UNIFESP

UNIVERSIDADE FEDERAL DE SÃO PAULO

1933

IQSC USP

V CINE-CMSC Workshop

Clustering Methods for Selecting Representative Samples in Chemical Databases

Student:
Felipe V. Calderan

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

July 10th, 2020

Motivation

Provide a faster method of **materials screening**, so that experts have access to large-scale analysis of material properties.

How?

Using **Machine Learning** methods, specifically **supervised clustering**, to obtain molecules that represent a larger set. Only these molecules will need to be analyzed using costly methods like **Density Functional Theory**, instead of all the others.

Supervised Clustering?

For a supervised clustering system, two algorithms working together are needed:

- Clustering algorithm
- Optimization algorithm

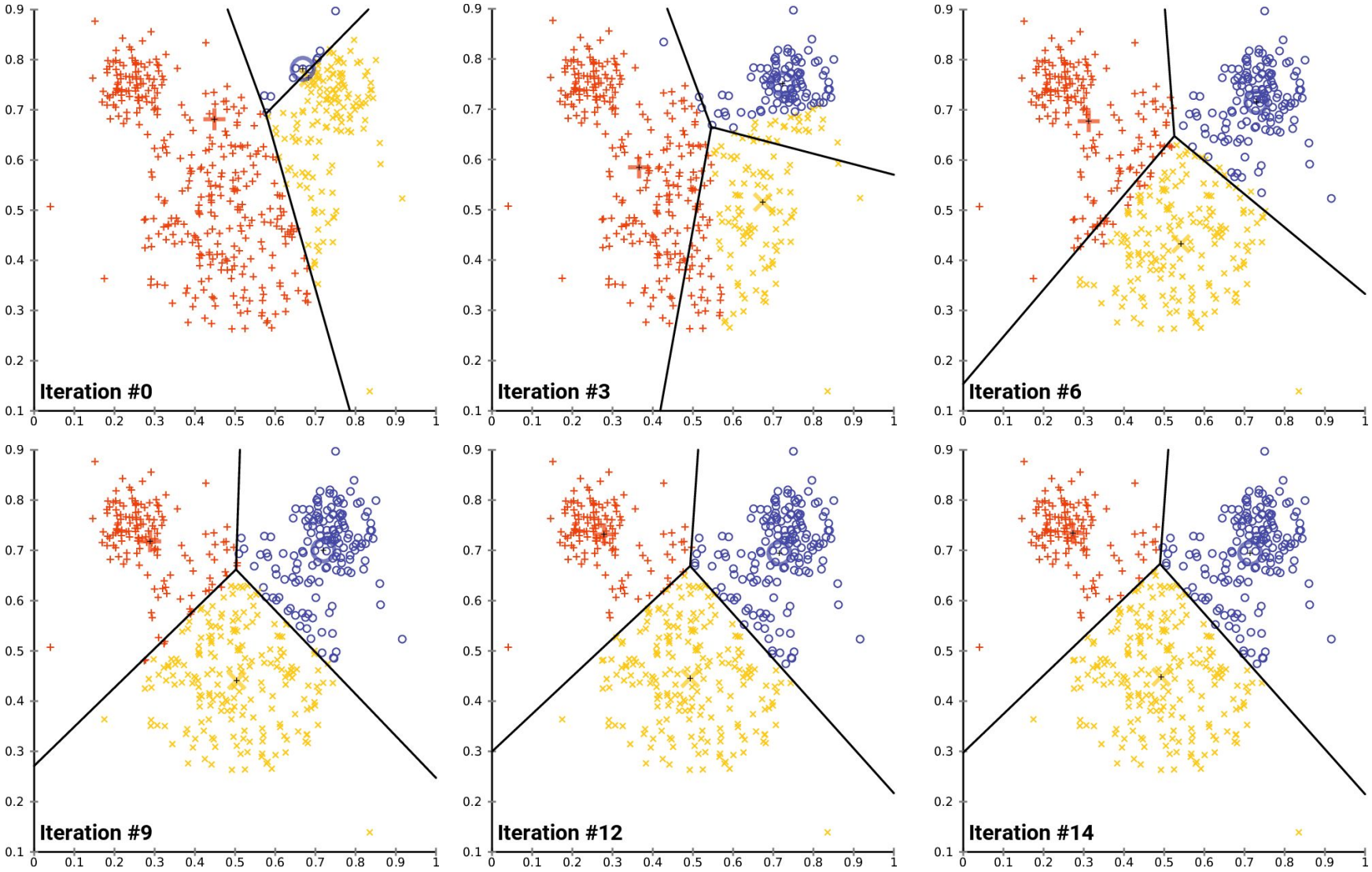
For the purposes of this presentation, it will be shown how **K-Means** and **Simulated Annealing** work together to bias the clustering process, in order to satisfy the needs of the specialist.

