CINE
CENTER FOR INNOVATION
ON NEW ENERGIES

Supervised Clustering for Selecting Representative Samples in Chemical Databases

Felipe V. Calderan

Advisors:
Marcos G. Quiles
Juarez L. F. Da Silva

Computational Materials Science & Chemistry

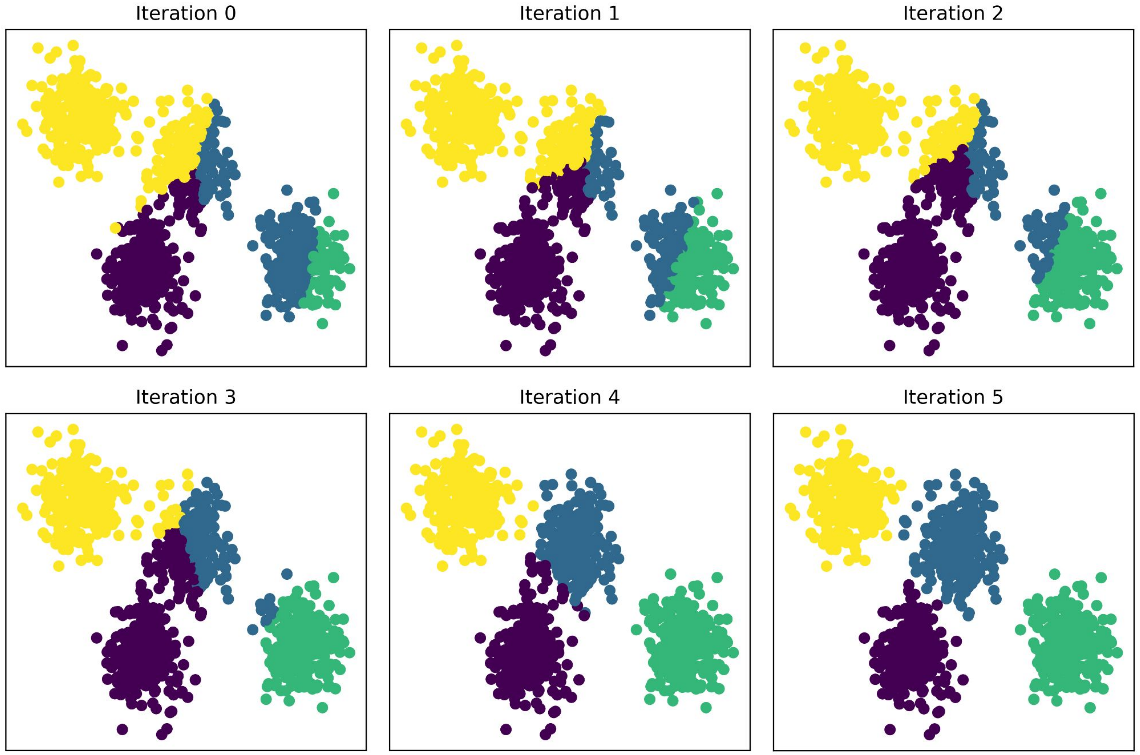
Objectives

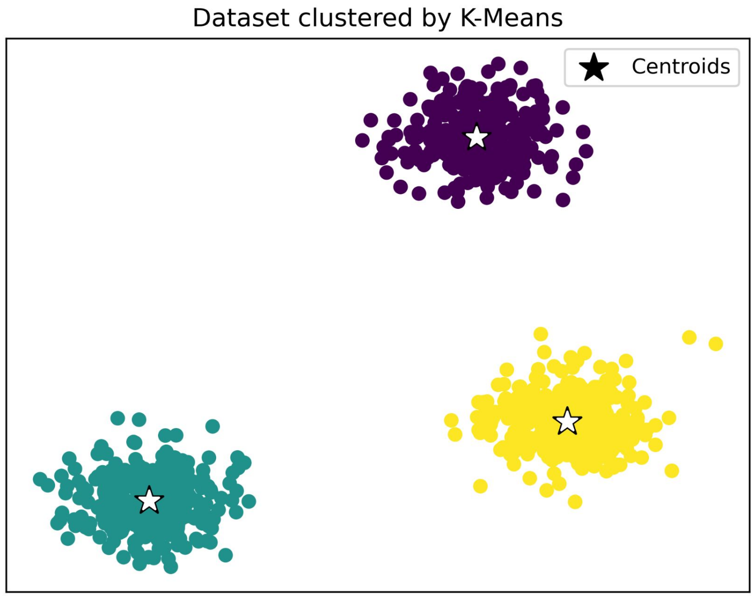
Provide a toolbox for faster material screening, so experts have access to large-scale material property analysis.

This can be achieved through the use of supervised clustering to obtain representative samples that will be analyzed rather than the complete dataset.

K-Means

K-Means is one of the most classic clustering methods in the literature.





Pseudo-code:

