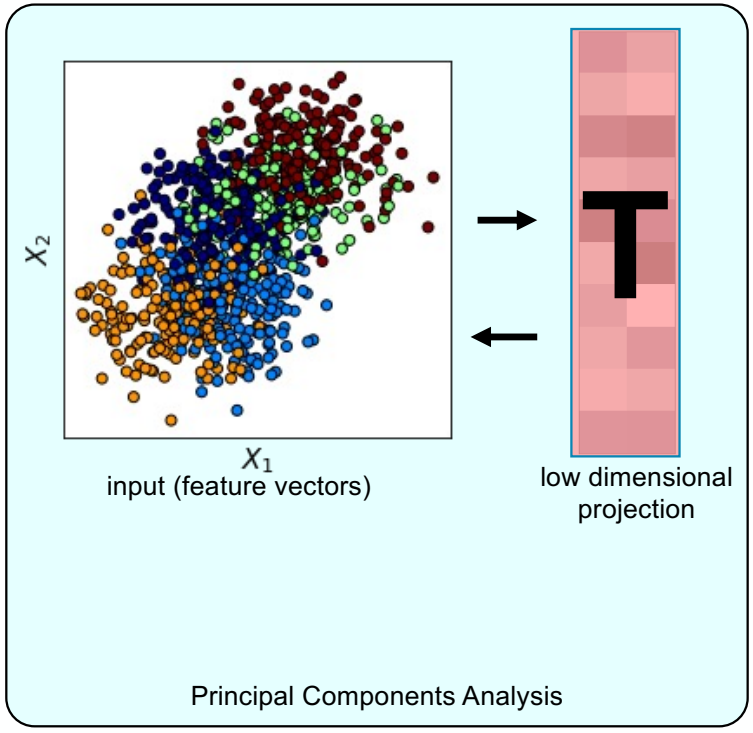
A couple words on notation…

$\mathbf{X} = \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \\ \dots \end{bmatrix}$	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
$\mathbf{K} = \begin{bmatrix} \mathbf{k}(\mathbf{x}_1, \mathbf{x'}_1) & \cdots & \mathbf{k}(\mathbf{x}_1, \mathbf{x'}_N) \\ \vdots & \ddots & \vdots \\ \mathbf{k}(\mathbf{x}_N, \mathbf{x'}_1) & \cdots & \mathbf{k}(\mathbf{x}_1, \mathbf{x'}_N) \end{bmatrix}$	A matrix containing the similarity kernel between two datasets

$\mathbf{P}_{AB}$	A matrix that projects from space $\mathbf{A}$ to space $\mathbf{B}$
$\mathbf{T} = \mathbf{X}\mathbf{P}_{XT}$	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.



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

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)

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

gram matrix

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

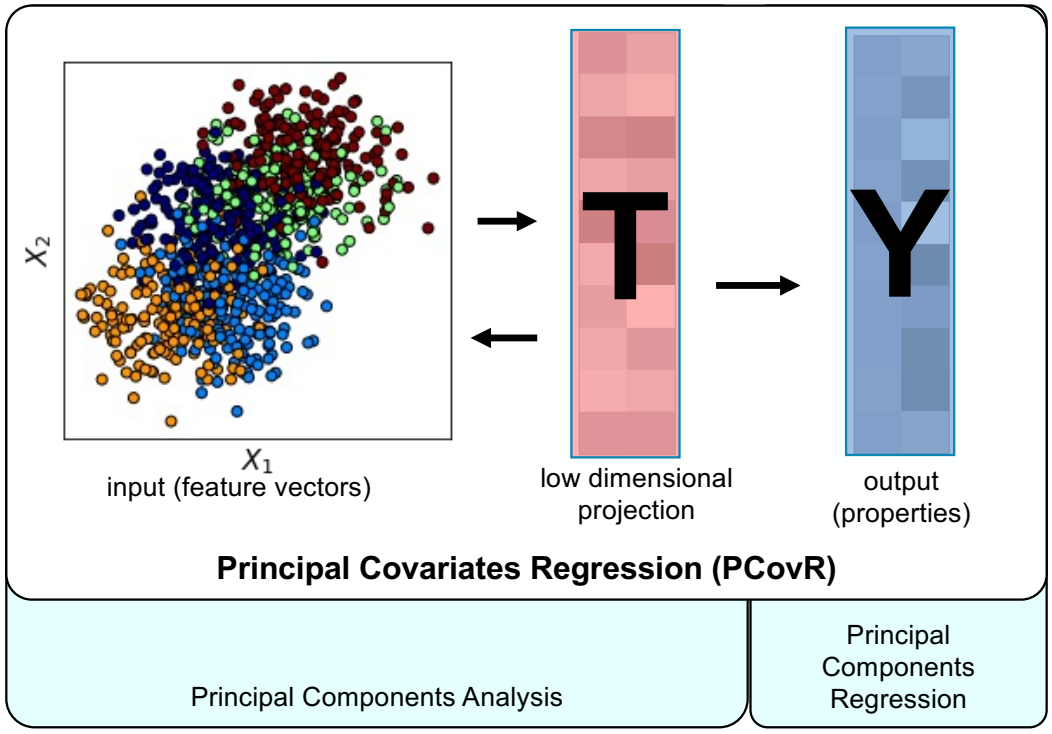
covariance matrix

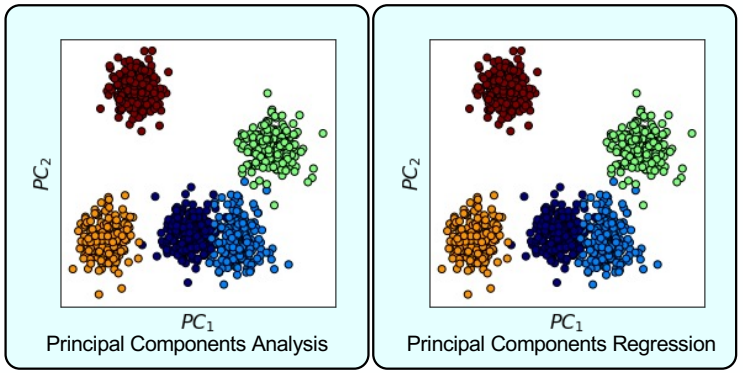
S. de Jong, H.A.L. Kiers, Chemom. intell. lab. syst. 14 (1992) 155-164.
scikit-matter.readthedocs.io

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

Principal Covariates Regression (PCovR)

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



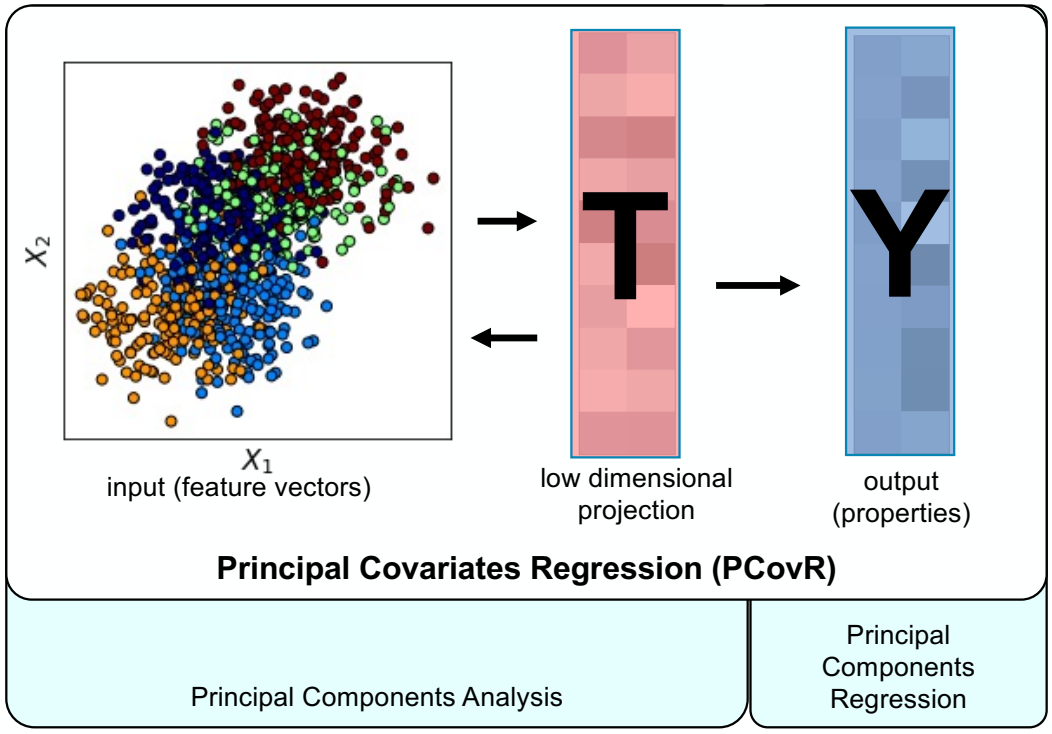


S. de Jong, H.A.L. Kiers, Chemom. intell. lab. syst. 14 (1992) 155-164.
scikit-matter.readthedocs.io

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

Principal Covariates Regression (PCovR)

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



loss in reconstructing X

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

loss in reconstructing Y

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

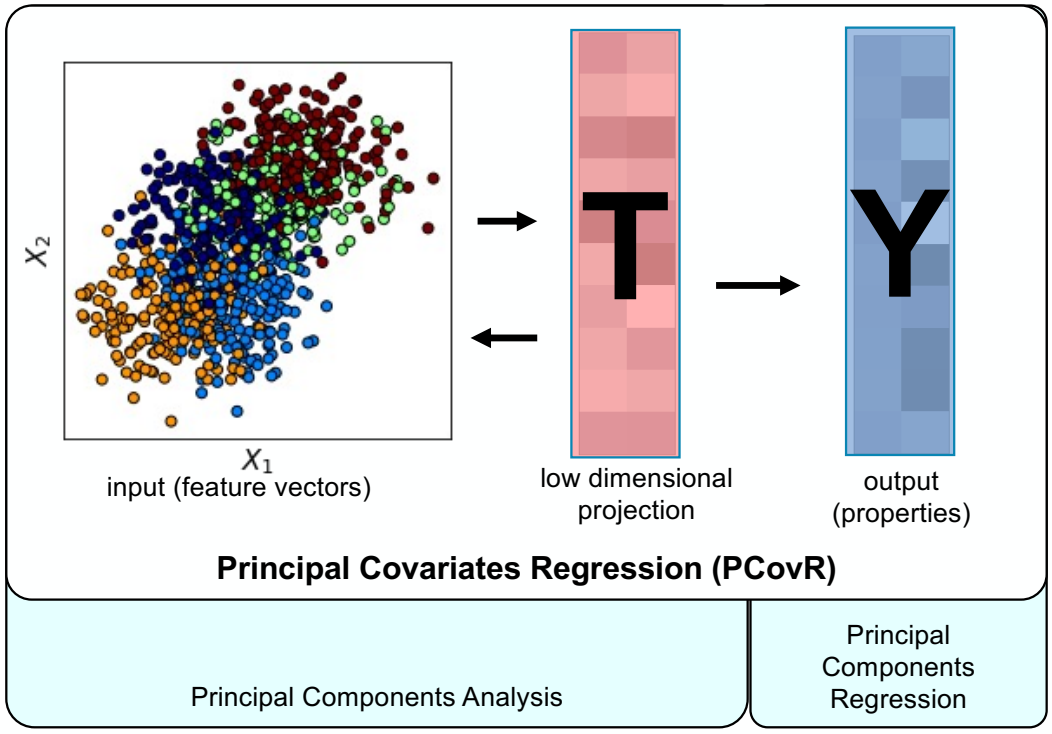
$$\mathbf{K} \rightarrow \tilde{\mathbf{K}} \qquad \mathbf{C} \rightarrow \tilde{\mathbf{C}}$$

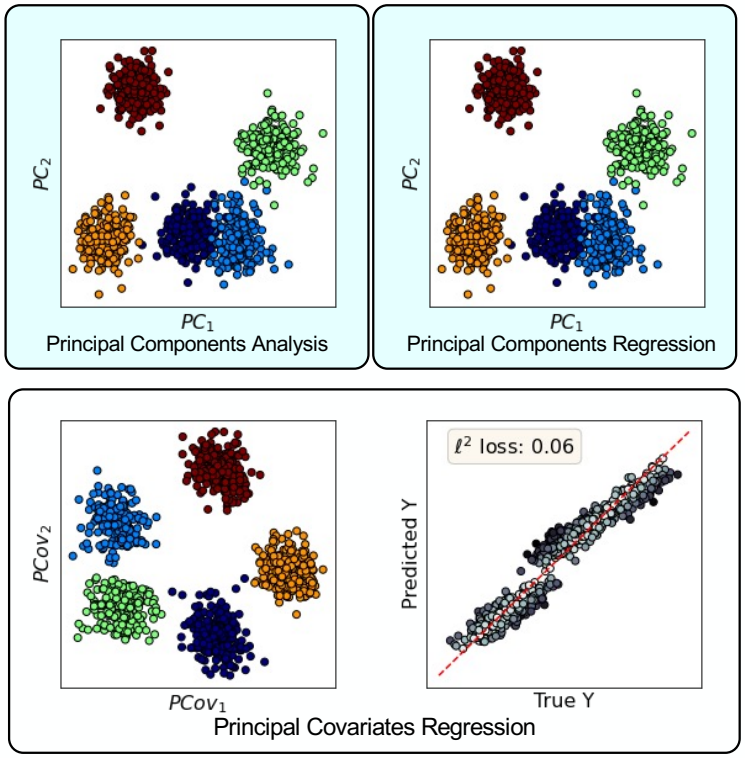
$$\tilde{\mathbf{K}} = \alpha \mathbf{X}\mathbf{X}^{\mathbf{T}} + (1-\alpha)\hat{\mathbf{Y}}\hat{\mathbf{Y}}^{\mathbf{T}} \qquad \tilde{\mathbf{C}} = (\mathbf{C}^{-1/2}\mathbf{X}^{\mathbf{T}})\tilde{\mathbf{K}}(\mathbf{X}\mathbf{C}^{-1/2})$$

