V CINE-CMSC Workshop

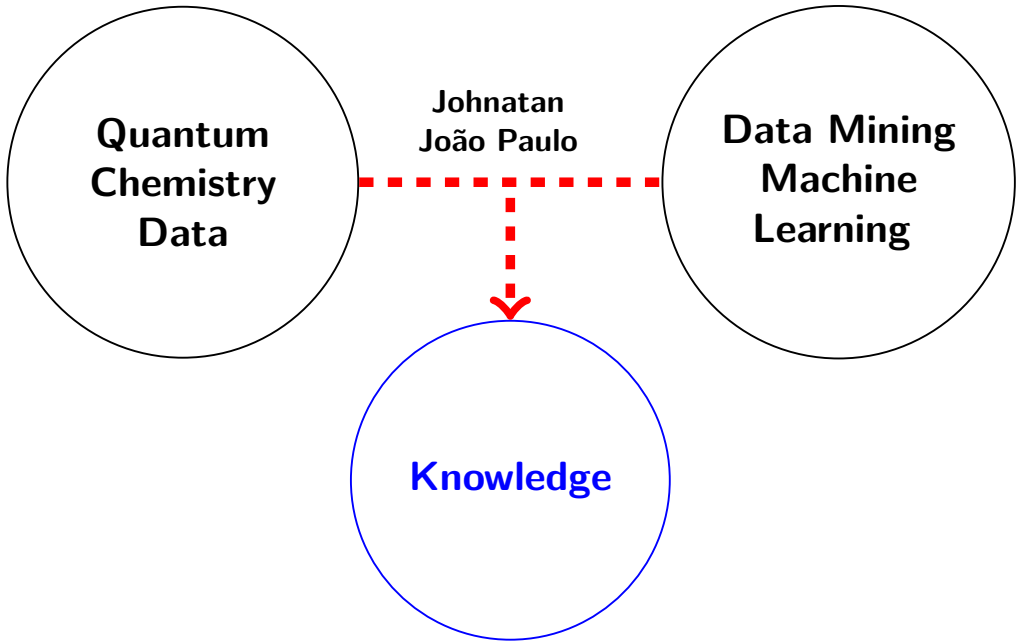
A Feature Engineering and Correlation-based Framework to Knowledge Extraction from Quantum Chemistry Datasets: the Nanoclusters Examples.

Johnatan Mucelini

Advisor: Juarez L. F. Da Silva

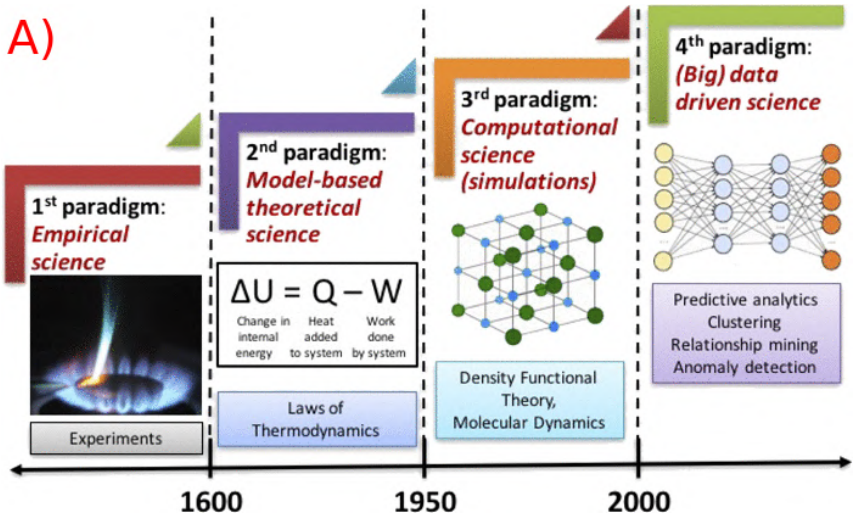
Quantum Chemistry Collaborators:
Priscilla Felício-Sousa;
Paulo de Carvalho Dias Mendes;
Karla F. Andriani.

Machine Learning Collaborators:
Marcos G. Quiles;
Ronaldo C. Prati.