1. Choose k centroids to match k random elements from the database

2. Assign each element to the nearest centroid

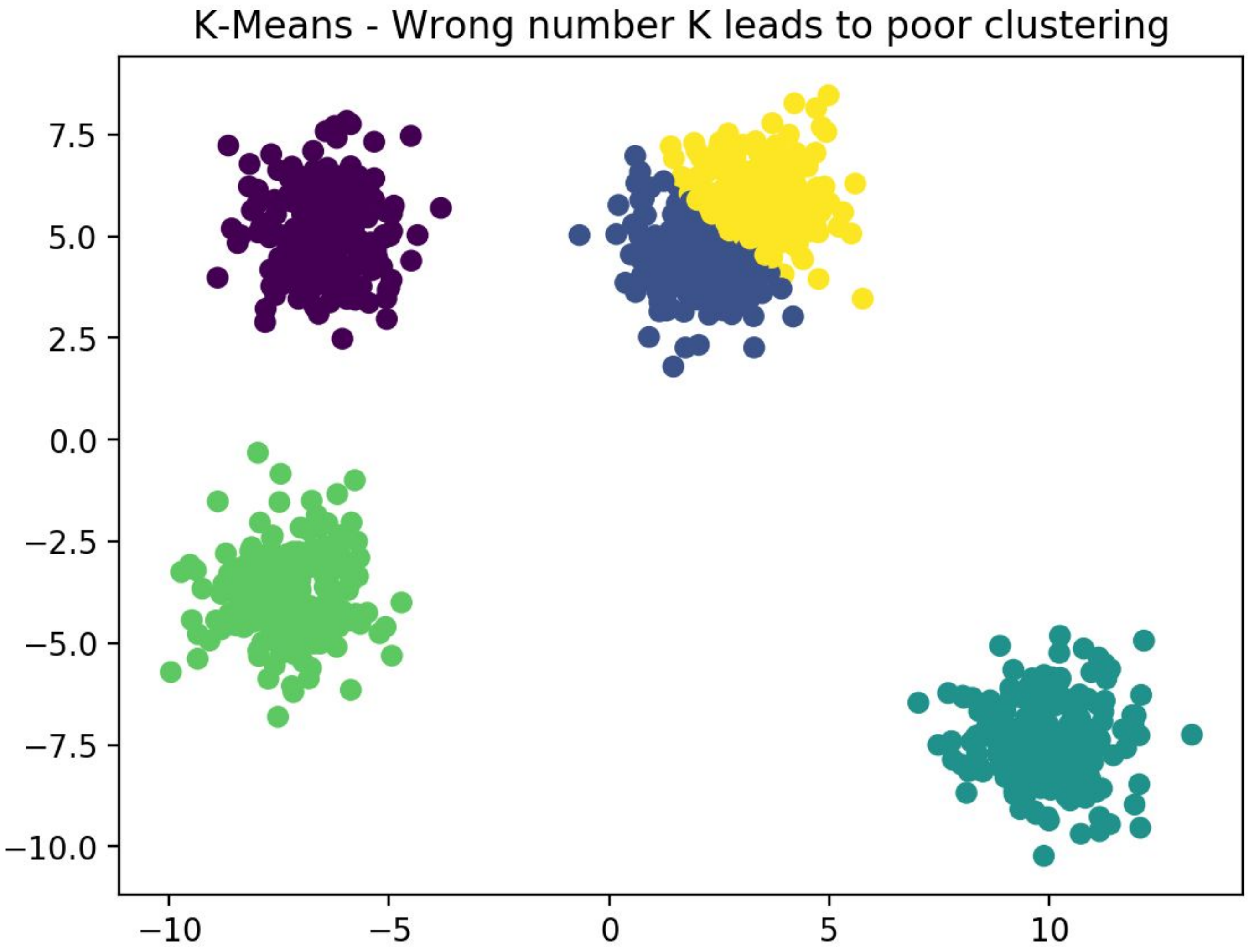
3. Recalculate the centroid of each cluster as the center of mass of its members

4. While the convergence criterion isn't met, repeat from step 2

Problems Tackled

- How many clusters should be found (what is the value of **K**)?

- What happens if some features are more important than others for my specific needs?



Supervised Clustering

K-Means is supervised by 2 other optimization algorithms.

Optimize number K of clusters with brute-force and a quality measure

→

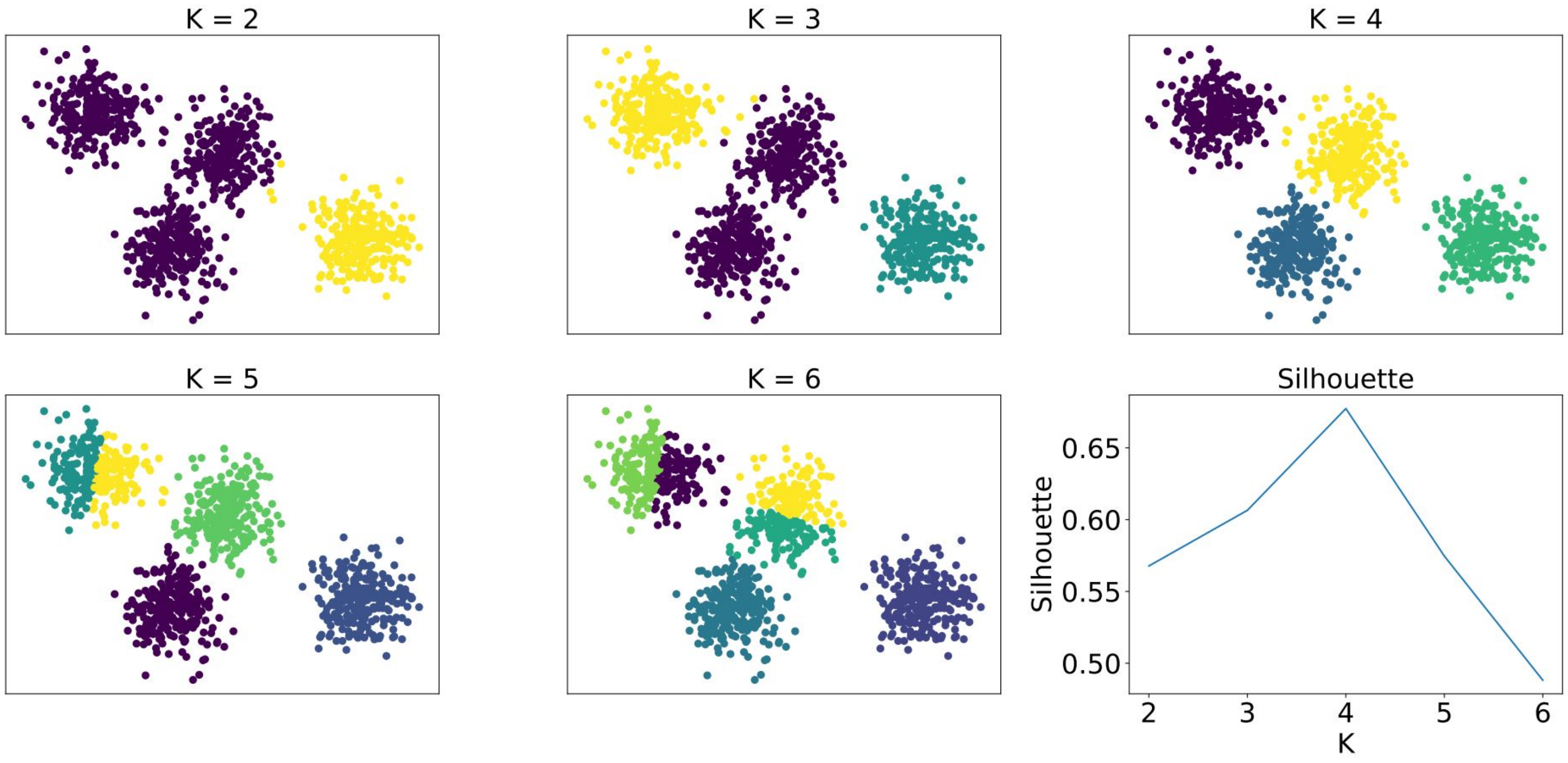
Minimize sum of intra-cluster variances for a given property with Basin-Hopping