K-Means

Clusters dataset's items based on similarity.

1. Choose k centroids to match k random elements from the database
2. Assign each element to the nearest (most similar) 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

Clustering Methods for Selecting Representative Samples in Chemical Databases



Simulated Annealing

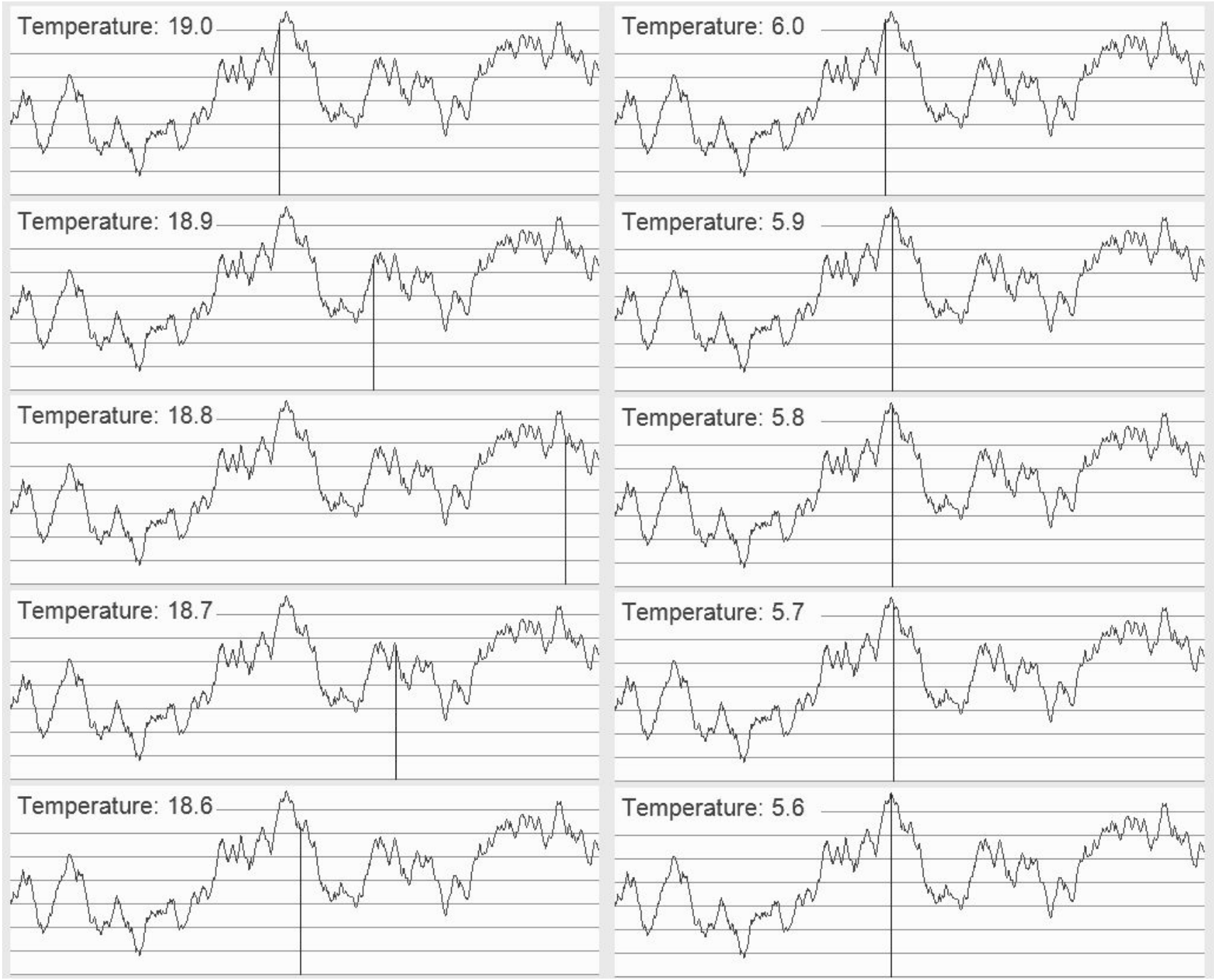
Aims to find global optima of functions

1. Initialize the temperature t as a high value and the current state c0 as a set of random values

2. Select a new state c1 from a neighbor of the current state and calculate $\Delta C = c1 - c0$

3. If $\Delta C \leq 0$ or $\exp(-\Delta C / t) > \text{random}(0, 1)$ then
   current_state <- new_state
   Else
   do nothing

4. Decrement t. If there is still no convergence, repeat the second step



Sim. Annealing Convergence By Kingpin13 - Own work, CC0, https://commons.wikimedia.org/w/index.php?curid=25010763