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

Principal Covariates Regression (PCovR)

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

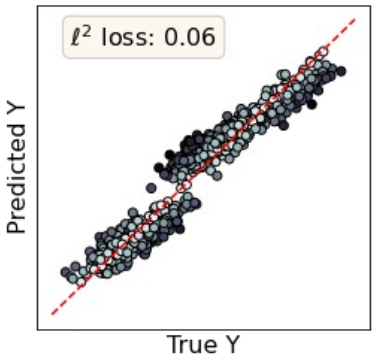


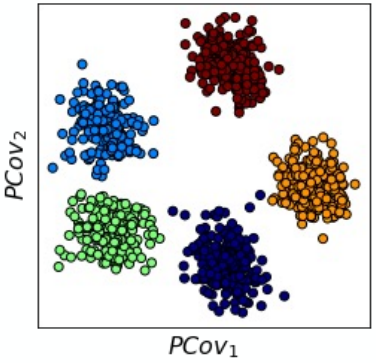


S. de Jong, H.A.L. Kiers, Chemom. intell. lab. syst. 14 (1992) 155-164.
scikit-matter.readthedocs.io

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

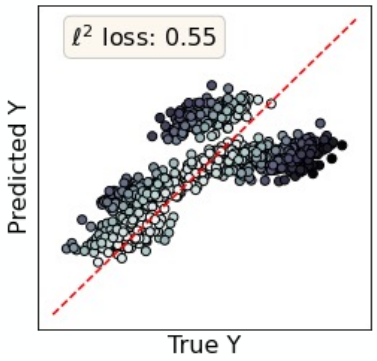
PCovR is controlled by a mixing parameter α that weights the regression and decomposition tasks.

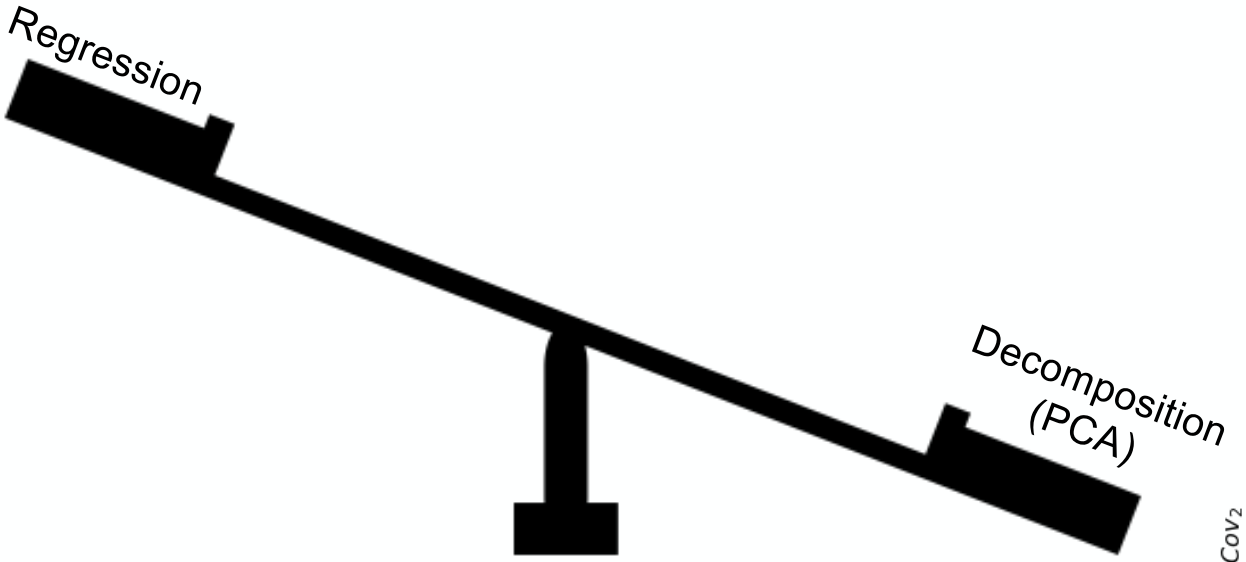




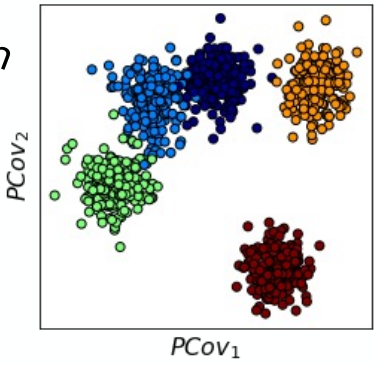
$$\alpha = 0.5$$

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

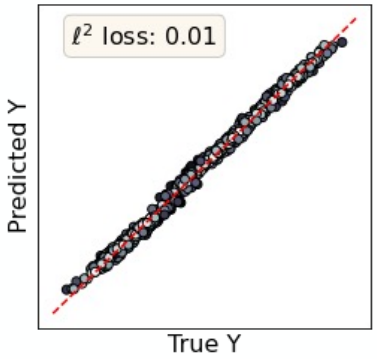


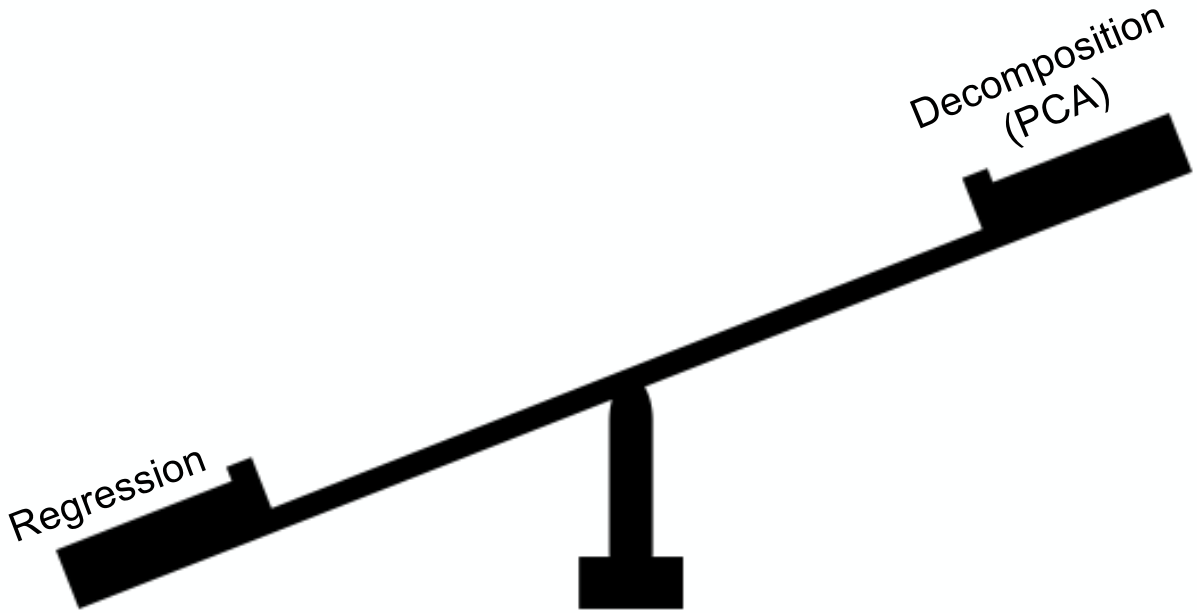


$\alpha = 0.1$

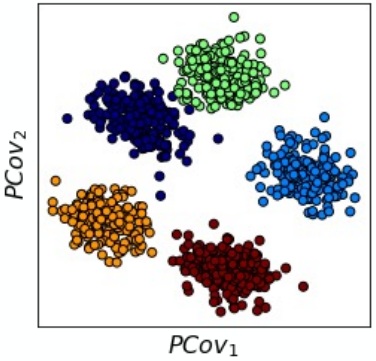


PCovR is controlled by a mixing parameter α that weights the regression and decomposition tasks.

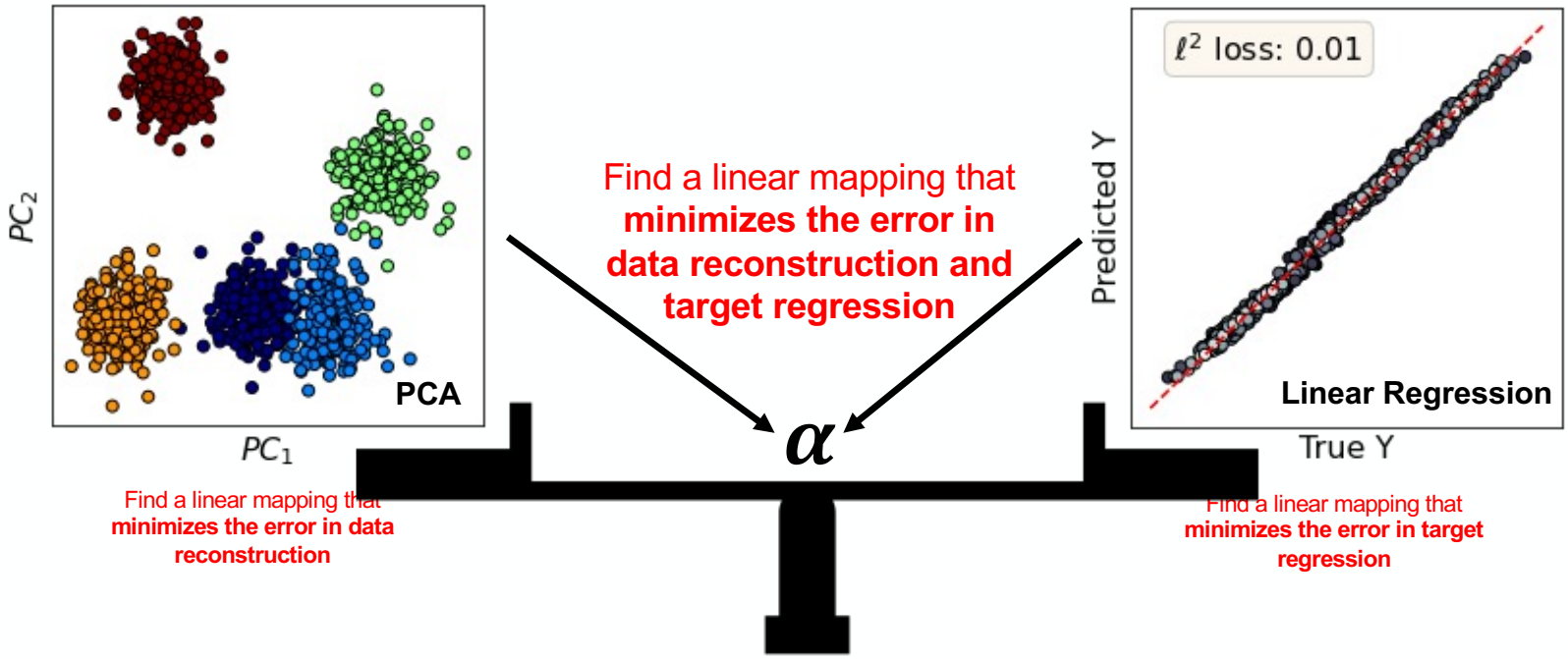




$\alpha = 0.9$



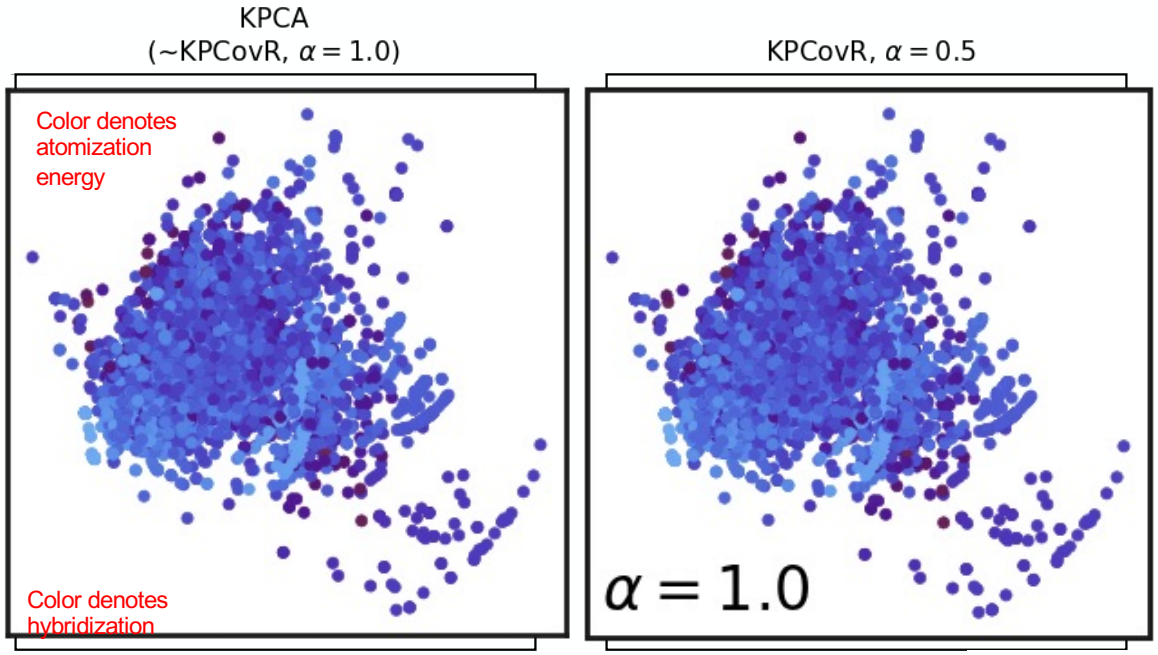
Principal Covariates Regression (PCovR) is akin to a Principal Components Analysis (PCA) but rotates the projection in hyperspace to correlate with a property of interest.