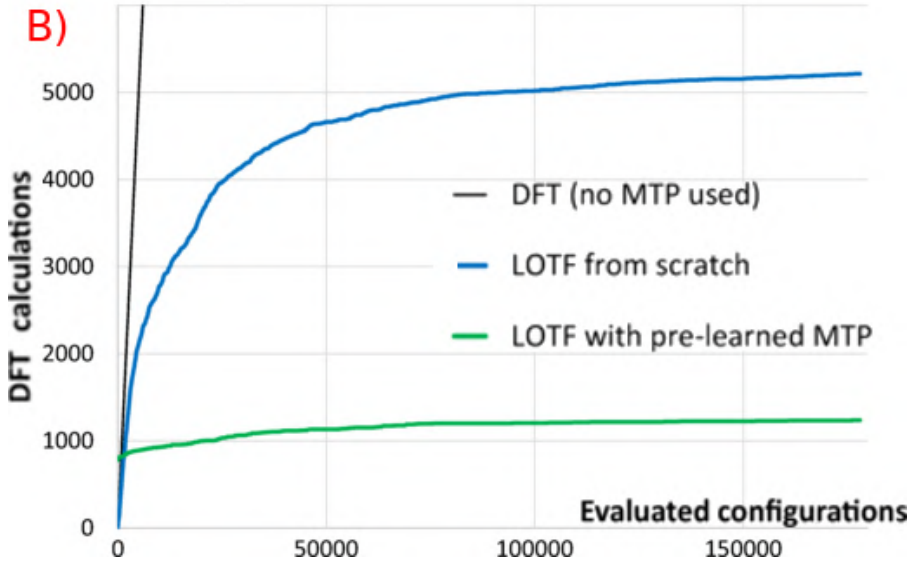
July, 2020

How to obtain insights from QC Data in the 4th paradigm?



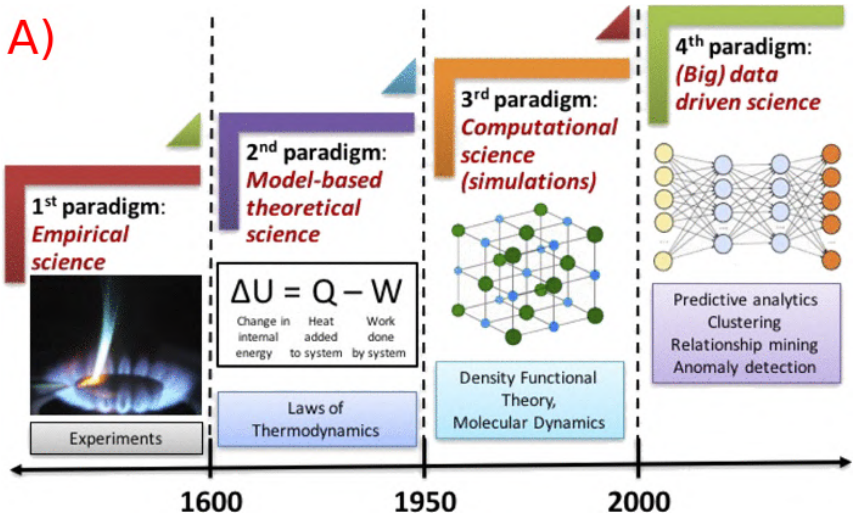
Font: *APL Materials*, **2016**, 4, 053208

The volume of data that we can build is large and will increase.

