Machine Learning and Data Mining Algorithms Applied in Computational Materials Science

Felipe V. Calderan[1], Marcos G. Quiles[1], Juarez L. F. Da Silva[2]
[1] Institute of Science and Technology, Federal University of São Paulo
[2] São Carlos Institute of Chemistry, University of São Paulo

Computational Materials Science & Chemistry

Acknowledgments

FAPESP

UNIFESP
UNIVERSIDADE FEDERAL DE SÃO PAULO

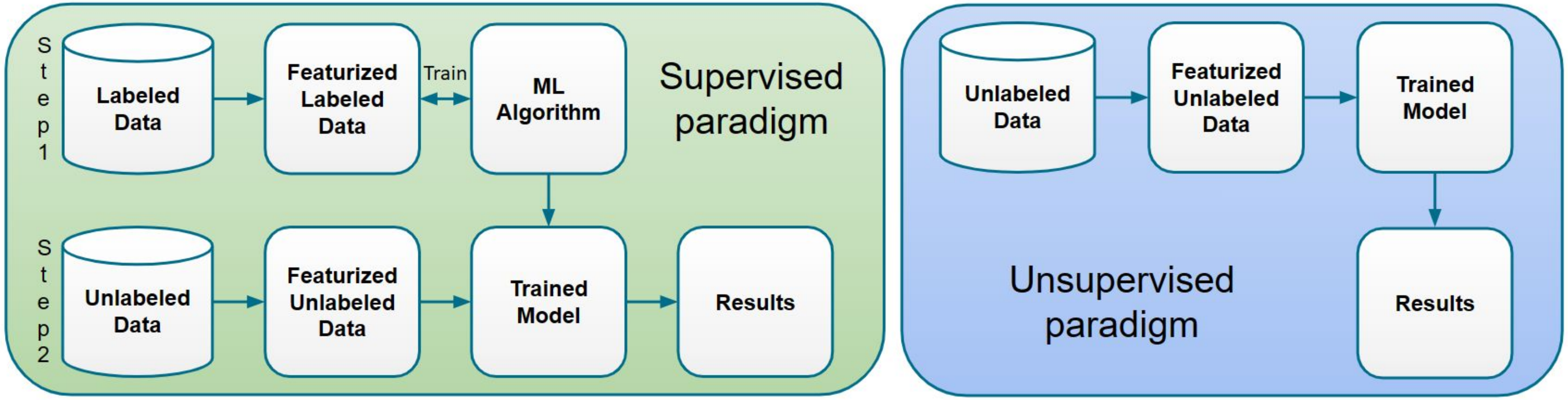
CINE
CENTER FOR INNOVATION ON NEW ENERGIES

CIDAG
Computational Intelligence and Data Analysis Group

QTNano
Juarez L. F. Da Silva Group

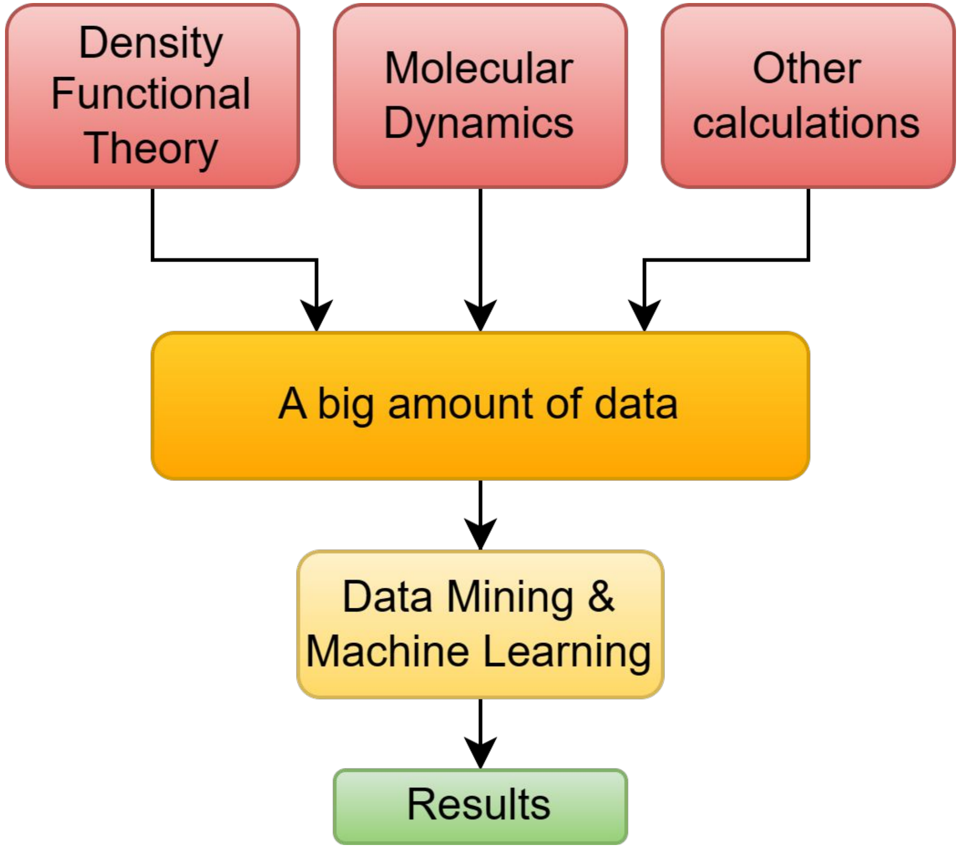
Machine Learning and Data Mining

- Machine Learning (ML) is a subset of Artificial Intelligence (AI) that studies algorithms that **learn from data** instead of being explicitly programmed.
- Typical uses for ML include prediction, classification and clustering.
- Data Mining (DM) **extracts useful information** from datasets by using statistical techniques.