BAH, **RKC**, GF, MC. *Mach. Learn.: Sci. Technol. 1 045021.*

Kernel Principal Covariates Regression

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

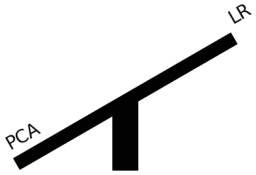


$$\tilde{\mathbf{K}} = \alpha \mathbf{X}\mathbf{X}^{\mathrm{T}} + (1-\alpha)\hat{\mathbf{Y}}\hat{\mathbf{Y}}^{\mathrm{T}}$$

gram matrix,
a.k.a. “linear kernel”

non-linear kernel

$$\mathrm{K}_{ij} = \mathrm{k}(\mathbf{x}_i, \mathbf{x}_j) = e^{-\gamma \|\mathbf{x}_i - \mathbf{x}_j\|^2}$$

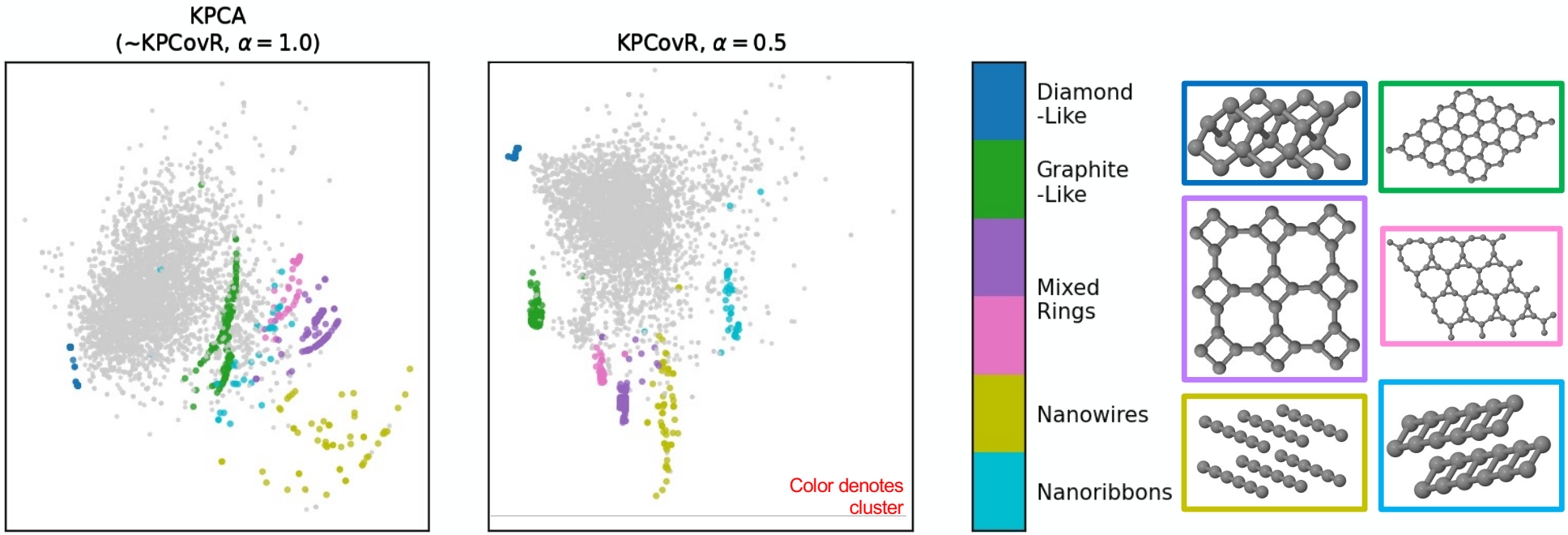


Inputs: SOAP features of 10,000 AIRSS carbon crystals
Target: energies in [eV / atom]
Kernel Parameters: RBF kernel, γ=$10^{-3.8}$
(1/1) train / test split

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-matter.readthedocs.io

Kernel Principal Covariates Regression

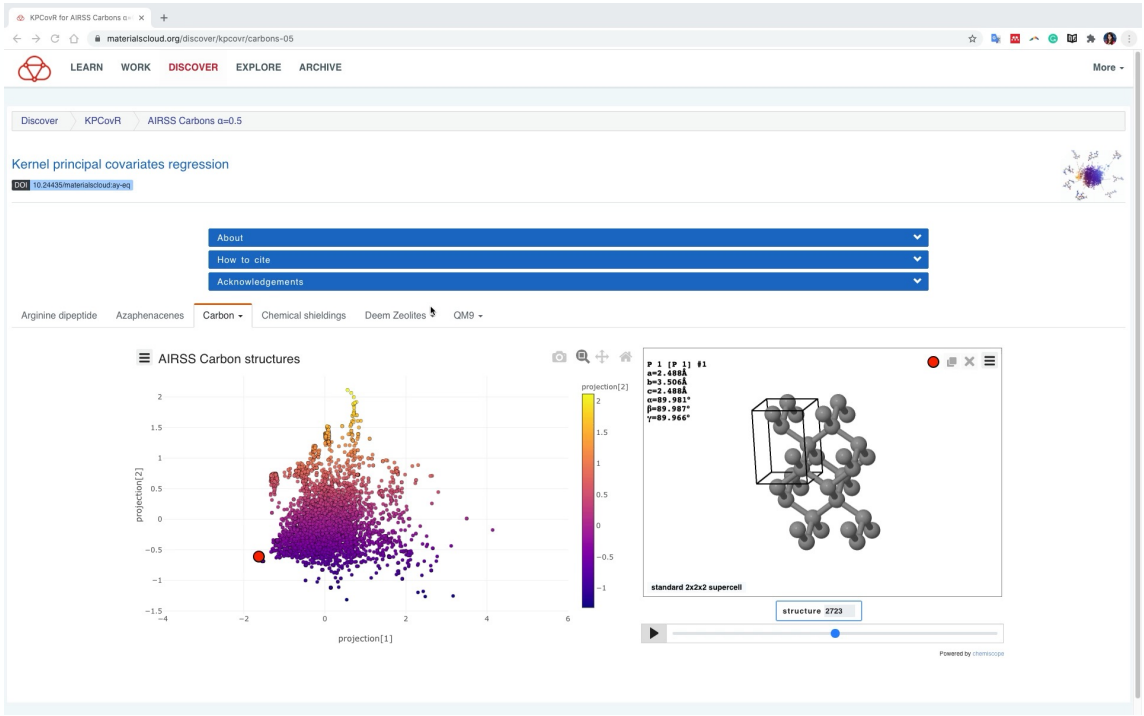
Determines a low-dimension projection from a similarity kernel, considering target data when constructing the projection.



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-matter.readthedocs.io

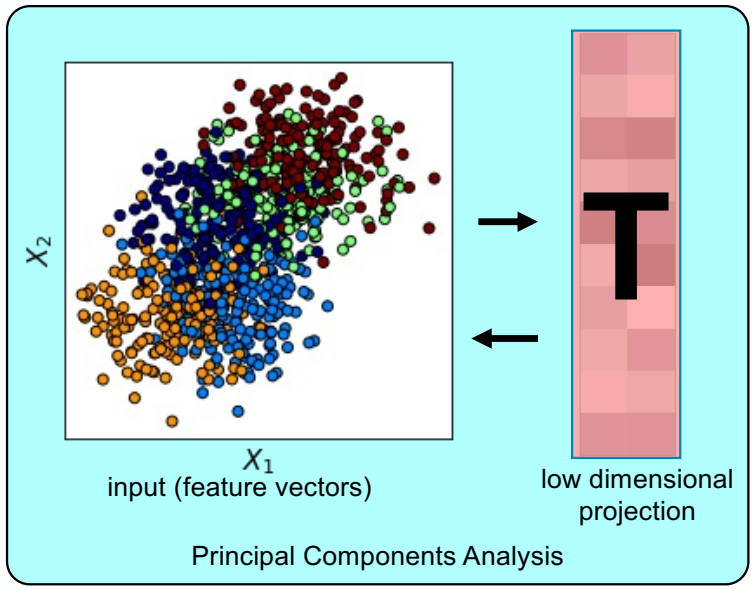
The dataset shown is discoverable via MaterialsCloud and *chemiscope*.

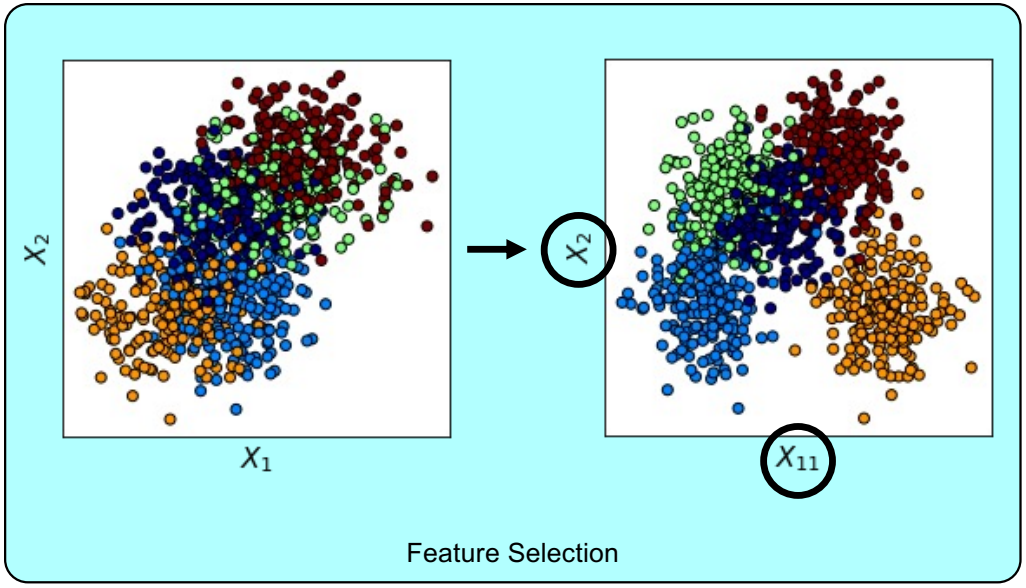


KPCovR chemiscope

What if the features carry inherent meaning?

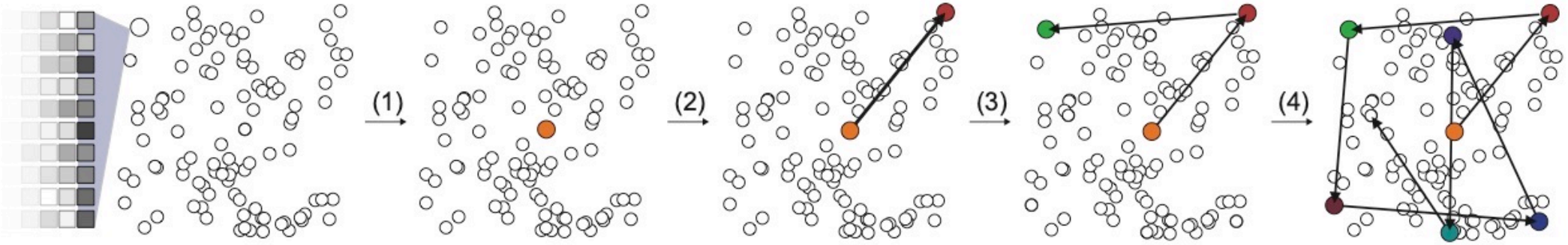
Many dimensionality reduction techniques construct a *new* set of features, but what if you want to just work with a subset of the old set?