K-Means + Simulated Annealing

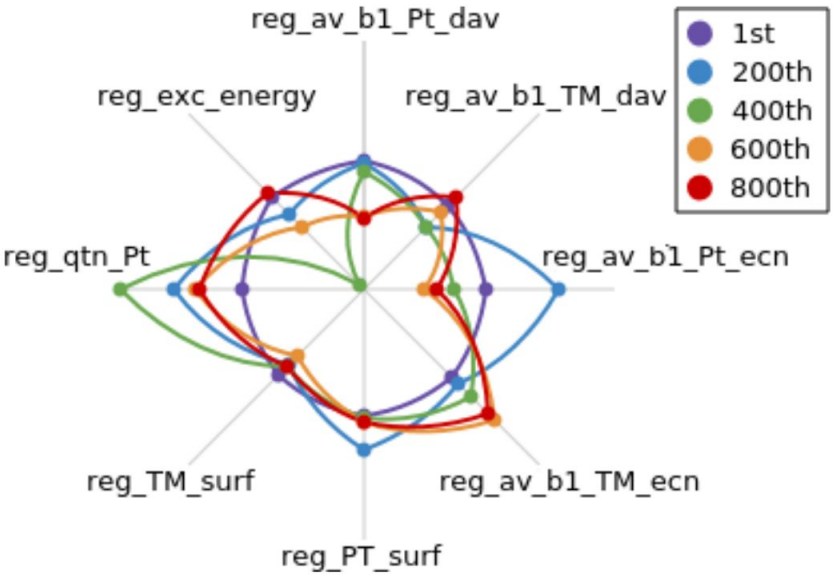
K-Means receives an array of features, where each feature has a weight **w**, so that clustering can be done. From the configuration generated by K-Means, a result **v** is extracted. This result is precisely what will be optimized by Simulated Annealing. The optimization algorithm control the weights **w** given the value **v**.

$$f\left(K\text{–}Means\left(\begin{bmatrix} v_{11}w_1 & v_{12}w_2 \\ v_{21}w_1 & v_{22}w_2 \\ v_{31}w_1 & v_{32}w_2 \end{bmatrix}\right)\right) = v$$

$$Simulated\ Annealing \Downarrow \quad \Downarrow \quad \Uparrow$$

K-Means + Simulated Annealing

Thus, it is possible to bias the clusters following an objective. For example: minimize the maximum variance in the number of elements per cluster.

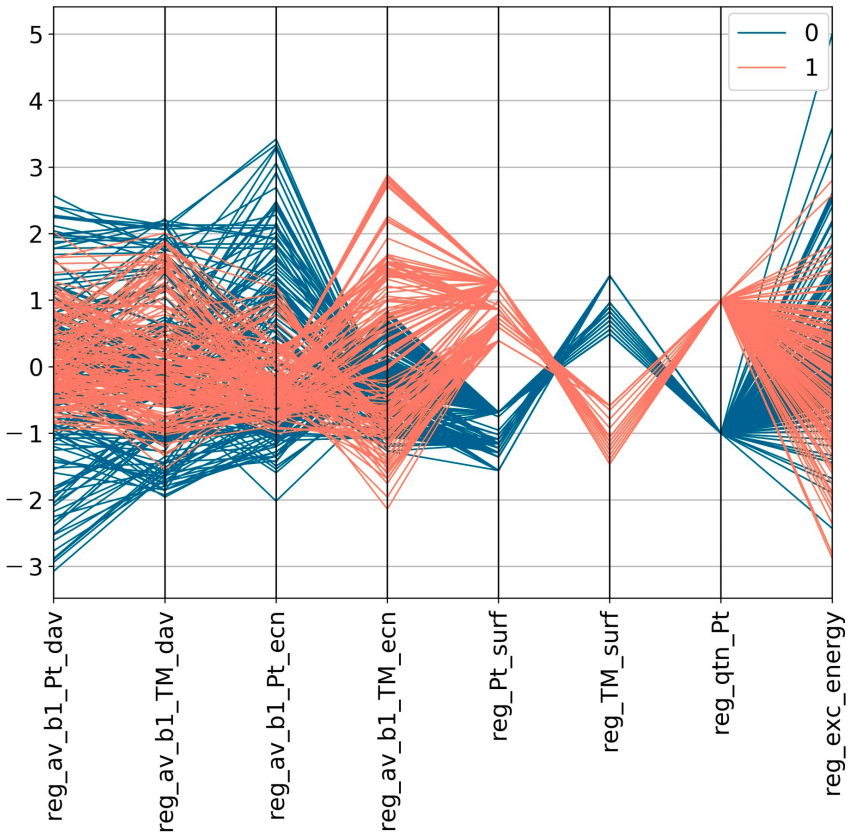


The figure to the side shows the convergence of the weights of the features along the iterations.

Notice how, in the beginning, all features had the same weight and these weights were transformed until they converged.