- ML is used, within DM, as a way to perform pattern recognition and data visualization.



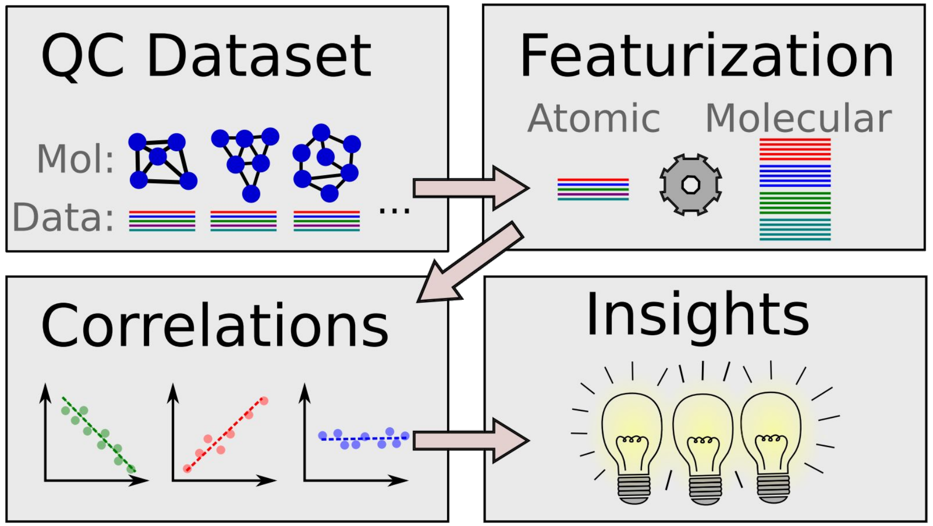
Combining Data Mining Techniques with Materials Science

- Materials Science (MS) generates huge amounts of data, due to Molecular Dynamics (MD), Density Functional Theory (DFT) and other calculations.
- This is a great scenario for Data Mining, in contrast to manual analyses.
- Here, We'll showcase cutting edge ML+DM tools built by the group that have potential to boost MS research.



Automatic features-property relationship analysis

Finding **features that correlate well with a target property** is an essential part of building a good ML model. This framework proposes a featurization process to convert atomistic features into molecular properties (AtoMF), which have their relationship with a target property automatically tested. Correlation techniques tested: Pearson, Spearman and Kendall. We selected Spearman due to the values being more spread inside [-1, 1] range and less sensitive to outliers.



QC Dataset

Mol:

Data:

...