→

Result: Good clustering quality and property highlighted

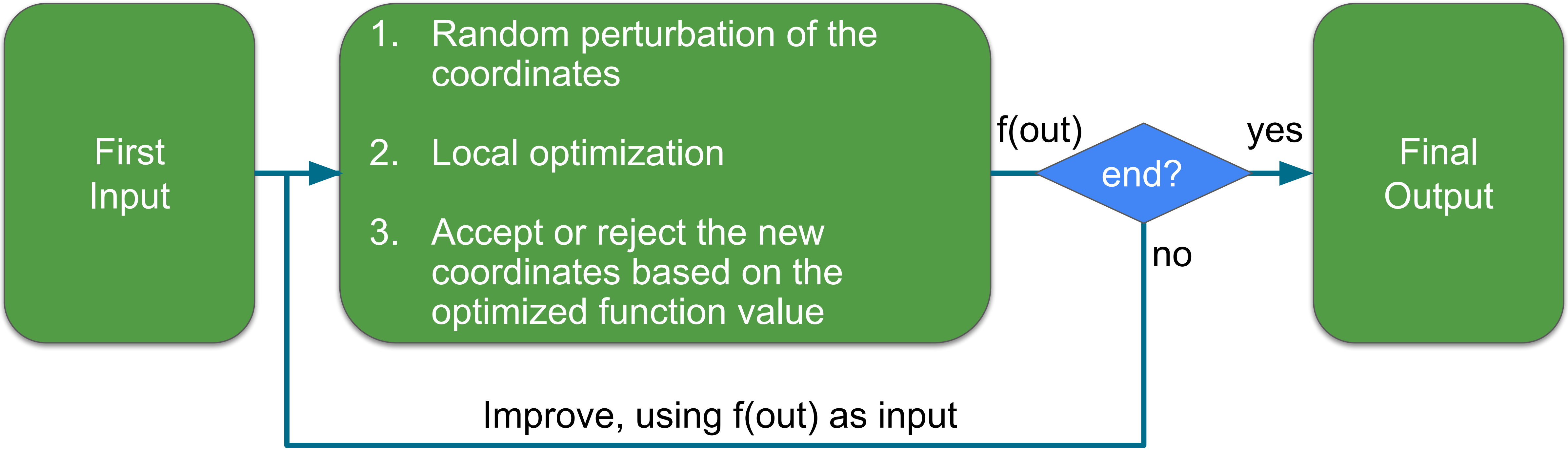
Brute-force Optimization

Objective: find the best K according to a clustering quality measure.



Basin-Hopping Optimization

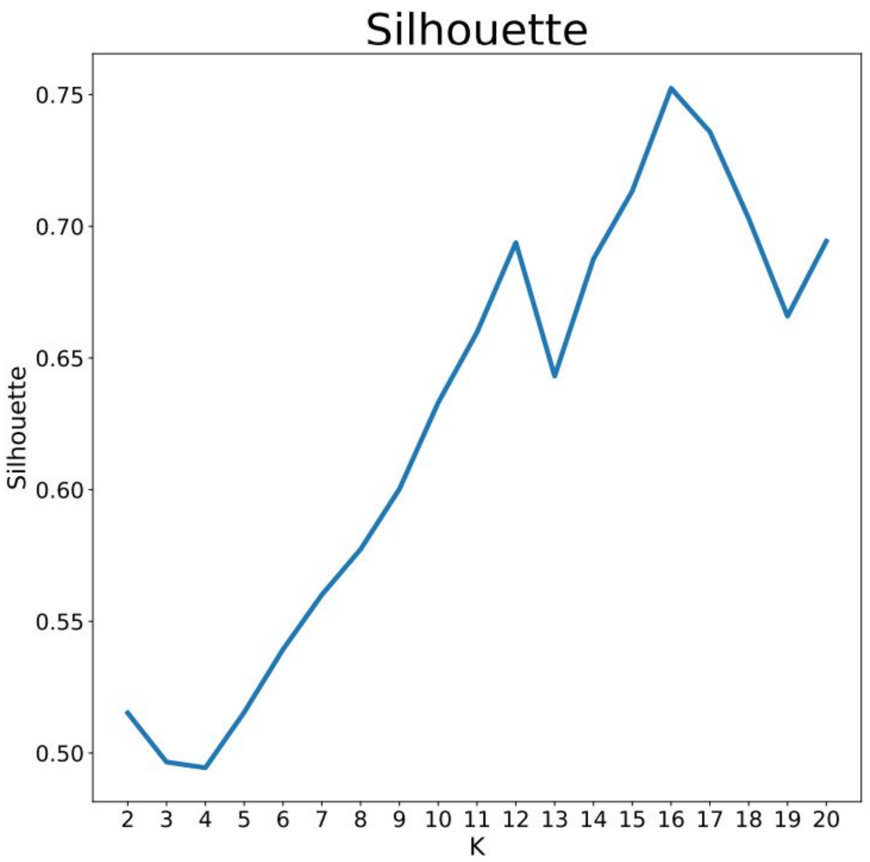
Objective: find the clustering configuration that best optimizes a criterion