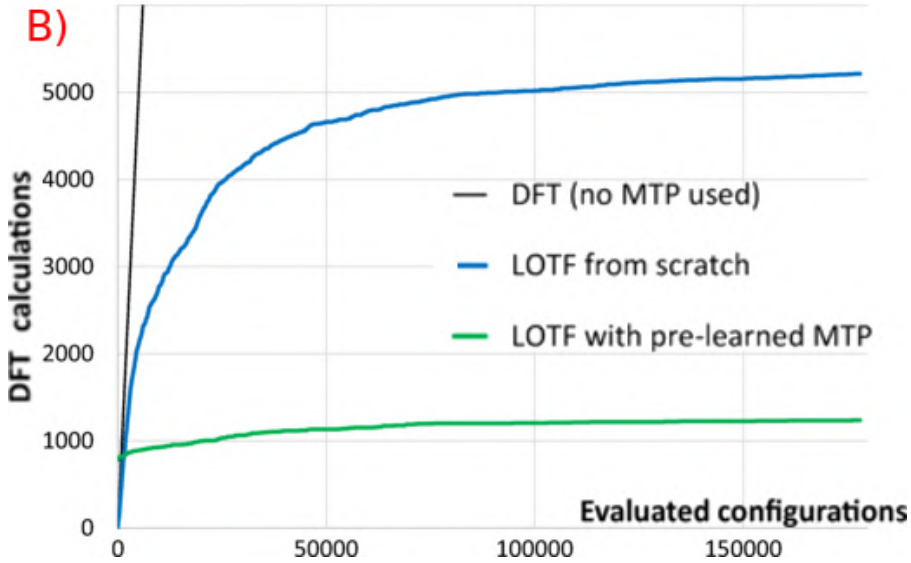


Font: *Phys. Rev. B*, **2019**, 99, 064114

How to obtain insights from QC Data in the 4th paradigm?



Font: *APL Materials*, **2016**, 4, 053208



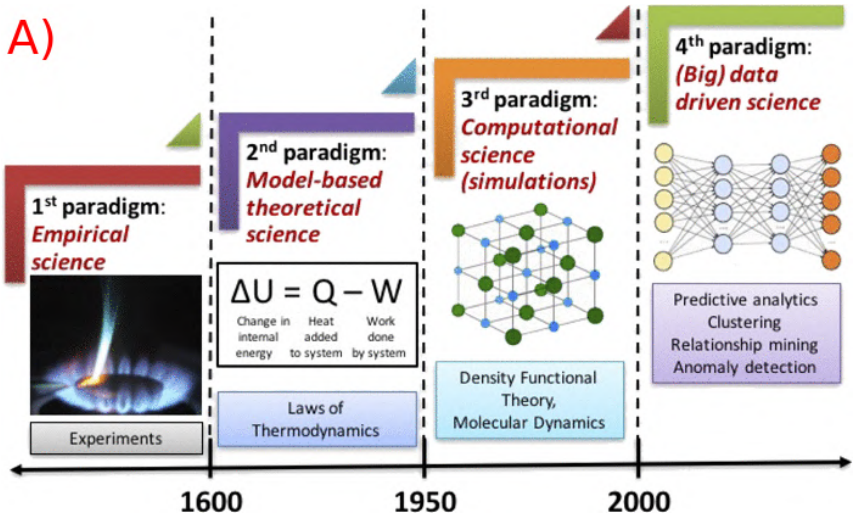
Font: *Phys. Rev. B*, **2019**, 99, 064114

The volume of data that we can build is large and will increase.

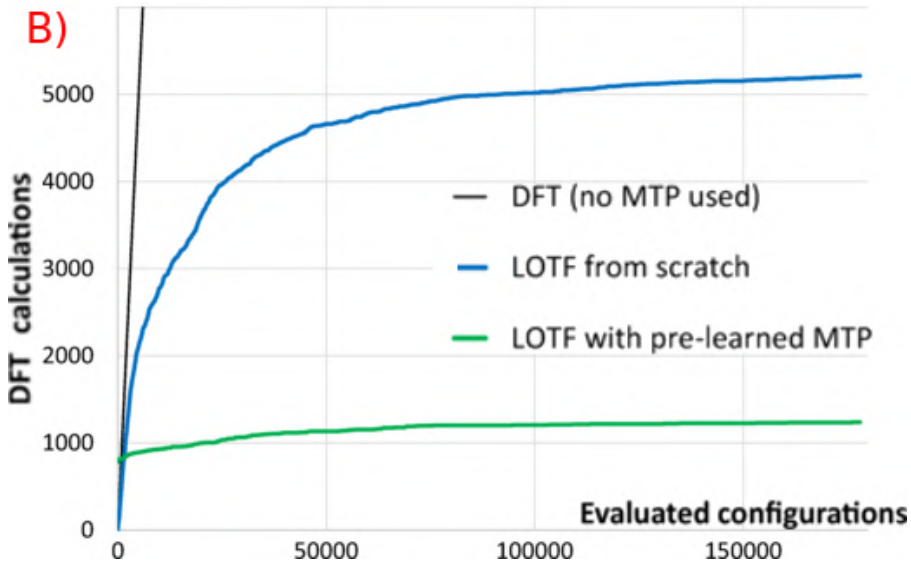
How we obtain knowledge/insights QC data at this moment?