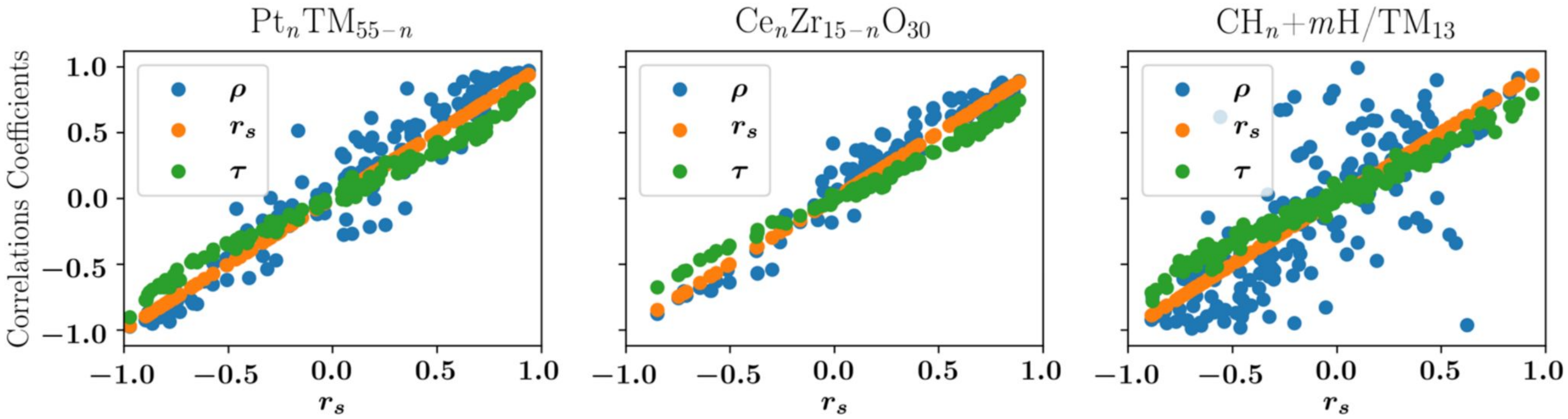
Featurization

Atomic

Molecular

Correlations

Insights

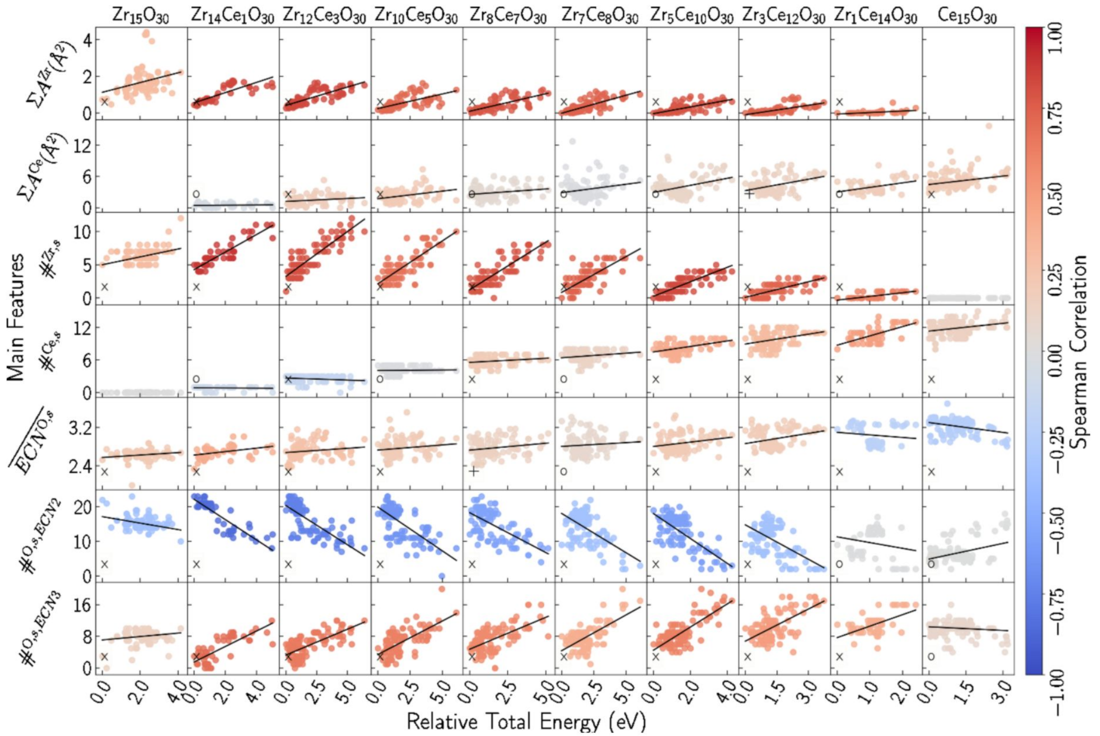


$\mathrm{Pt}_n\mathrm{TM}_{55-n}$

$\mathrm{Ce}_n\mathrm{Zr}_{15-n}\mathrm{O}_{30}$

$\mathrm{CH}_n+m\mathrm{H}/\mathrm{TM}_{13}$

Samples of insights obtained

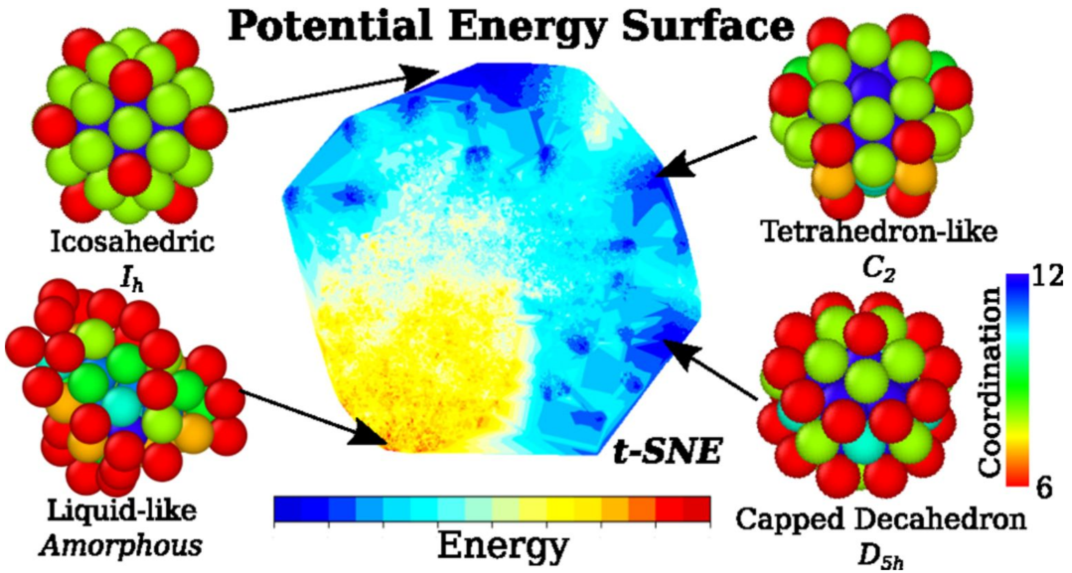


For $Ce_nZr_{15-n}O_{30}$, feature ΣA^{Zr} has a bigger magnitude and has more significant correlations than ΣA^{Ce}, showing that Zr atoms are preferred over Ce for core sites or surface sites with little vacuum exposition. Similar patterns shown with $\#^{Zr,S}$ and $\#^{Ce,S}$.