Results

$CeZrO_4$

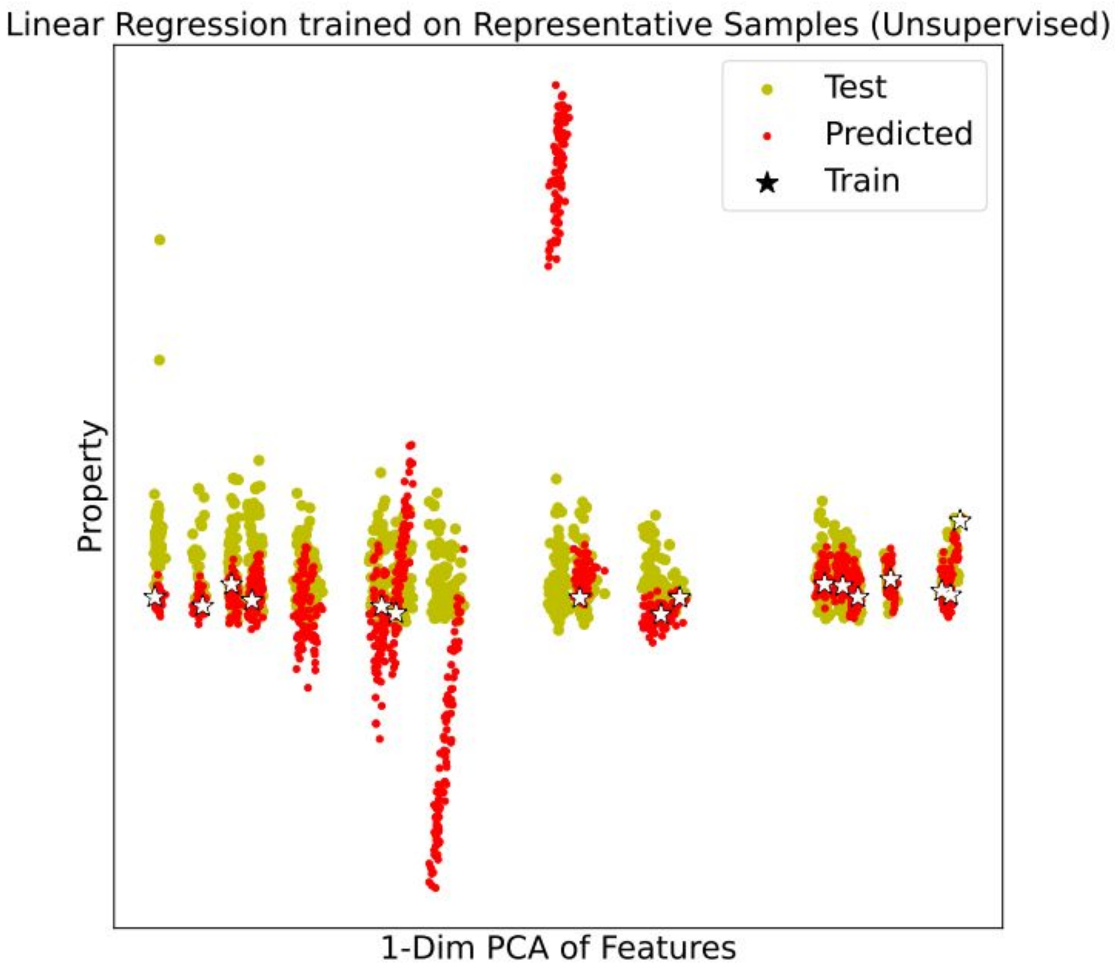
Description	Value
Best number K of groups/representative samples	16
Sum of intra-cluster variances before optimization	17.028679
Sum of intra-cluster variances after optimization	16.835181
Linear Regression MSE before optimization	24.032095
Linear Regression MSE after optimization	14.270303