- Analyze only the most stable structures.
- Visual description of atomic systems.
- Trends are visually identified and mensured.

How to obtain insights from QC Data in the 4th paradigm?



Font: *APL Materials*, **2016**, 4, 053208



Font: *Phys. Rev. B*, **2019**, 99, 064114

The volume of data that we can build is large and will increase.

How we obtain knowledge/insights QC data at this moment?

- Analyze only the most stable structures.
- Visual description of atomic systems.
- Trends are visually identified and mensured.

New QC data mining processes are desired!

- To reduce the human **time spended** and human **bias** introduced;
- To explore **all data** available (including **big data**);

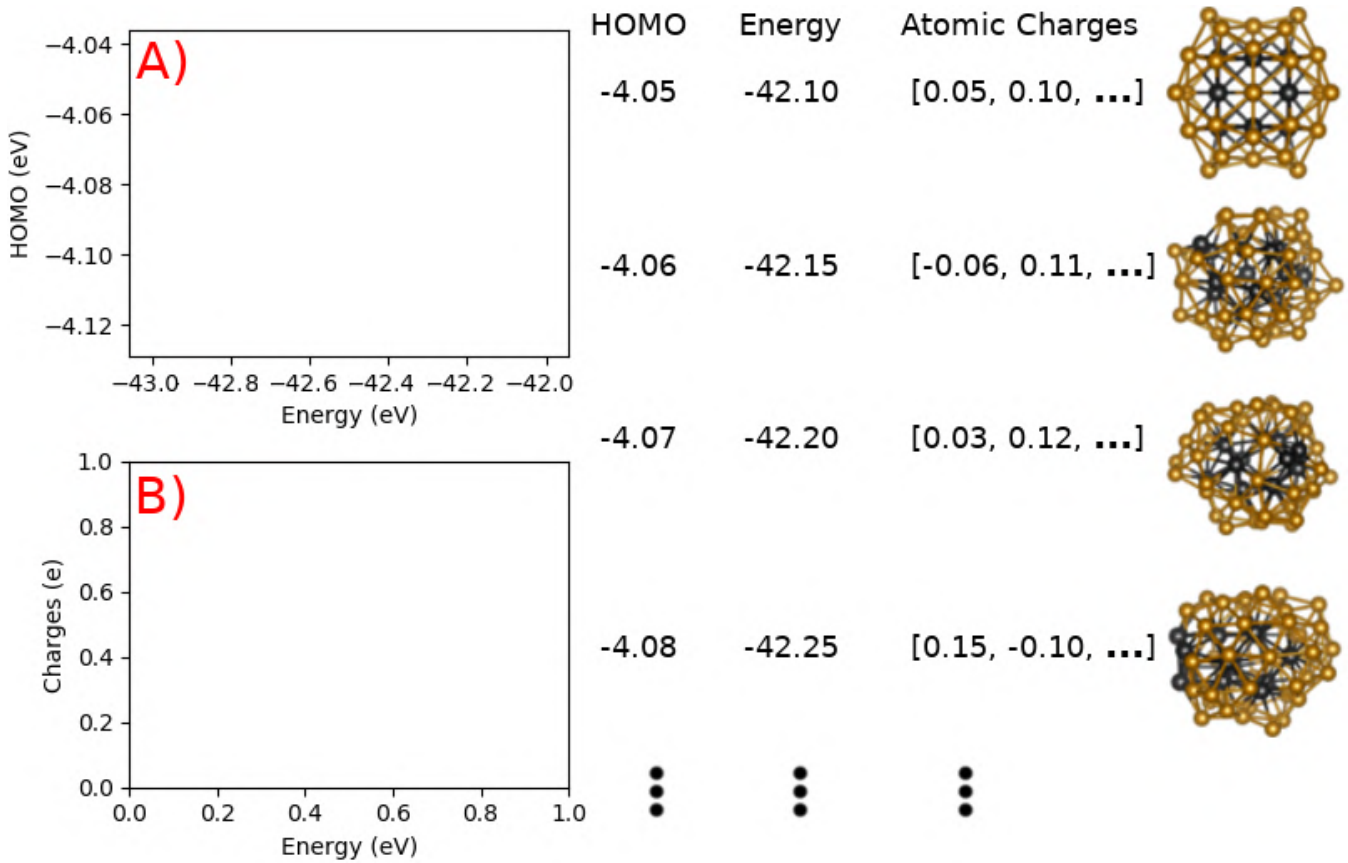
Types of Features with Physical Meaning

Molecular info (single value):

- Energy;
- HOMO.