To understand the diagram generated by the framework, it is important to point out that Direct correlations indicate that as a particular feature increases in value, so does the target property, while inverse correlations mean that as a particular feature increases in value, the target property decreases and vice versa.

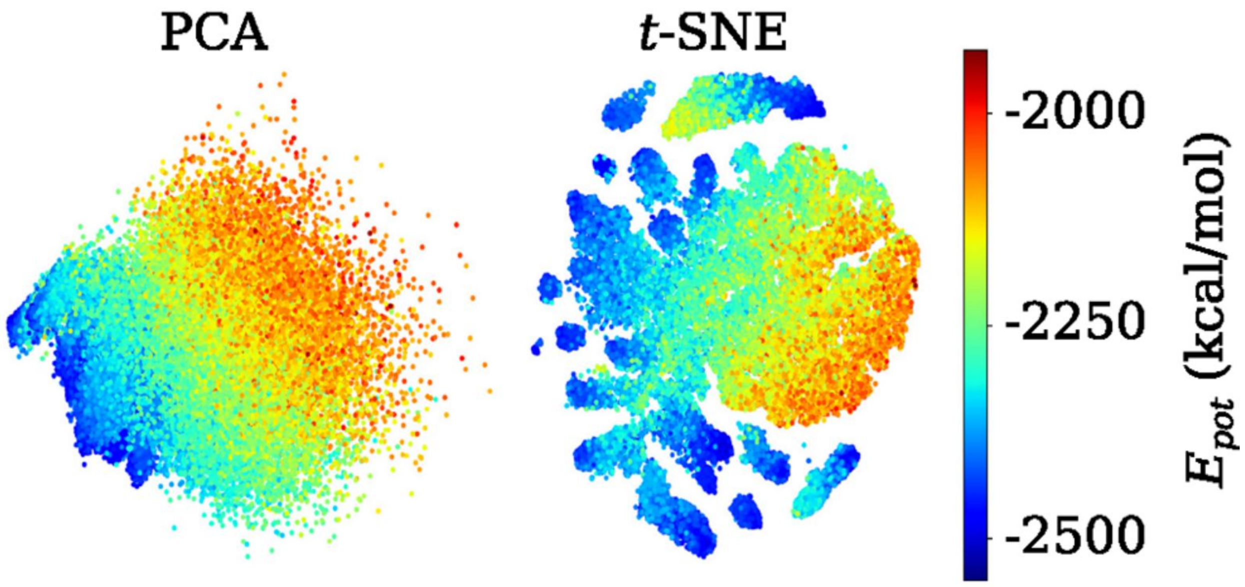
With this in mind, we have that the change of the amount of Zr in core sites or sites with little vacuum exposition has a bigger impact in the Relative Total Energy value of the system than the change of Ce.