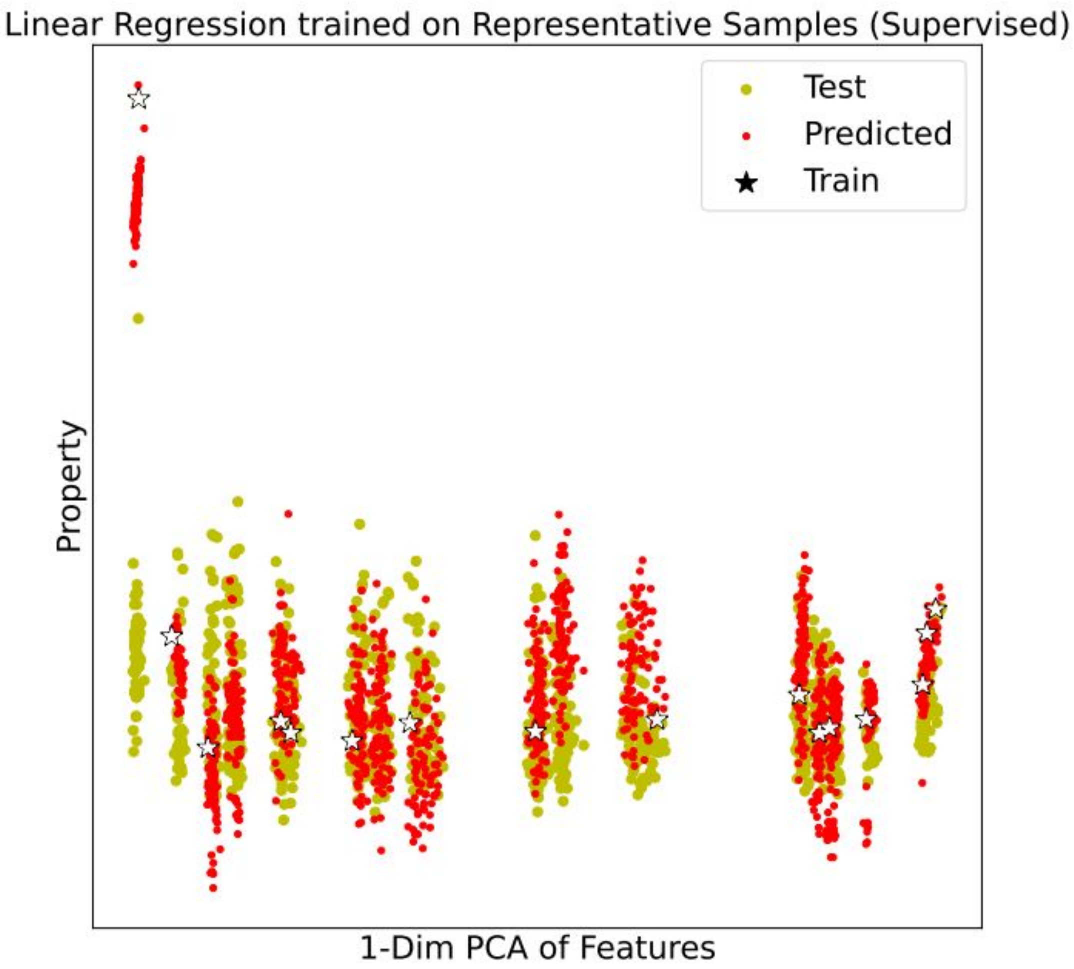


[Dataset] FELÍCIO-SOUSA, P. et al. Ab initio insights into the structural, energetic, electronic, and stability properties of mixed cenzr15no30 nanoclusters. Phys. Chem. Chem. Phys., The Royal Society of Chemistry, v. 21, p. 26637–26646, 2019.

Results

$CeZrO_4$

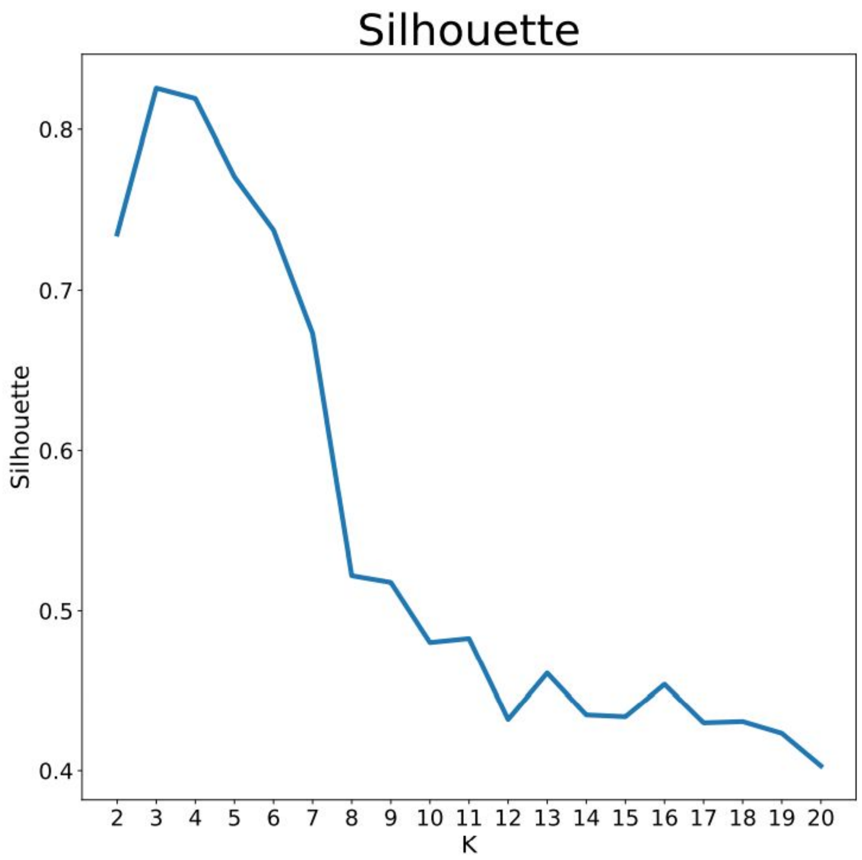




Results

55-Atom Pt-Based Core–Shell Nanoalloys

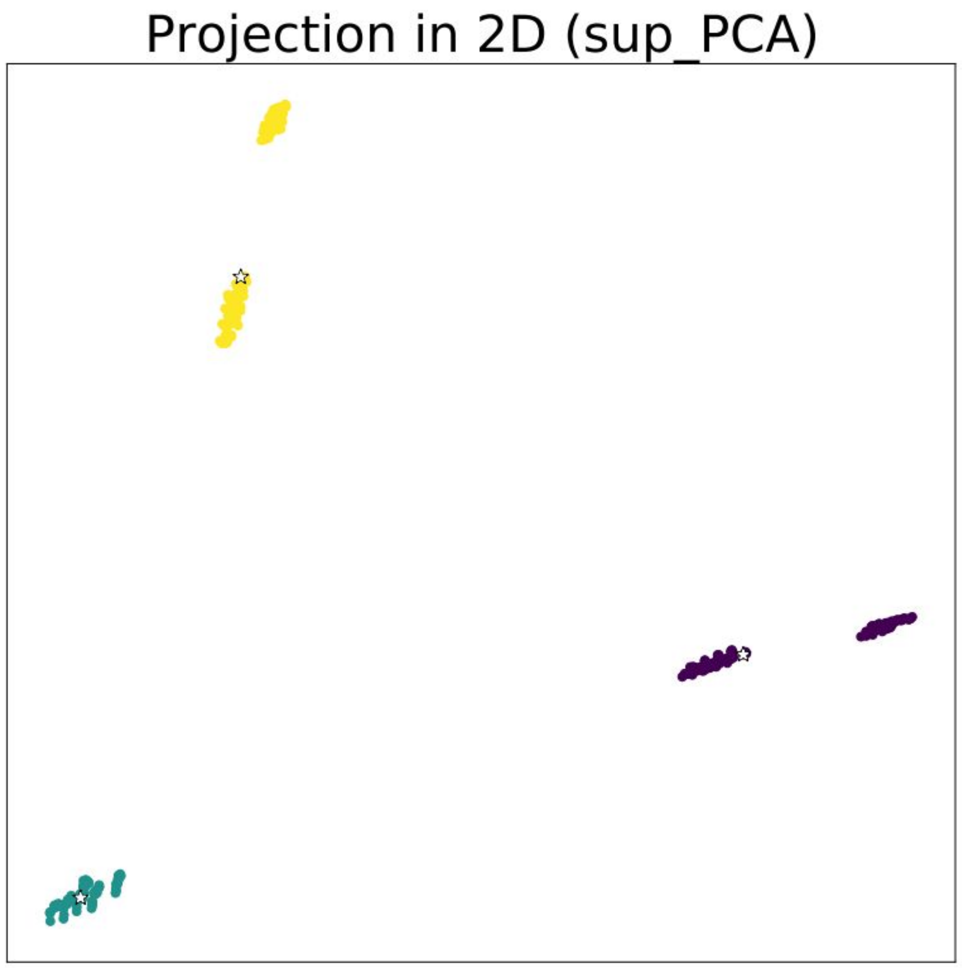
Description	Value
Best number K of groups/representative samples	3
Sum of intra-cluster variances before optimization	0.027594
Sum of intra-cluster variances after optimization	0.027594
Linear Regression MSE before optimization	0.009169
Linear Regression MSE after optimization	0.009169