Atomic info (n values):

- Effective Coordination Number;
- Atomic Charges.



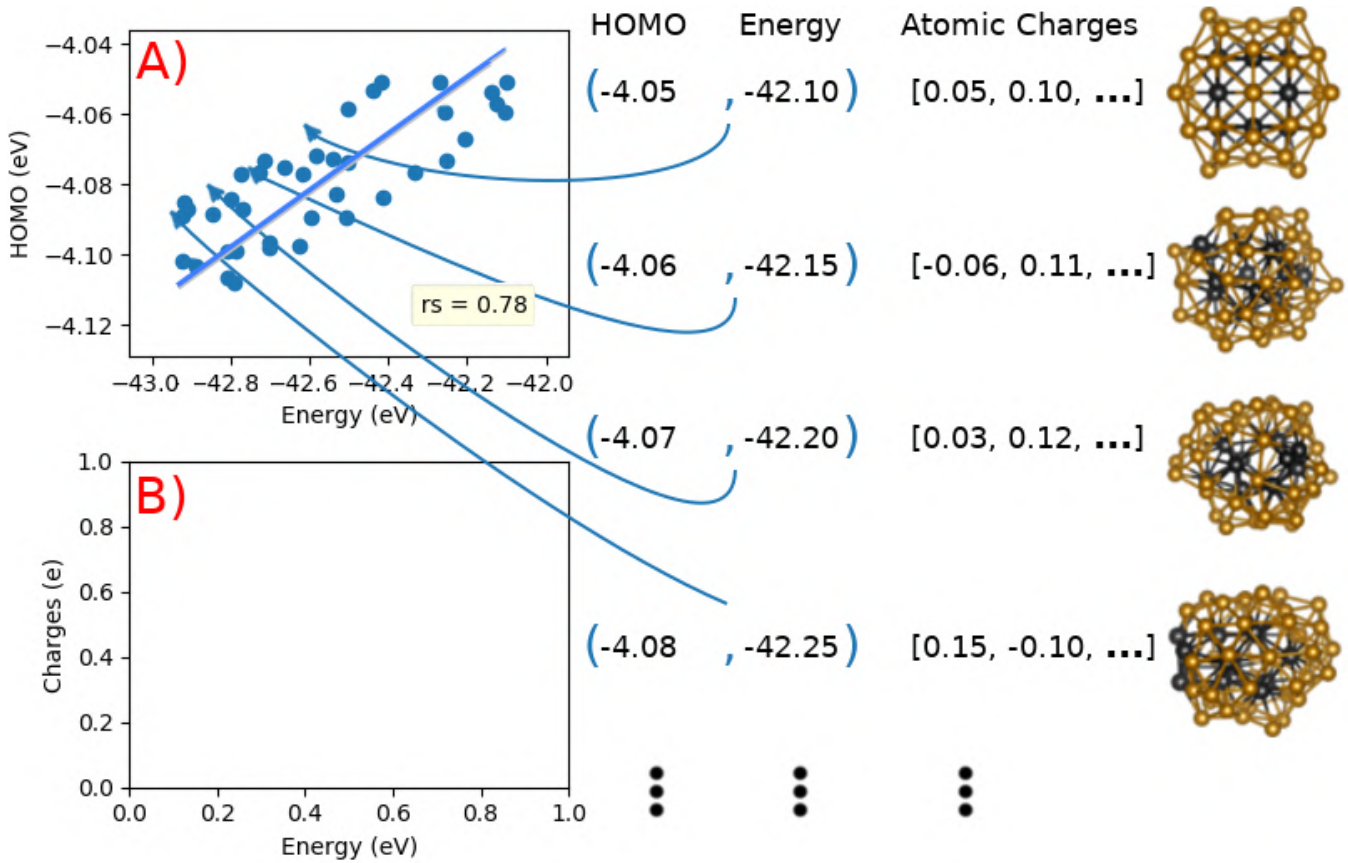
Types of Features with Physical Meaning

Molecular info (single value):

- Energy;
- HOMO.