Farthest Point Sampling (FPS)

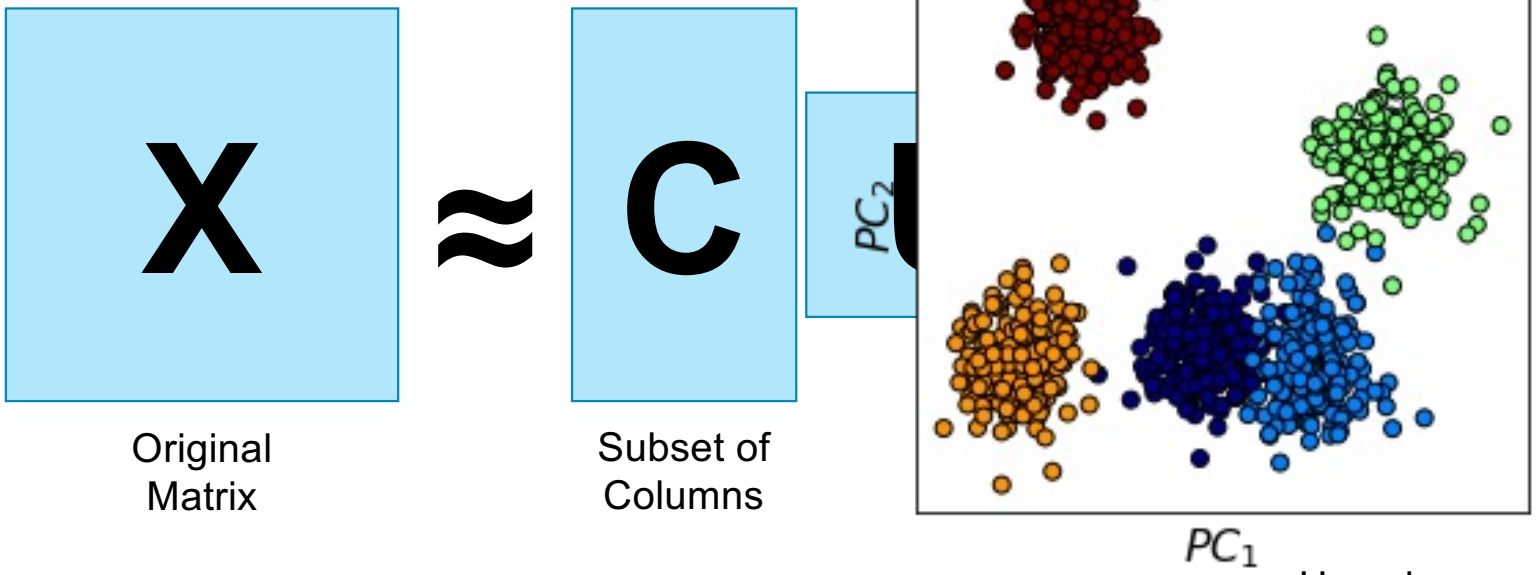
FPS aims to select a diverse subset of features or samples that cover the greatest portion of sample or feature space.



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



Feature selection based on CUR:

1. For each column, 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!

How do we calculate π?

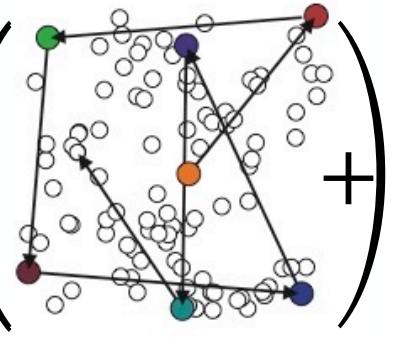
$PC_1 = AX_1 + BX_2 + CX_3 \ldots$

weights of our features in the PCA

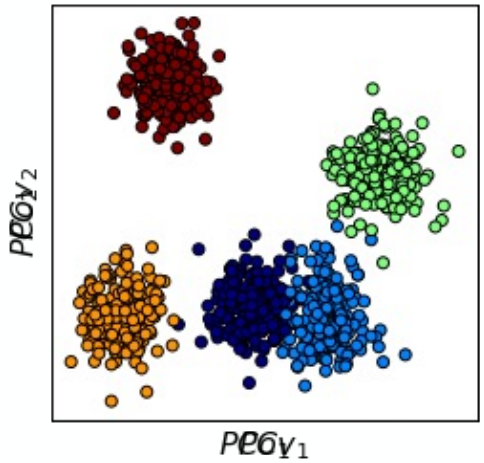
PCov-FPS and Pcov-CUR

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

Farthest Point Sampling (FPS)

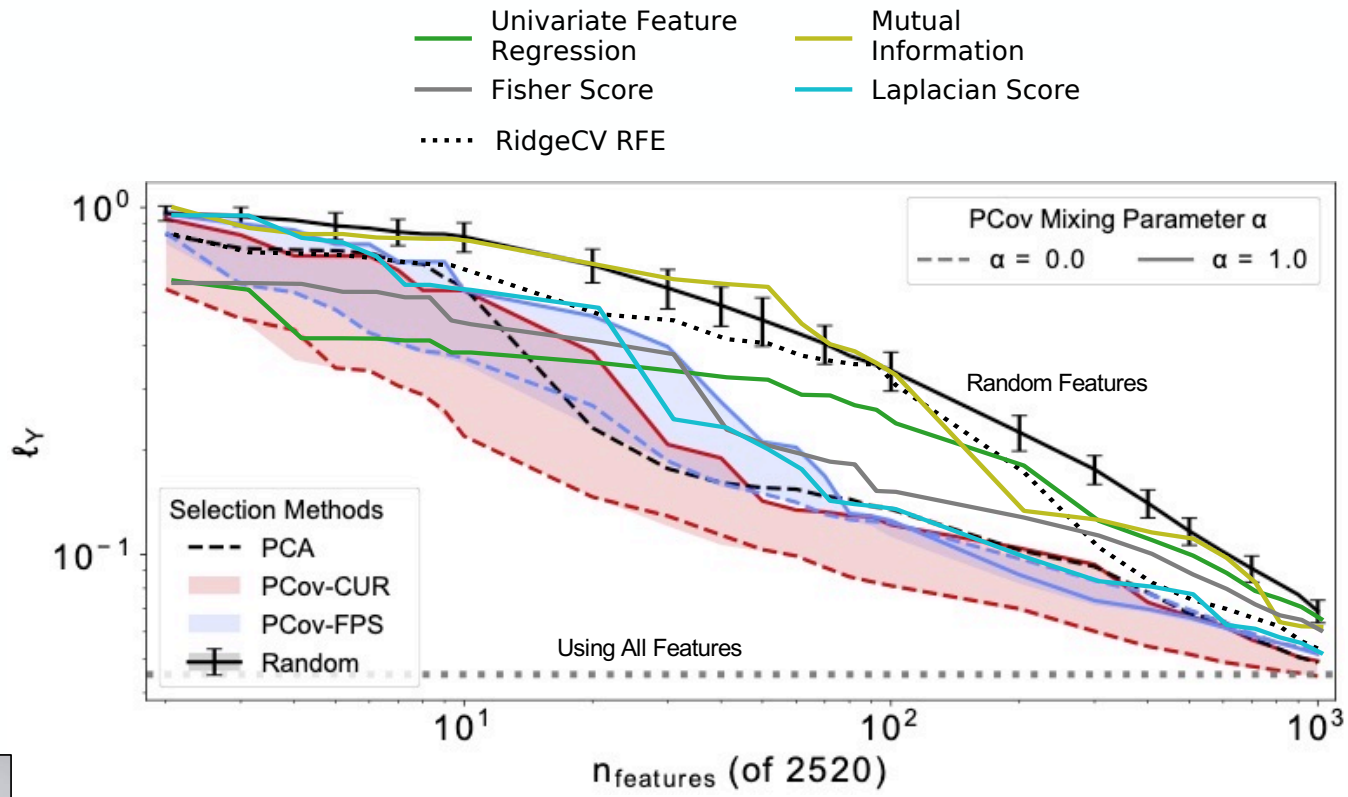
$$\mathbf{d} = f\left(\text{  } + \text{Contribution to Regression Weights} \right)$$

CUR Decomposition

$$\pi = f\left(\text{  } \right)$$

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

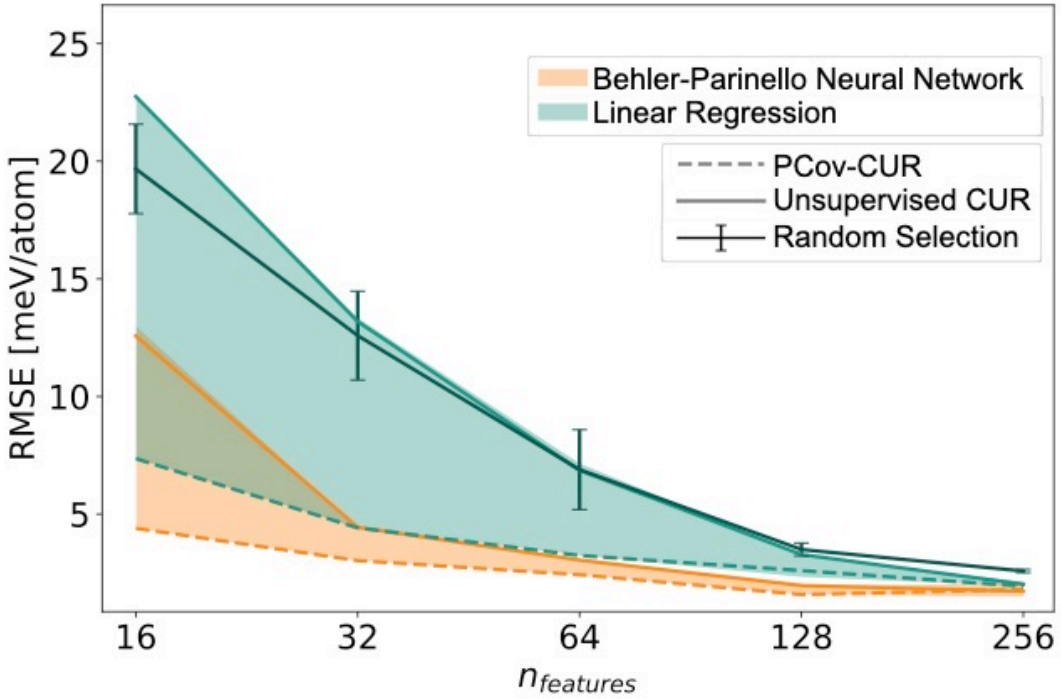
RKC, et al 2021 Mach. Learn.: Sci. Technol. 2 035038
scikit-matter.readthedocs.io

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.

A linear model with well-selected features can perform comparably to a NN with previous state-of-the-art-selected features.

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



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

We can access these functions using the open-source *scikit-matter*

pip install skmatter

```
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)
```

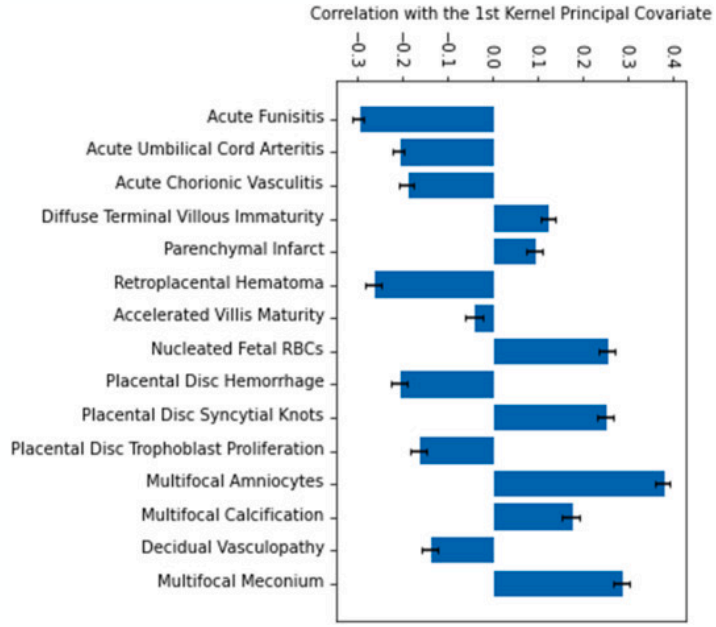
scikit-matter is a collection of scikit-learn compatible utilities that implement methods born out of the materials science and chemistry communities.