Potential Energy Surface exploration and phase transition mapping



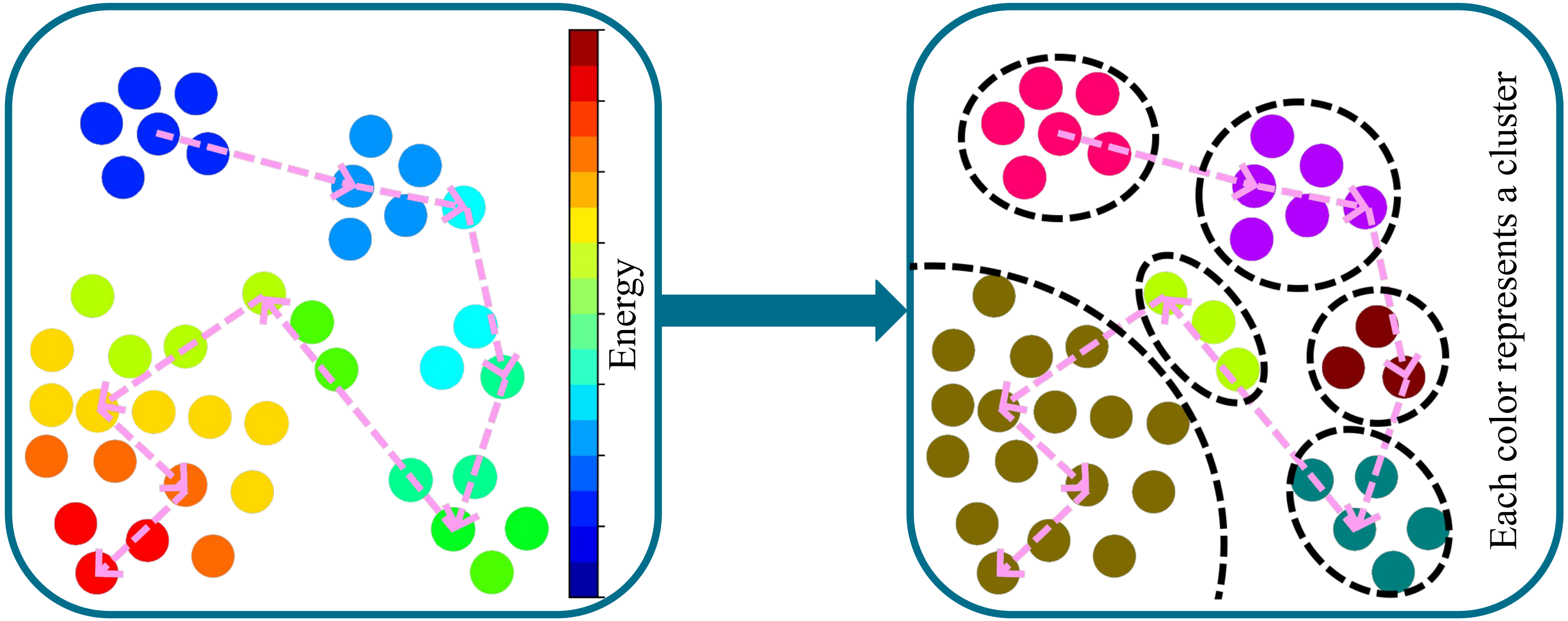
Exploration of the Potential Energy Surface (PES) enables the discovery of the most important morphologies and phase transitions. Considering the variety of experimental conditions and the sensitivity of nanocluster properties given the geometry, this tool can be used for the development of many technologies.

The analysis of PES has many challenges, such as the number of minima present and a good MD trajectory representation, both of which benefit from having a good high-dimensional projection into a low dimension.



Machine Learning in the inner workings

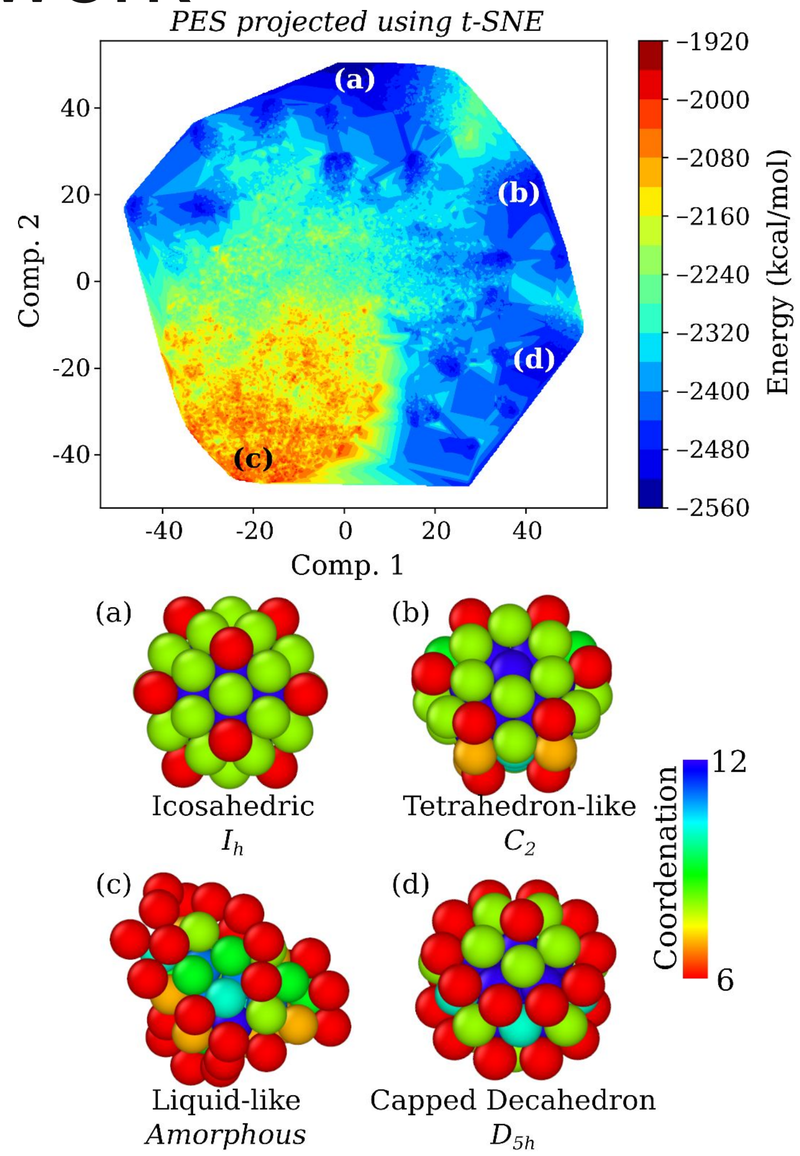
Having the PES and trajectories projected by **t-SNE**, the framework uses **DBSCAN+kNN** to cluster the data and suggest phase transitions. Transitions are validated by a post-processing algorithm based on chemical knowledge.