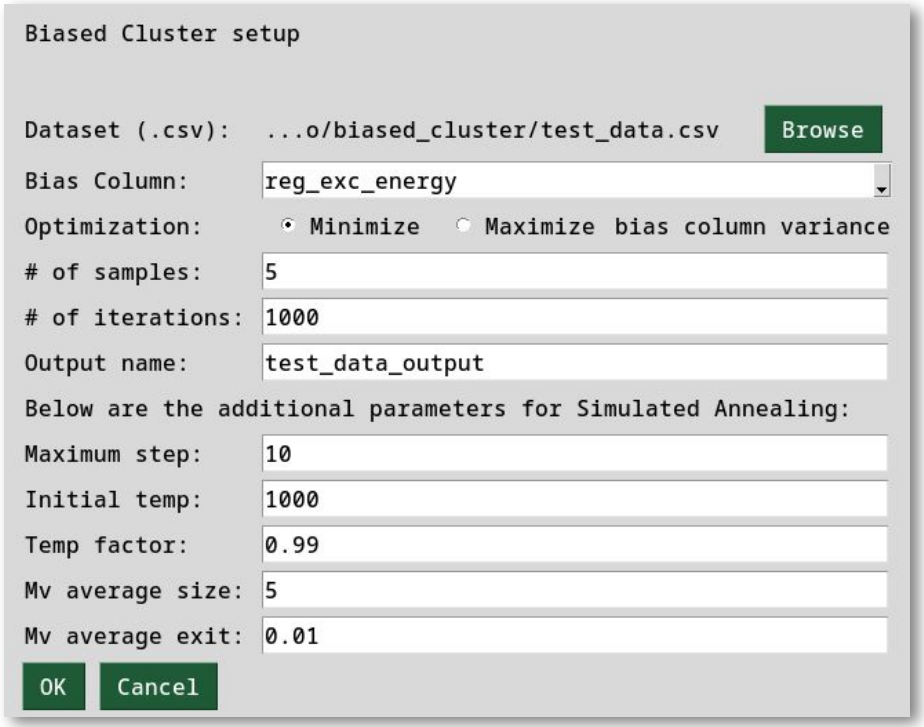
Example Result

In the graph we can see a disparity in **reg_qtb_Pt** (amount of platinum), **reg_TM_surf** (amount of transition metals on the surface) and **reg_Pt_surf** (amount of platinum on the surface). That is, the molecules were grouped taking into account mainly the amount of platinum and transition metals. Increasing the number of clusters would cause subdivisions into groups given by less prevalent characteristics such as **reg_av_b1_Pt_ecn** and **reg_av_b1_TM_ecn** (average number of neighbors of Pt or TM in the molecule).



Toolbox in development

A toolbox for executing the supervised/biased clustering algorithm from graphical or command line interfaces is being developed.



Acknowledgements

UNIFESP
UNIVERSIDADE FEDERAL DE SÃO PAULO
1933

PIBIC
CNPq

CINE
CENTER FOR INNOVATION ON NEW ENERGIES

C M S C
Computational Materials Science and Chemistry

QTNano

IQSC USP

FAPESP

Thank you!

References:

(1) Jain, A. K. Data clustering: 50 years beyond K-means. Pattern Recognition Letters 2010, 31, 651–666.
(2) Cha, S.-H. Comprehensive Survey on Distance/Similarity Measures Between Probability Density Functions. Int. J. Math. Model. Meth. Appl. Sci. 2007, 1 .
(3) Jain, A. K.; Murty, M. N.; Flynn, P. J. Data Clustering: A Review. ACM Computing Surveys 1999, 31, 265–323.
(4) Yang, X.-S. Introduction to Mathematical Optimization: From Linear Programming to Metaheuristics; Cambridge International2 Science Publishing, 2008.
(5) van Laarhoven P.J.M., A. E. Simulated annealing. Simulated Annealing: Theory and Applications. 1987; pp 7–15.
(6) Wagstaff, K.; Cardie, C.; Rogers, S.; Schrödl, S. Constrained K-means Clustering with Background Knowledge. 2001; pp 577–584.
(7) Auffarth, B. Clustering by a genetic algorithm with biased mutation operator. 2010; pp 1 – 8.
(8) Rose, K. Deterministic annealing, clustering, and optimization. Ph.D. thesis, California Institute of Technology, 1991. (9) Preparata, F. P.; Shamos, M. Computational Geometry: An Introduction; SpringerVerlag New York, 1985.
(10) Abdi, H.; Williams, L. J. Principal component analysis. Wiley Interdisciplinary Reviews: Computational Statistics 2010, 433–459.
(11) Wegman, E. J. Hyperdimensional Data Analysis Using Parallel Coordinates. Journal of the American Statistical Association 1990, 85, 664–675