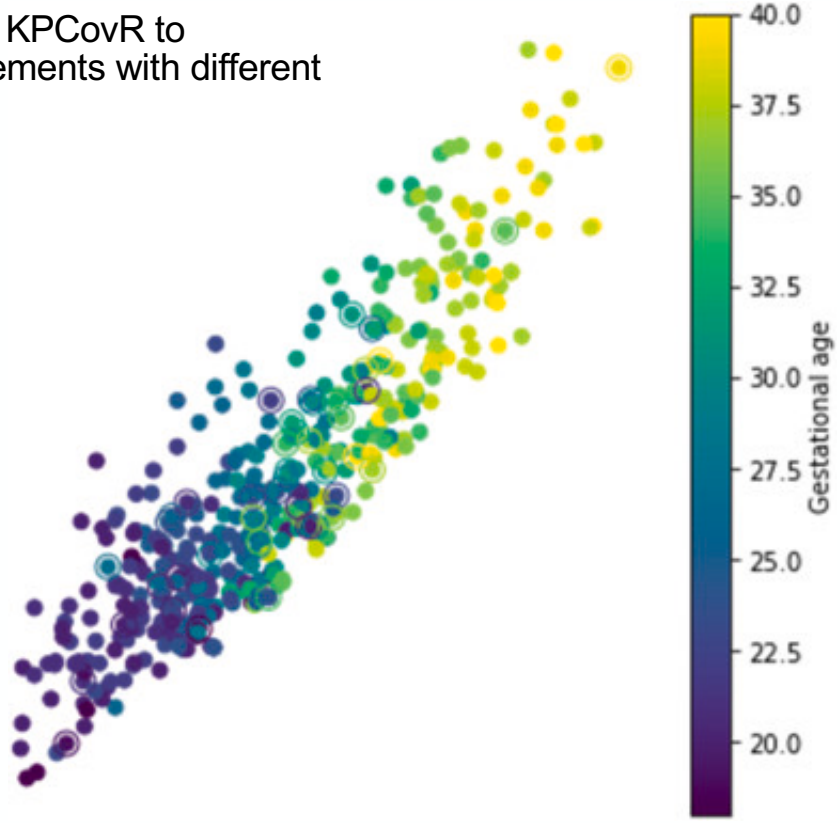
scikit-matter.readthedocs.io

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

Kernel Principal Covariates Regression can be useful beyond chemical contexts.

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

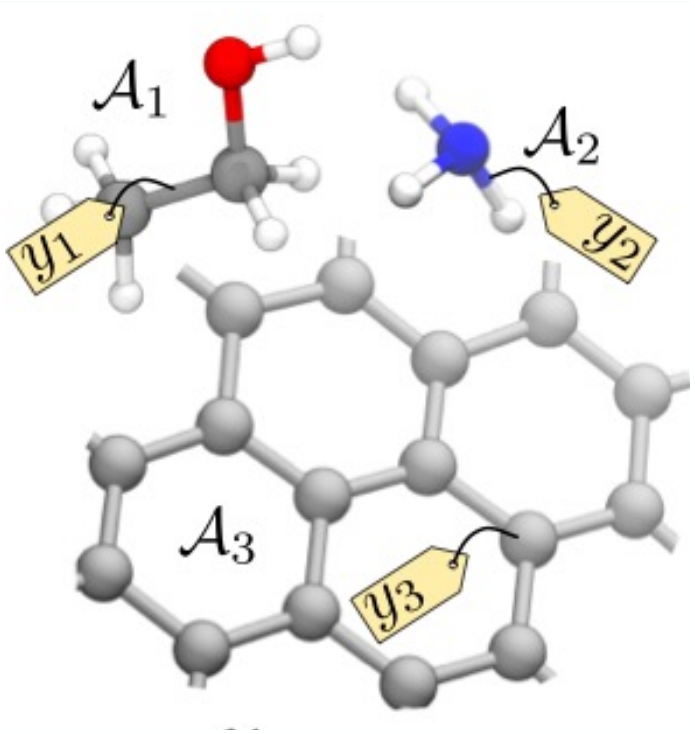


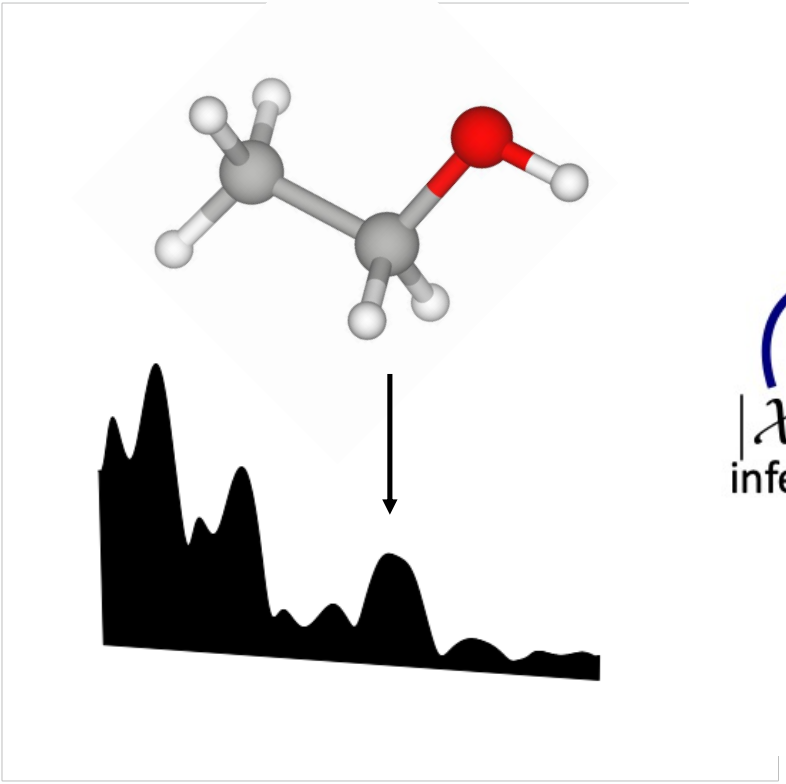


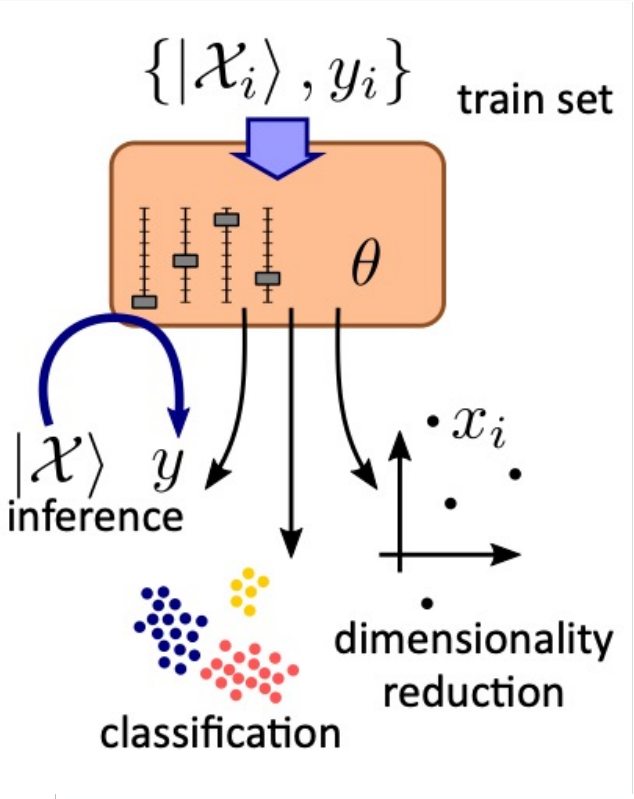
THE WARREN ALPERT Medical School

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, https://doi.org/10.1016/j.placenta.2023.04.005.

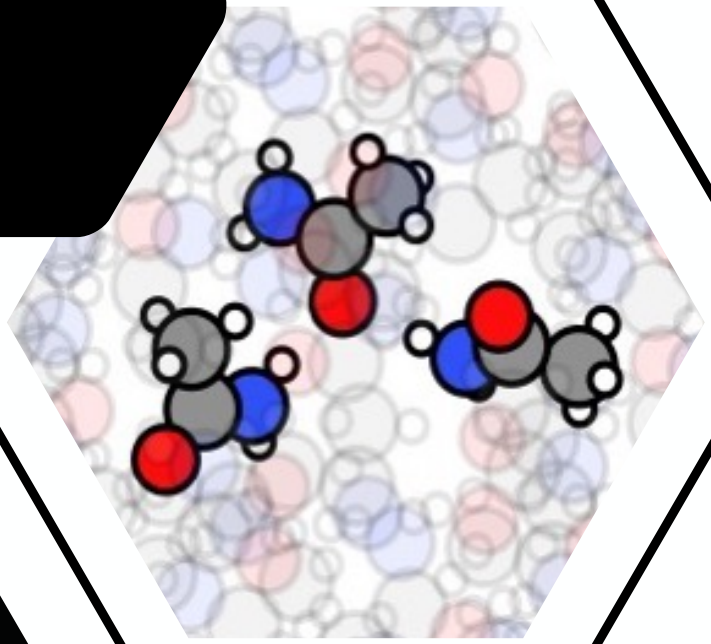
Chemical Data	Numerical Representation	Machine Learning Model



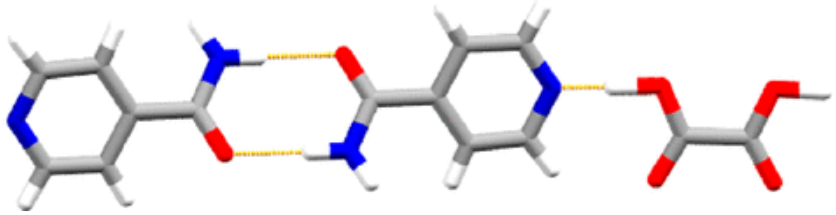




In a molecular crystal, what atoms or groups of atoms guide the crystallization process?

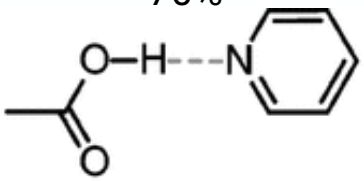


- In experiment, molecular packings are often rationalized by supramolecular synthons – libraries of intermolecular interactions common to molecular crystals.
- The hierarchies of these synthons is typically determined by mixing molecules with the synthons of interest and observing the interactions of the product.



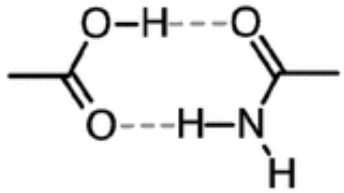
(isonicotinamide)·(oxalic acid) cocrystal
CSD ref. code: ULAWAF

70%



carboxylic-pyridine,
amide/amide

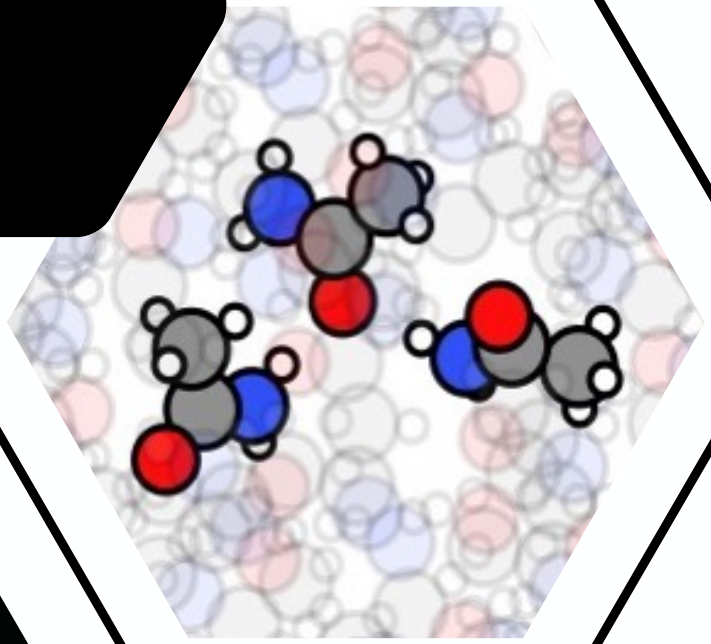
30%



carboxylic-amide

- Computationally, this is done by querying CSD and determining the prevalence of each synthon interaction.

In a molecular crystal, how do different atoms or groups of atoms contribute to the lattice energy?



Chemical Data

Maria Pahnova
Marvel INSPIRE Intern
Incoming UW-Madison PhD

CSD-10k*, ~10000 geometry-optimized (but weird) molecular crystals, pre-partitioned into training/testing sets

ShiftML2 (1.0.0) (2022). https://doi.org/10.5281/zenodo.7097427

↓

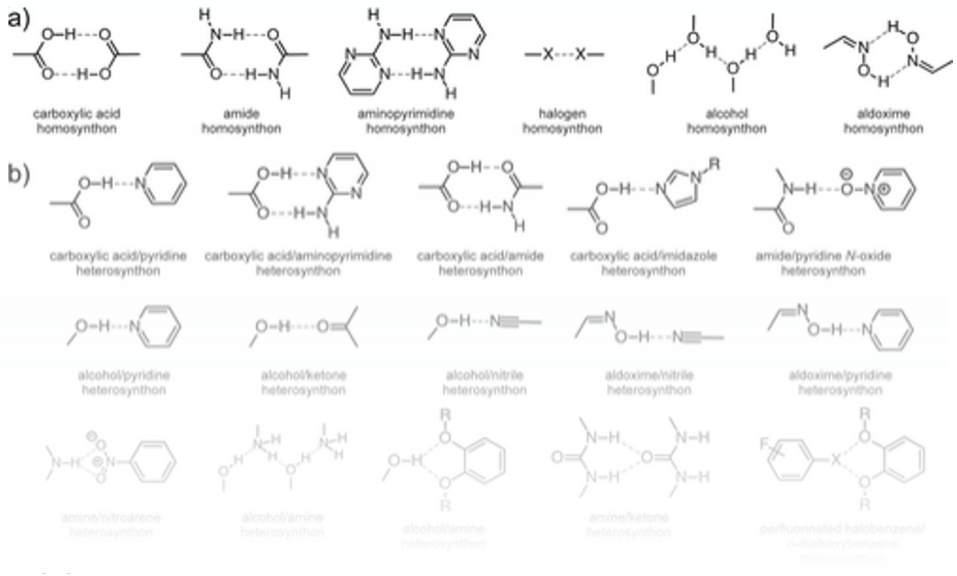
Select only C, H, N, O, S

↓

Remove crystals where binding is non-trivially-defined (polymers, charged/zwitterionic molecules, etc.)