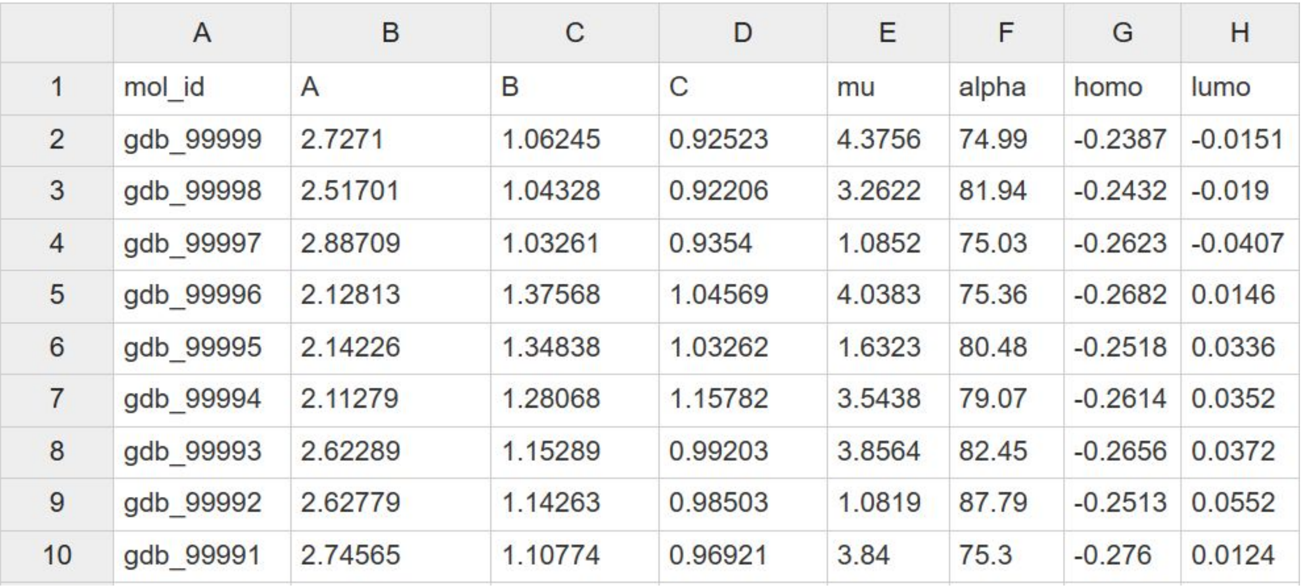
Applications for the theoretical framework

- PES projection plots with MD trajectories
- Phase and geometrical transitions
- Stability of a geometrical configuration
- Ability to classify atomic arrangements by 2D surface analysis (as depicted to the right).

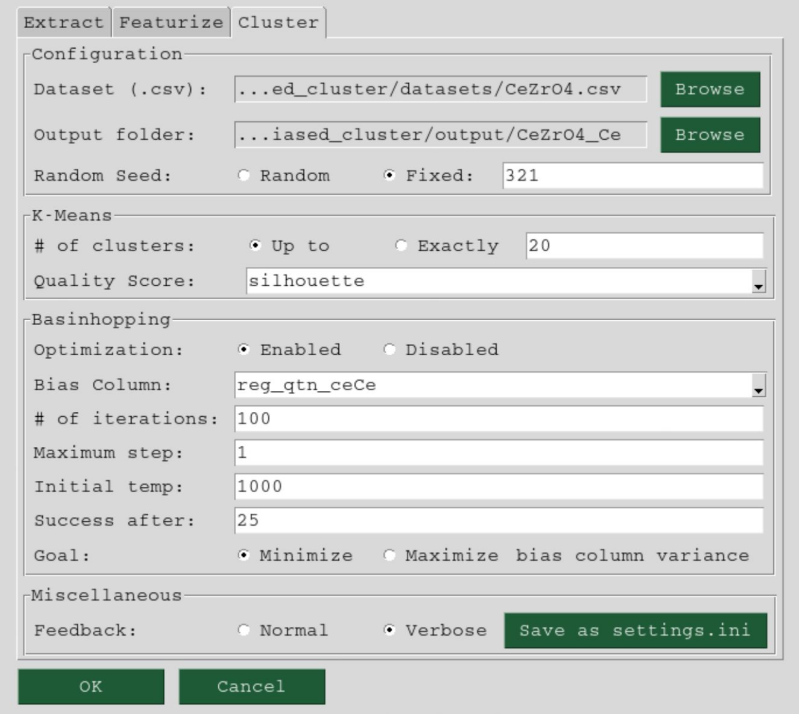
The paper validates these insights on $\boldsymbol{Cu_n}$ MD dataset for $\boldsymbol{n = 13, 38, 55, 75, 98, 102}$ and $\boldsymbol{147}$.