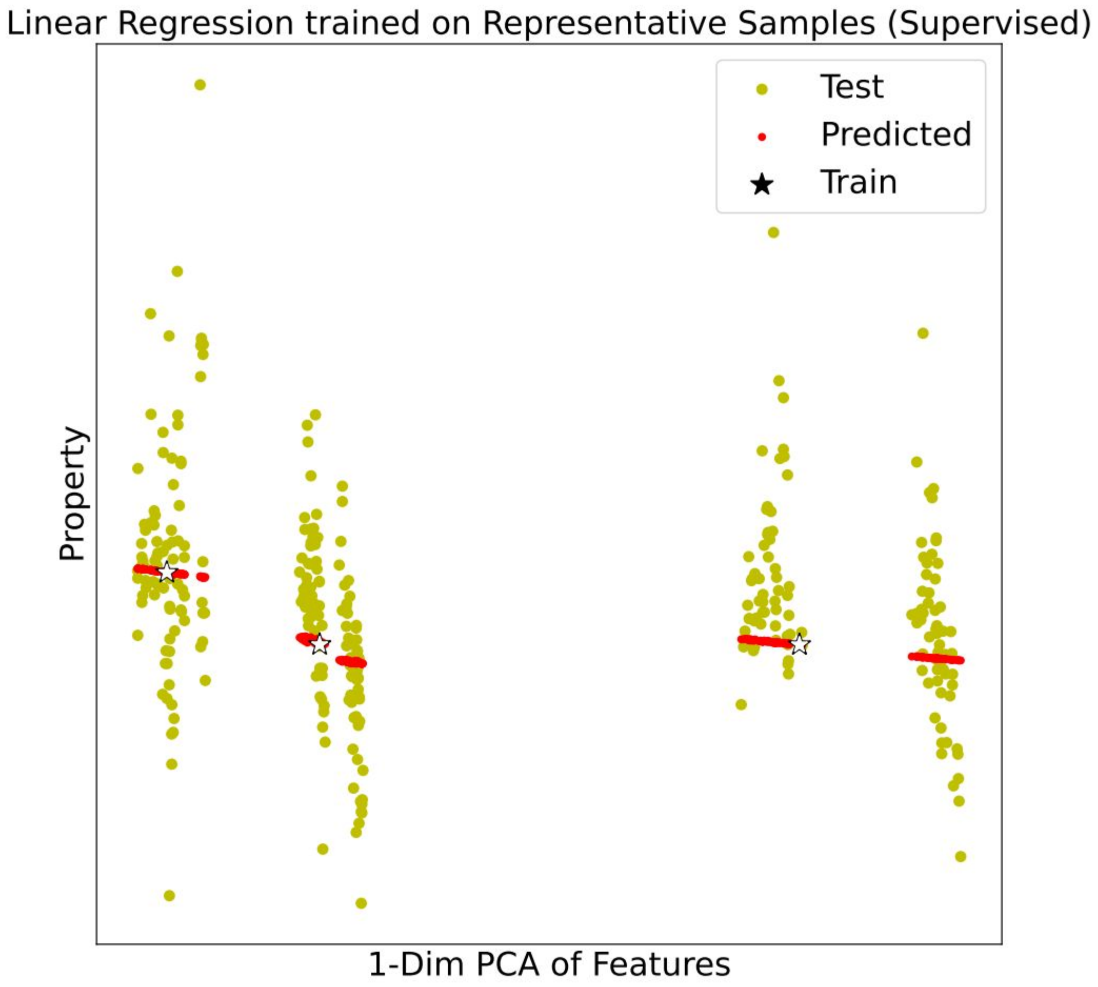


[Dataset] MENDES, P. C. D. et al. Ab initio screening of pt-based transition-metal nanoalloys using descriptors derived from the adsorption and activation of co2. Phys. Chem. Chem. Phys., The Royal Society of Chemistry, v. 23, p. 6029–6041, 2021.