↓

Determine relaxed geometries and energies of molecules in order to determine lattice energy

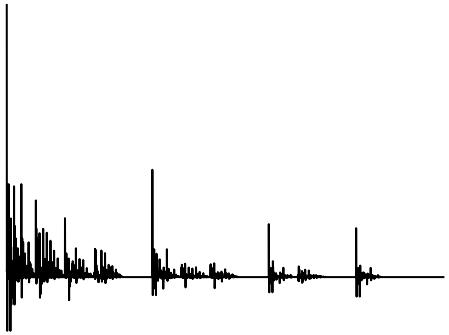


R.K. Cersonsky, et al, *Lattice energies and relaxed geometries for 2'707 organic molecular crystals and their 3'242 molecular components.*, Materials Cloud Archive **2023.5** (2023), doi: 10.24435/materialscloud:71-21. Visualization at: https://molmotifs.matcloud.xyz/

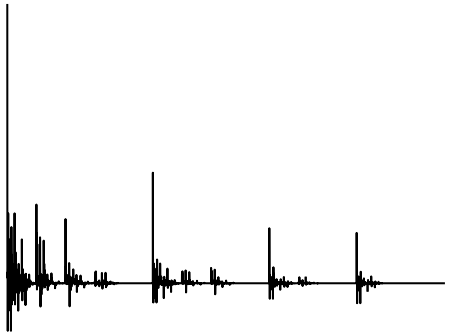
SOAP hyperparameters: n=6, l=4, interaction cutoff (7.0Å) with radial scaling, gaussian width (0.3Å)
Regression hyperparameters: 2707 train/551 test, sklearn.RidgeCV(cv=5, alphas=np.logspace(-12,-3,20))

Numerical Representation

Representation	Regression Correlation Coefficient (R^2)	RMSE (kJ/mol)
Crystal Environments	0.86	0.778



the crystalline environments -

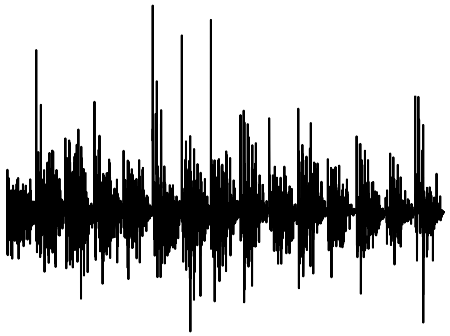


the gas-phase environments

= the ”remnant” environments

these results hold for implicit feature spaces as well, but for that, you need to read the paper:
R. K. Cersonsky, et al., 2023 *Chem. Sci.* **14**, 1272–1285.

By using an additive descriptor, we can estimate the contribution of each motif to the cohesive energy.



the “remnant” fingerprint averaged over the crystal

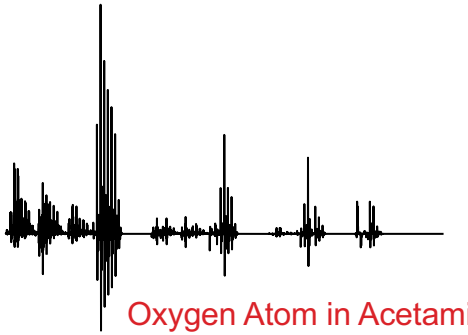
×

weights

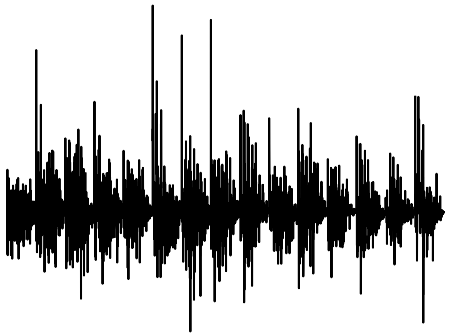
→

lattice energy of the crystal

By using an additive descriptor, we can estimate the contribution of each motif to the cohesive energy.

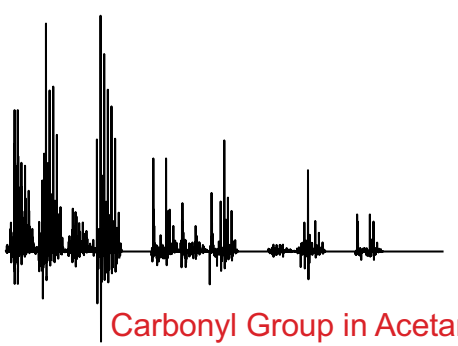


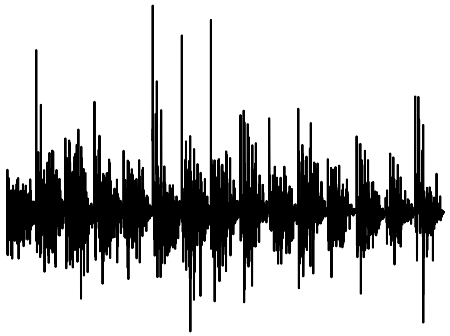
Oxygen Atom in Acetamide



the “remnant” fingerprint of **each atom** × weights → lattice energy **contribution of the atom**

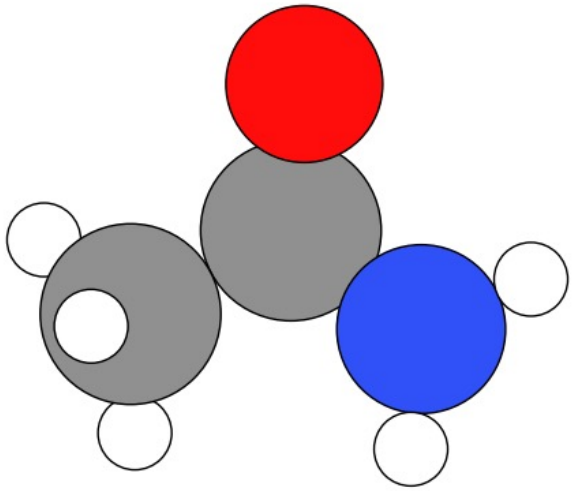
By using an additive descriptor, we can estimate the contribution of each motif to the cohesive energy.



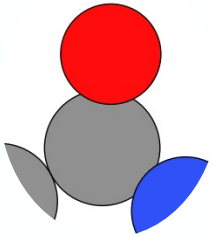


the “remnant” fingerprint **averaged over a collection of atoms** × weights → lattice energy **contribution of the collection**

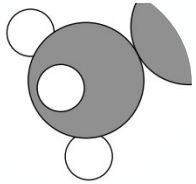
We used SMARTS string to automate labeling 48 popular molecular subgroups that are found in supramolecular synthons or tectons, resulting in approximately 70,000 motifs.



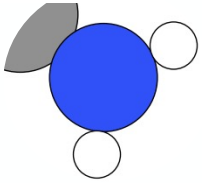
Canonical SMILES: CC(=O)N



Carbonyl
[C]=[O;X1]

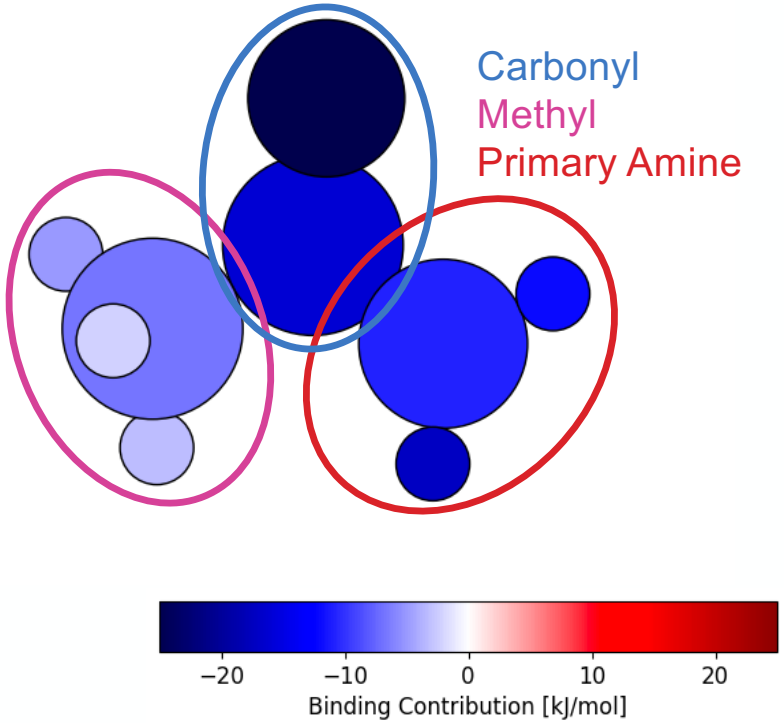


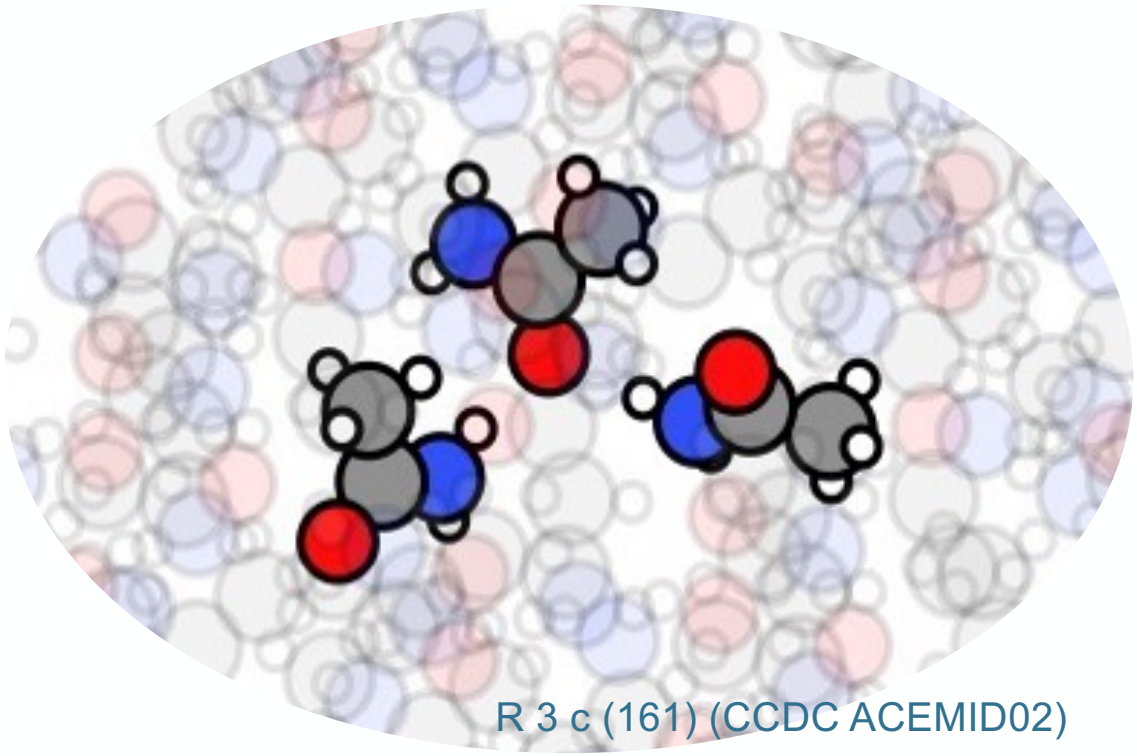
Methyl
[C;H3;X4]