Guided Clustering for Selecting Representatives Samples in Chemical Databases



	A	B	C	D	E	F	G	H
1	mol_id	A	B	C	mu	alpha	homo	lumo
2	gdb_99999	2.7271	1.06245	0.92523	4.3756	74.99	-0.2387	-0.0151
3	gdb_99998	2.51701	1.04328	0.92206	3.2622	81.94	-0.2432	-0.019
4	gdb_99997	2.88709	1.03261	0.9354	1.0852	75.03	-0.2623	-0.0407
5	gdb_99996	2.12813	1.37568	1.04569	4.0383	75.36	-0.2682	0.0146
6	gdb_99995	2.14226	1.34838	1.03262	1.6323	80.48	-0.2518	0.0336
7	gdb_99994	2.11279	1.28068	1.15782	3.5438	79.07	-0.2614	0.0352
8	gdb_99993	2.62289	1.15289	0.99203	3.8564	82.45	-0.2656	0.0372
9	gdb_99992	2.62779	1.14263	0.98503	1.0819	87.79	-0.2513	0.0552
10	gdb_99991	2.74565	1.10774	0.96921	3.84	75.3	-0.276	0.0124



Configuration

Dataset (.csv): ...ed_cluster/datasets/CeZrO4.csv

Output folder: ...iased_cluster/output/CeZrO4_Ce

Random Seed: ○ Random ● Fixed: 321

K-Means

# of clusters: ● Up to ○ Exactly 20

Quality Score: silhouette

Basinhopping

Optimization: ● Enabled ○ Disabled

Bias Column: reg_qtn_ceCe

# of iterations: 100

Maximum step: 1

Initial temp: 1000

Success after: 25

Goal: ● Minimize ○ Maximize bias column variance

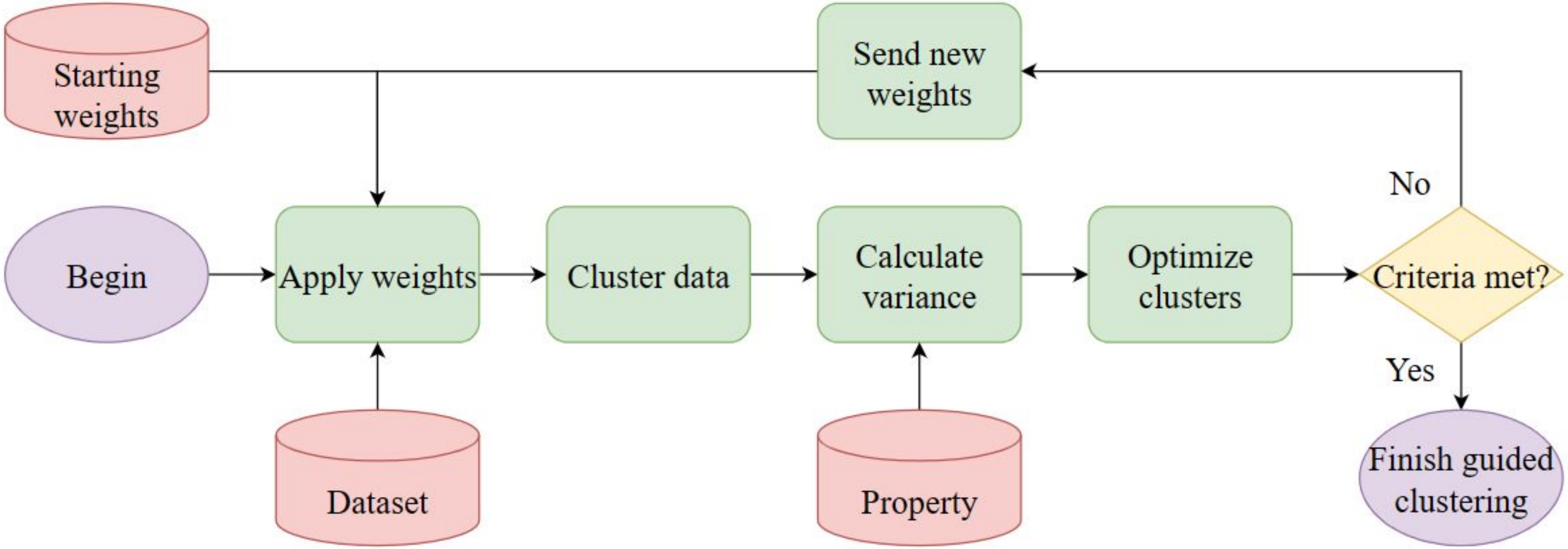
Miscellaneous

Feedback: ○ Normal ● Verbose

Analysis of a molecule or nanocluster can be expensive and time-consuming, experimentally and computationally. Selecting those that represent the best a large amount of similar entities is very interesting, since analyzing only them is