Results

55-Atom Pt-Based Core–Shell Nanoalloys





Results

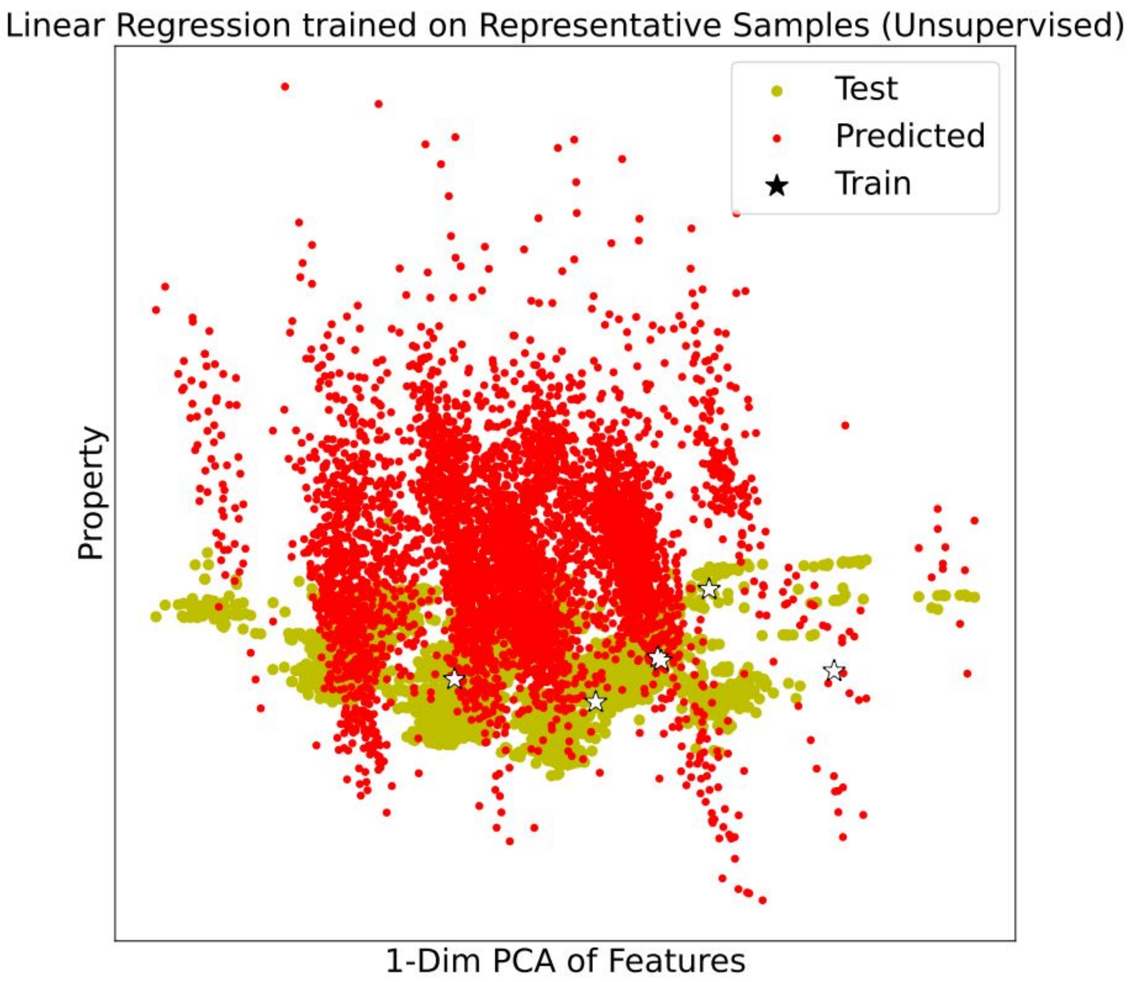
QM9

Description	Value
Defined number K of groups/representatives	6
Sum of intra-cluster variances before optimization	27.995522
Sum of intra-cluster variances after optimization	22.388382
Linear Regression MSE before optimization	94.712168
Linear Regression MSE after optimization	26.794895

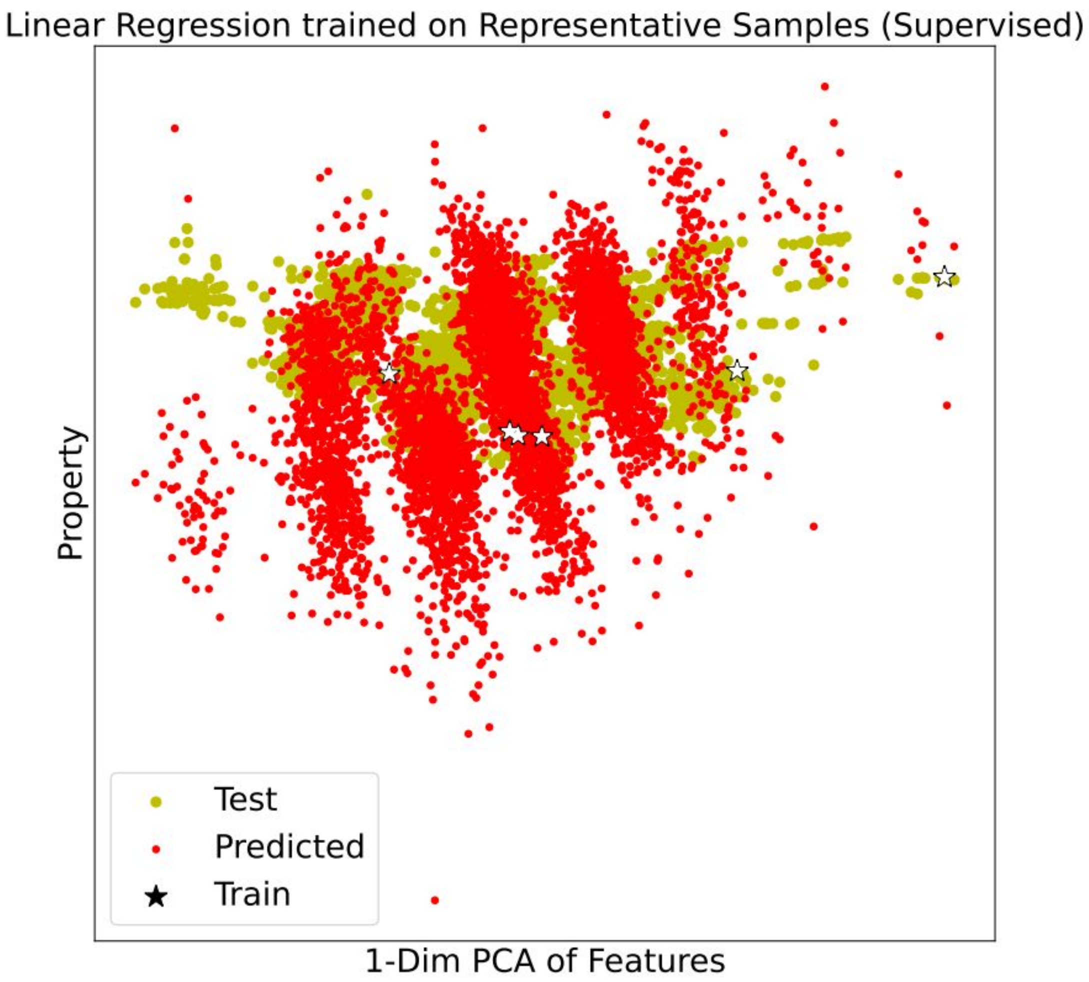
[Dataset] Quantum chemistry structures and properties of 134 kilo molecules. Scientific data, Nature Publishing Group, v. 1, p. 140022, 2014.