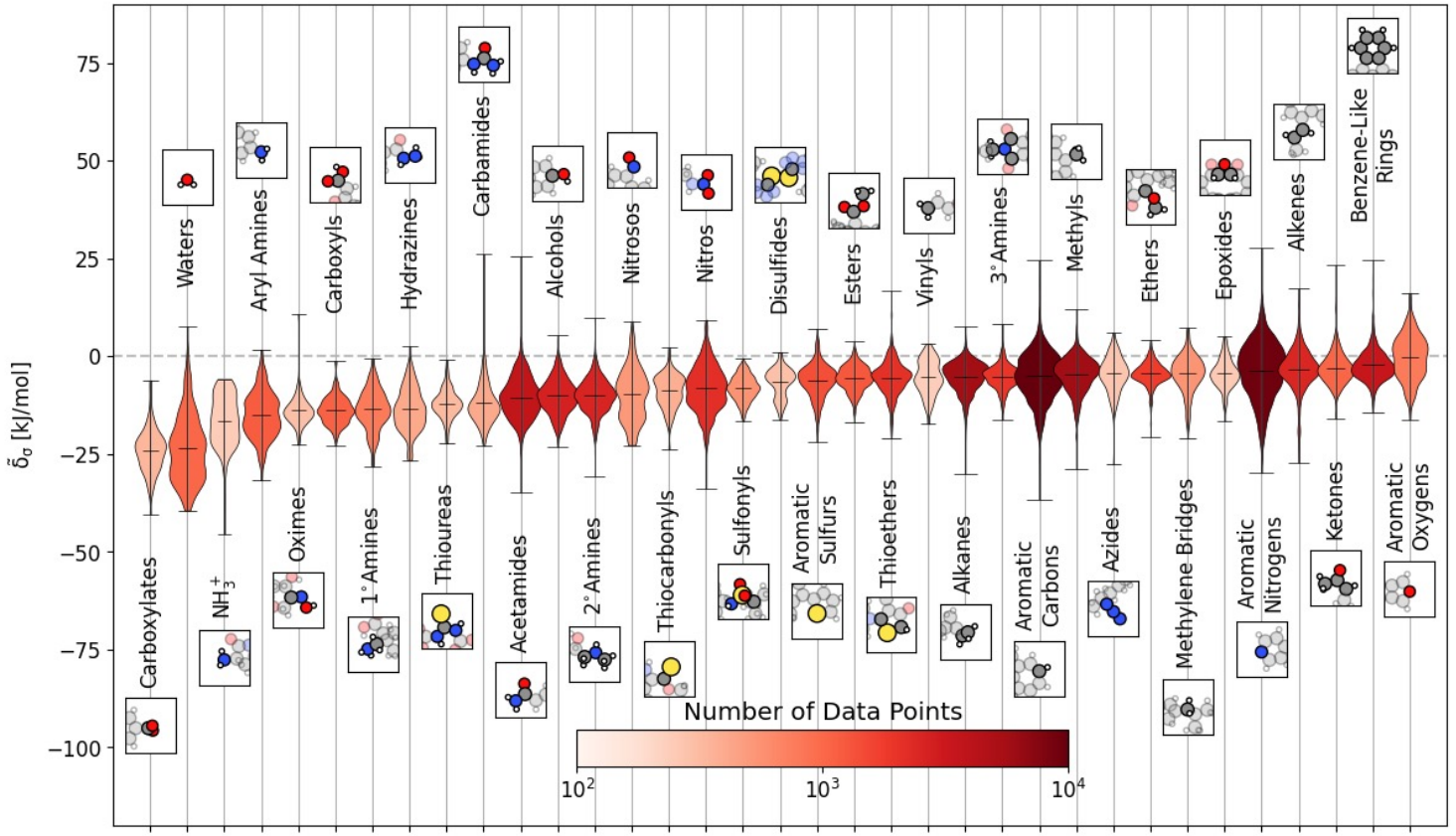
Primary Amine
[N;X3;H2]

With these categorizations, we can see the contribution of each sub-group.

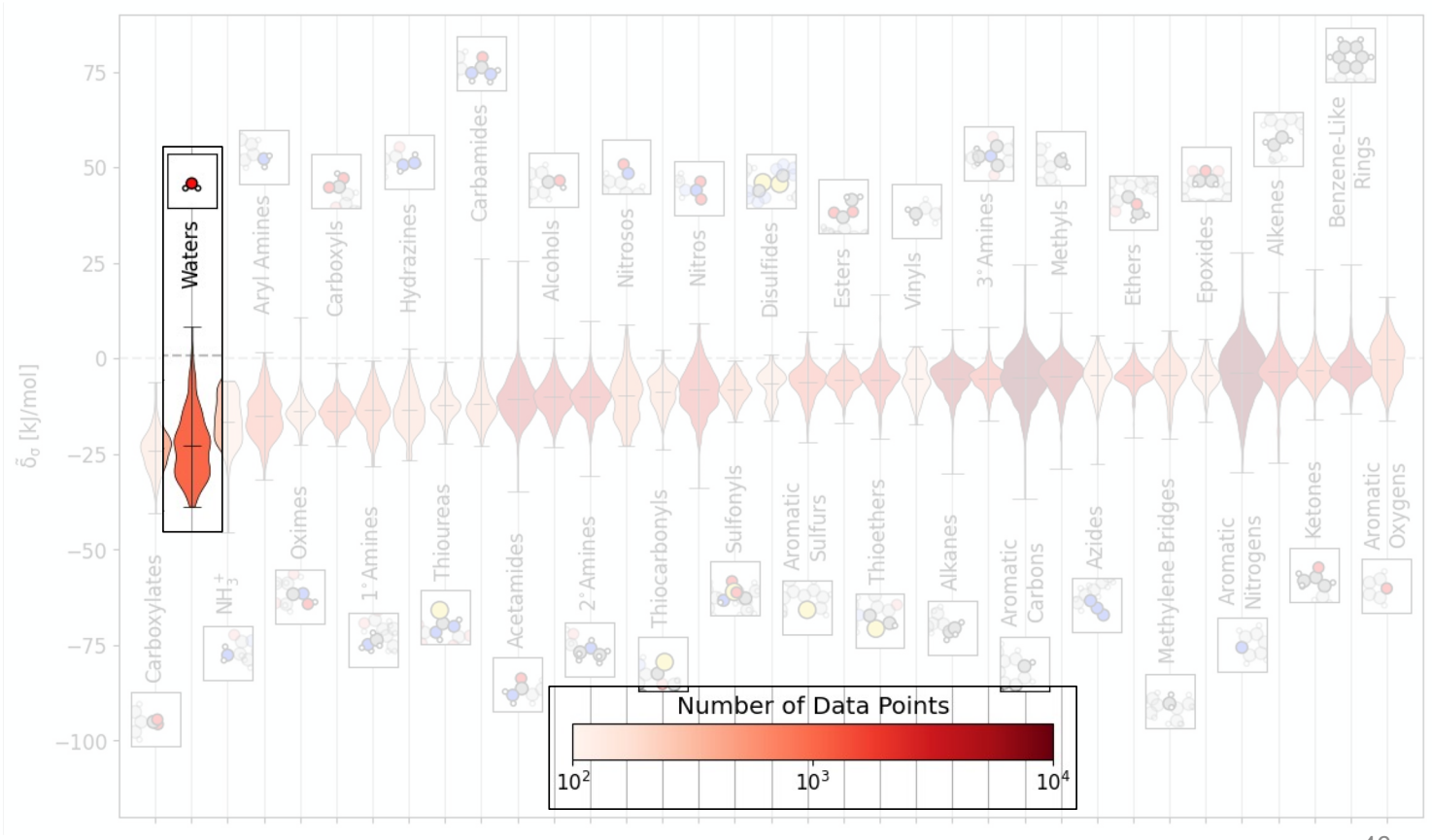




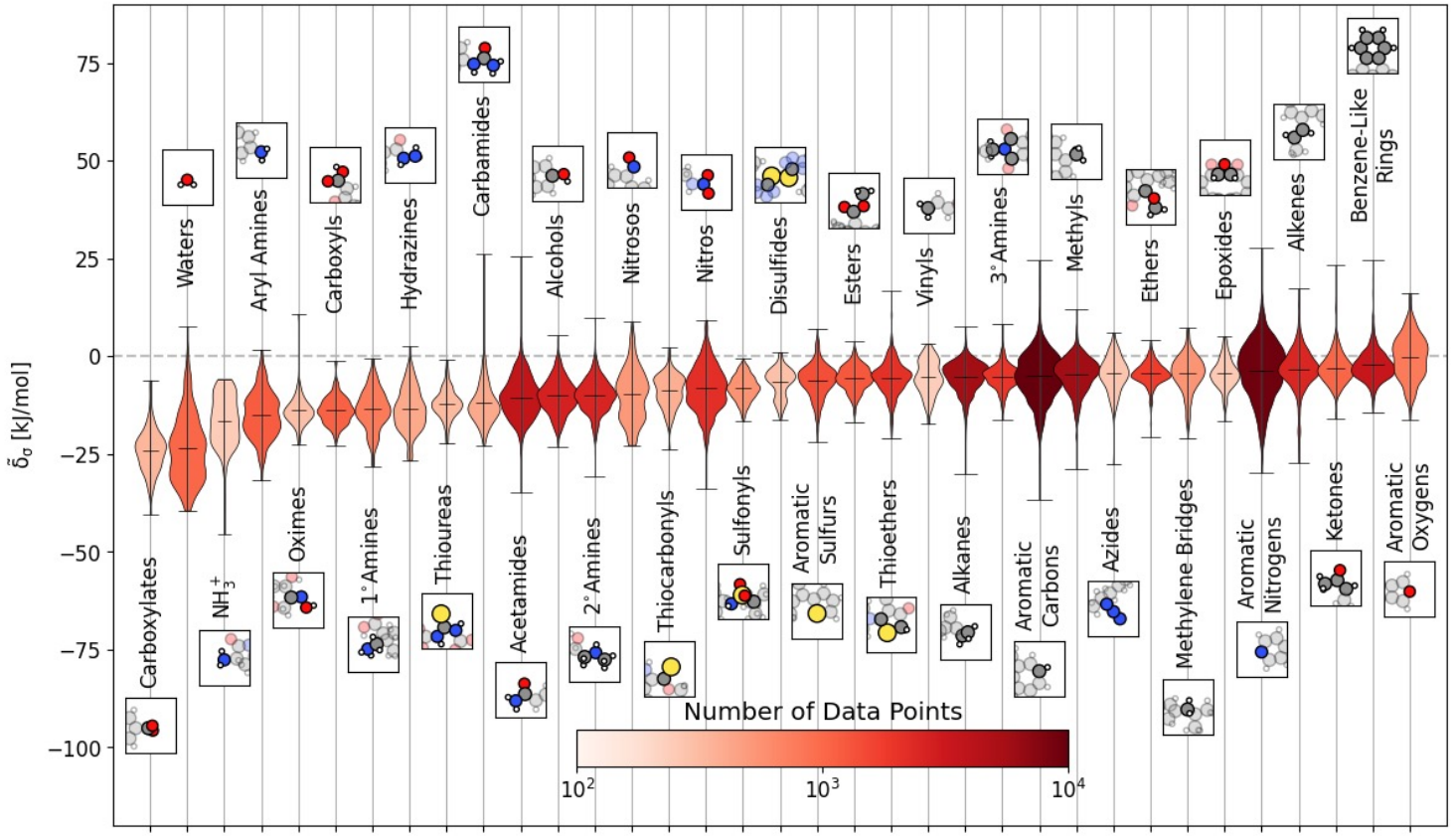
Each different subgroup results in a range of contributions…