cheaper, faster and gives a good overview of what the rest look like. This tool enables the selection of representatives and automates many of the tasks that would have to be done manually, such as selecting the number of clusters and scaling the features properly to highlight a desired property.

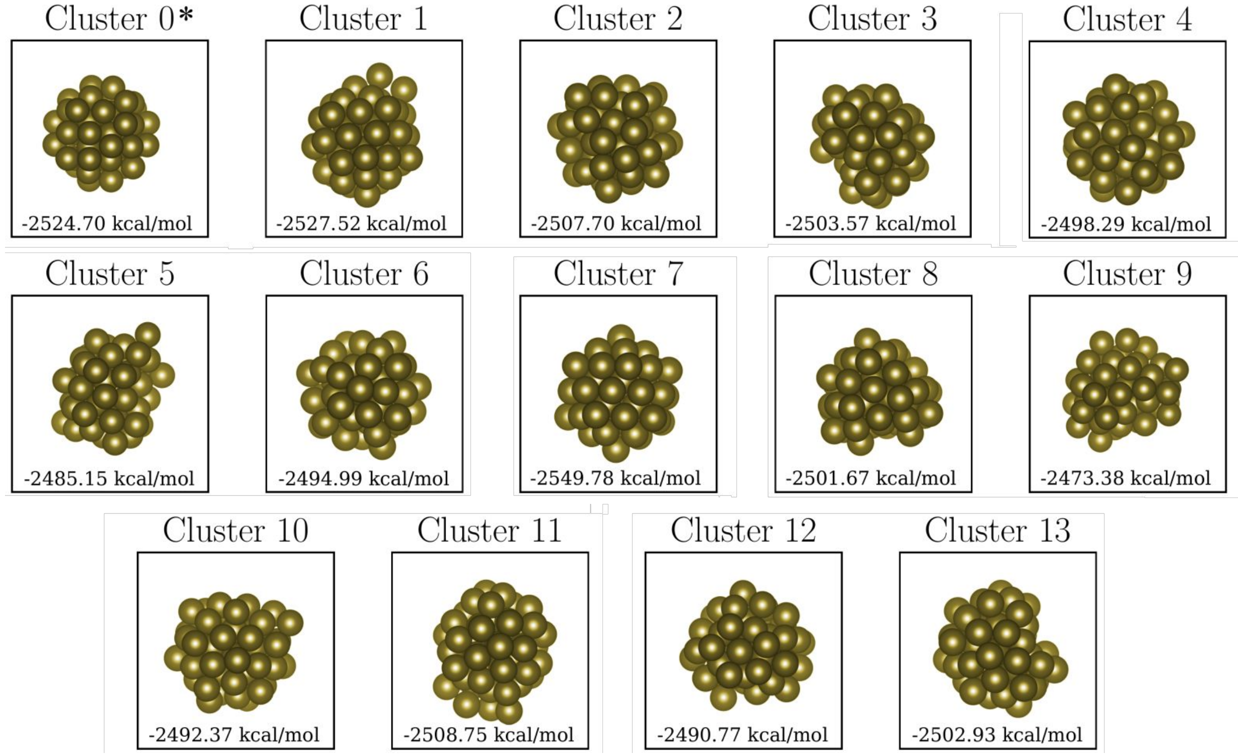
Finding better representatives

The toolbox suggests a value for K based on the **Silhouette Criterion**, and scales the values of the features by an iterative optimization depicted below.



Applications for Cu_{55} nanoclusters

- $K=14$ clusters or representatives minimizes E_{tot} intracluster variance.
- 3 of 4 clusters with low energies averages are associated to I_h, C_2 and D_{5h} shapes, which are highly stable.
- Automatic scaling has a big impact on the feature space. Comparing to non-scaled features, we have a Normalized Mutual Index (NMI) of 0.730.



Conclusion

There is an ever-growing amount of studies combining ML and MS, and as a result of our efforts in the domain of these studies, we can now:

- Automatically find correlations between features and properties
- Project PES and map phase/geometric transitions
- Find representative samples with underlying physicochemical knowledge

And there are still gaps in the literature that we can fill. As an example, several methods can be applied for representative sample selection, but there is no survey in the chemical/materials science fields.