Results

QM9



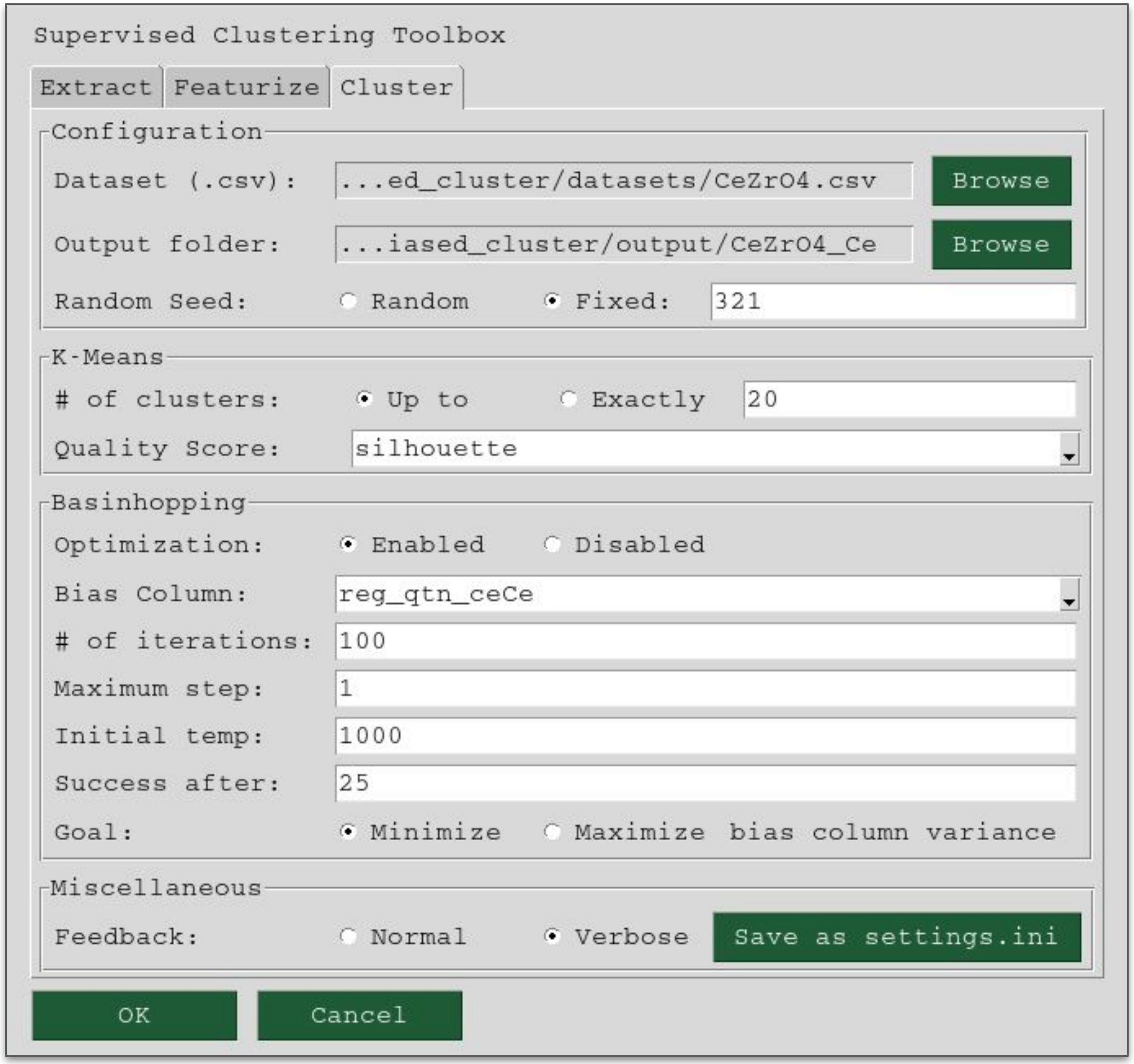
Linear Regression trained on Representative Samples (Unsupervised)
Test
Predicted
Train
Property
1-Dim PCA of Features



Linear Regression trained on Representative Samples (Supervised)
Test
Predicted
Train
Property
1-Dim PCA of Features

Toolbox and Further Development

- Easy to setup
- CLI
- GUI
- Multiplatform
- Multi-threaded



Conclusions

- Supervised clustering for selecting representative samples in databases.

- According to the analyses, it tends to outperform traditional clustering.

- Toolbox with Command Line / Graphical interface to run the algorithm.

Thank You!

UNIFESP
UNIVERSIDADE FEDERAL DE SÃO PAULO
IQSC USP
QTNano
Juarez L. F. Da Silva Group
FAPESP
CNPq
CINE
CENTER FOR INNOVATION ON NEW ENERGIES