Each different subgroup results in a range of contributions…

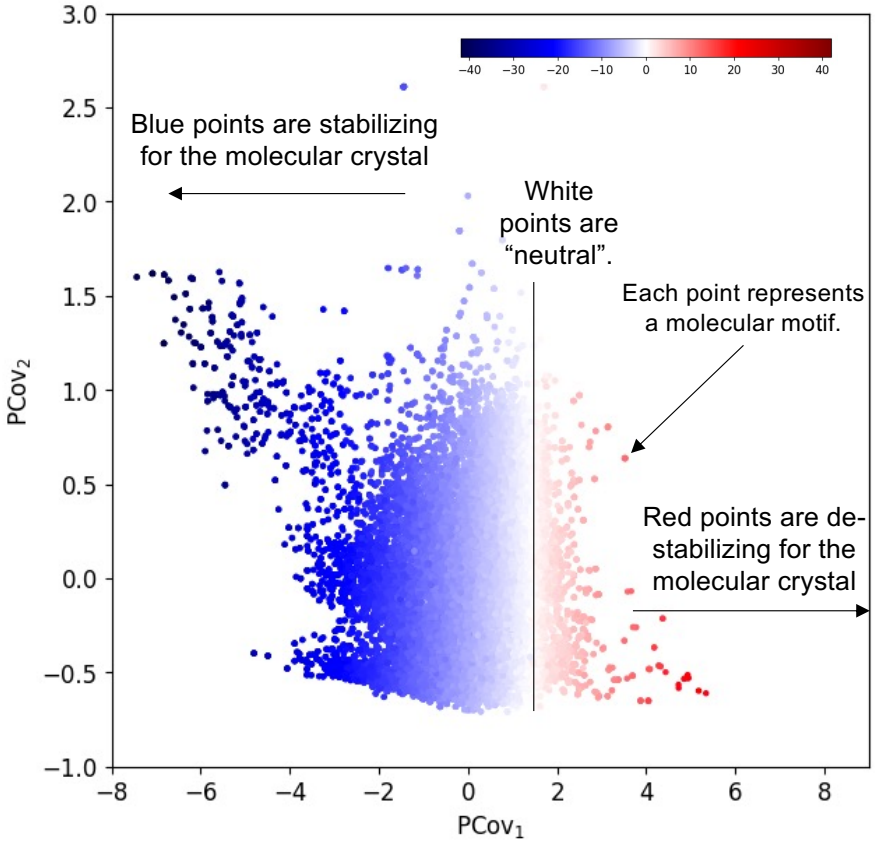


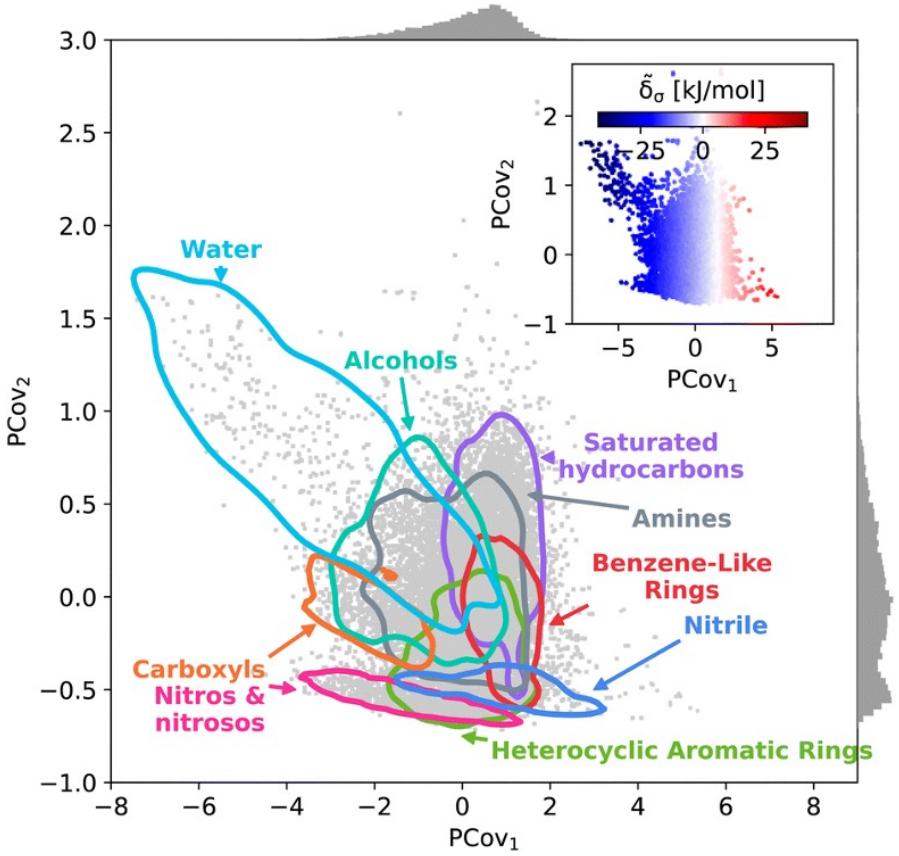
Each different subgroup results in a range of contributions…**but how do we interpret these results?**



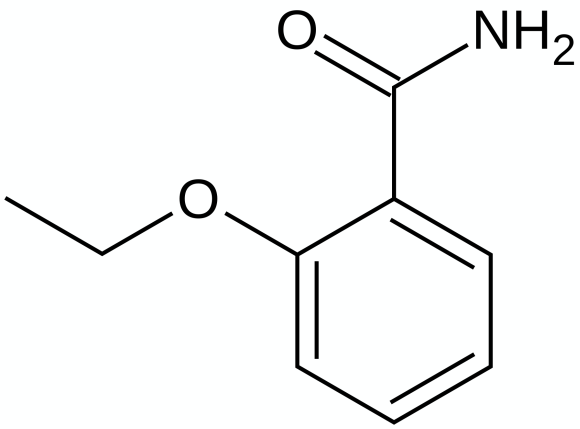
Model

We use a PCovR mapping to understand the similarities of intermolecular interactions.

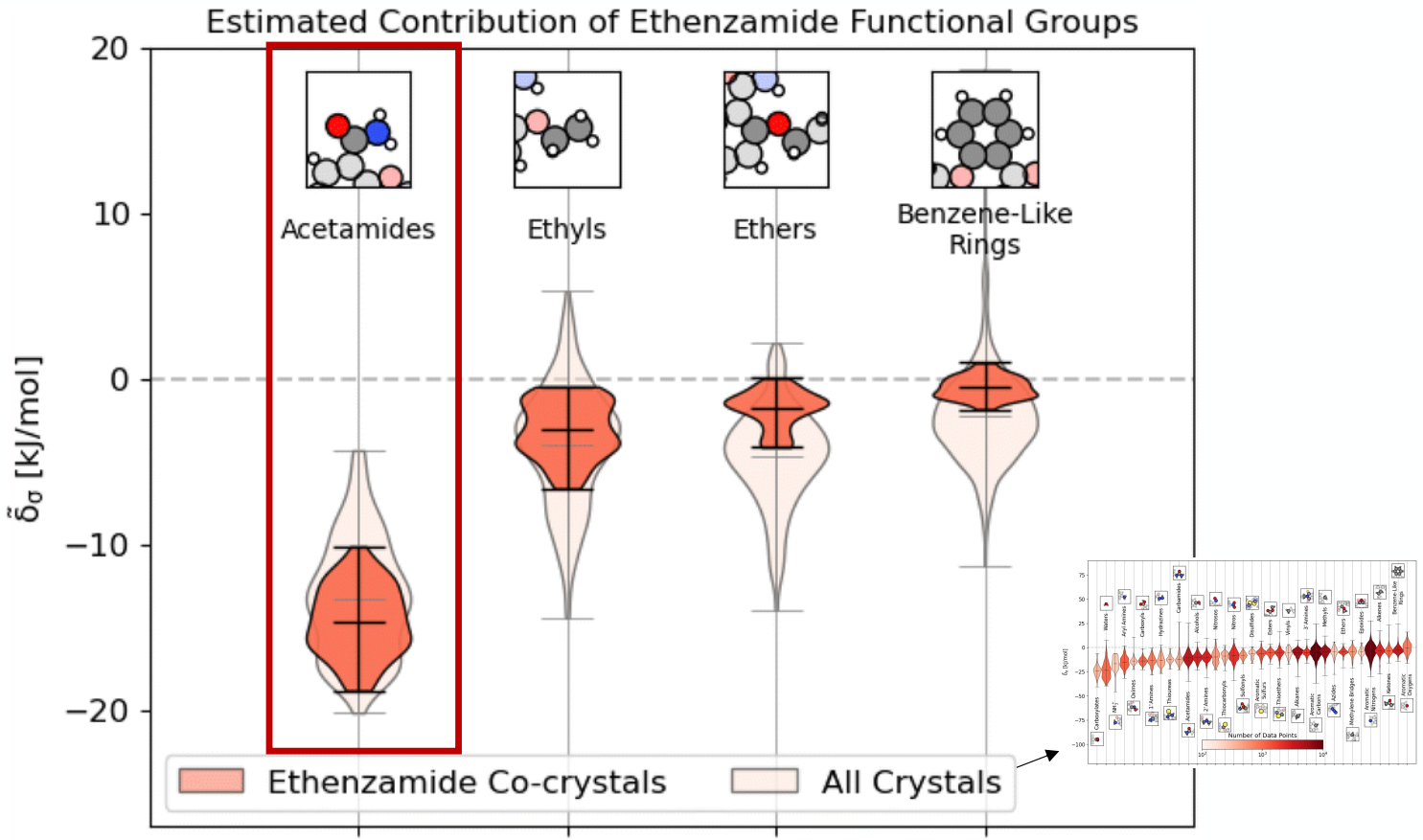




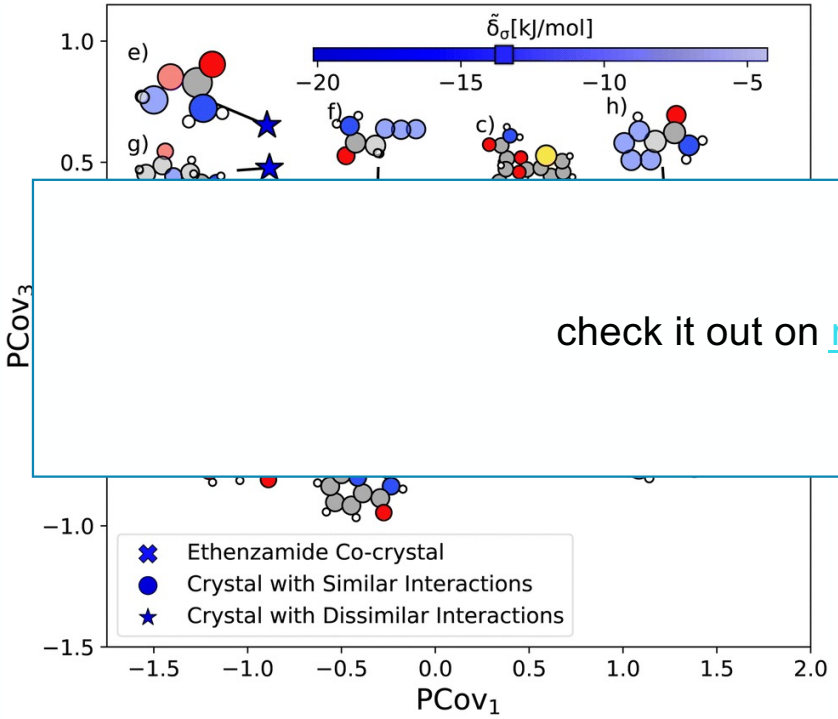
When we compare the range of interactions for a given molecule we want to co-crystal, we can see the engineerable range of interactions.



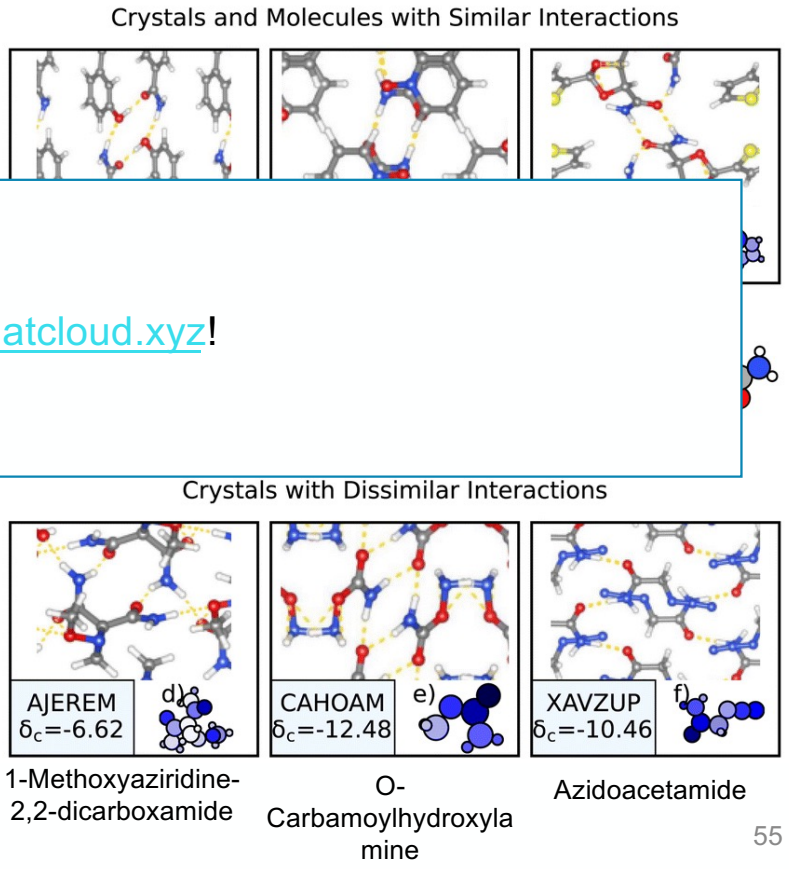
Ethenzamide,
a common analgesic



For example, we can choose new ethenzamide co-crystal formers by looking at interactions very similar or disimilar to those in the ethenzamide dataset.



check it out on molmotifs.matcloud.xyz!



BAH, **RKC**, GF, MC. *Mach. Learn.: Sci. Technol. 1 045021.*
GF, **RKC,** MC. (2020). Journal of Open Source Software, 5(51), 2117

Using ML methods engineered for interpretability, we can disentangle the structure-property paradigm, even in complex molecular systems.

Methods:

Mapping techniques for structure-property mappings: B. A. Helfrecht, **RKC**, et al. 2020 Mach. Learn.: Sci. Technol.1 045021.

Feature subselection: **RKC**, et al. 2021 Mach. Learn.: Sci. Technol. 2 035038.

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

Software:

`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 and Data:

R. K. Cersonsky, et al., 2023 *Chem. Sci.* **14**, 1272–1285.

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

R.K. Cersonsky, et al, *Lattice energies and relaxed geometries for 2'707 organic molecular crystals and their 3'242 molecular components.*, Materials Cloud Archive **2023.5** (2023), doi: 10.24435/materialscloud:71-21. Visualization at: https://molmotifs.matcloud.xyz/

If current trends do not change, fields such as chemical engineering and materials science will not reach gender parity any time soon. 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. Forthcoming in npj Computational Materials in ~1 week.

My group is hiring PhDs and postdocs!

In the Cersonsky Lab, our ultimate goal is to build a unified machine learning feature space and methodology for studying the thermodynamical behavior of multiscale and hierarchical materials.

rose.cersonsky@wisc.edu