Atomic info (n values):

- Effective Coordination Number;
- Atomic Charges.



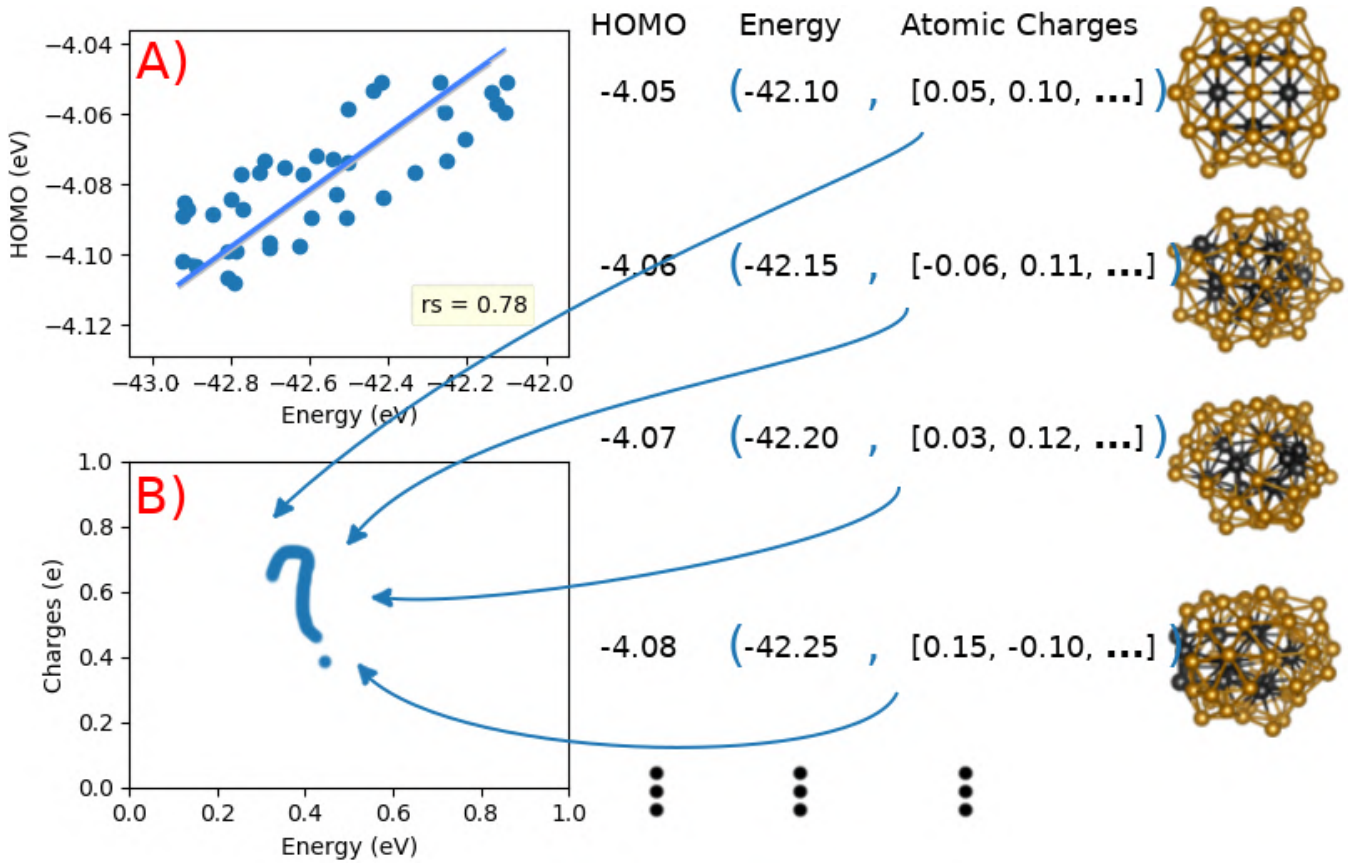
Types of Features with Physical Meaning

Molecular info (single value):

- Energy;
- HOMO.

Atomic info (n values):

- Effective Coordination Number;
- Atomic Charges.



Our Feature Engineering Process (Challenge)

Our process get values (operator) that describe the physical properties (bag) of similar atoms (class), **keeping the physical meaningful**.

Bag: A array with atomic data.

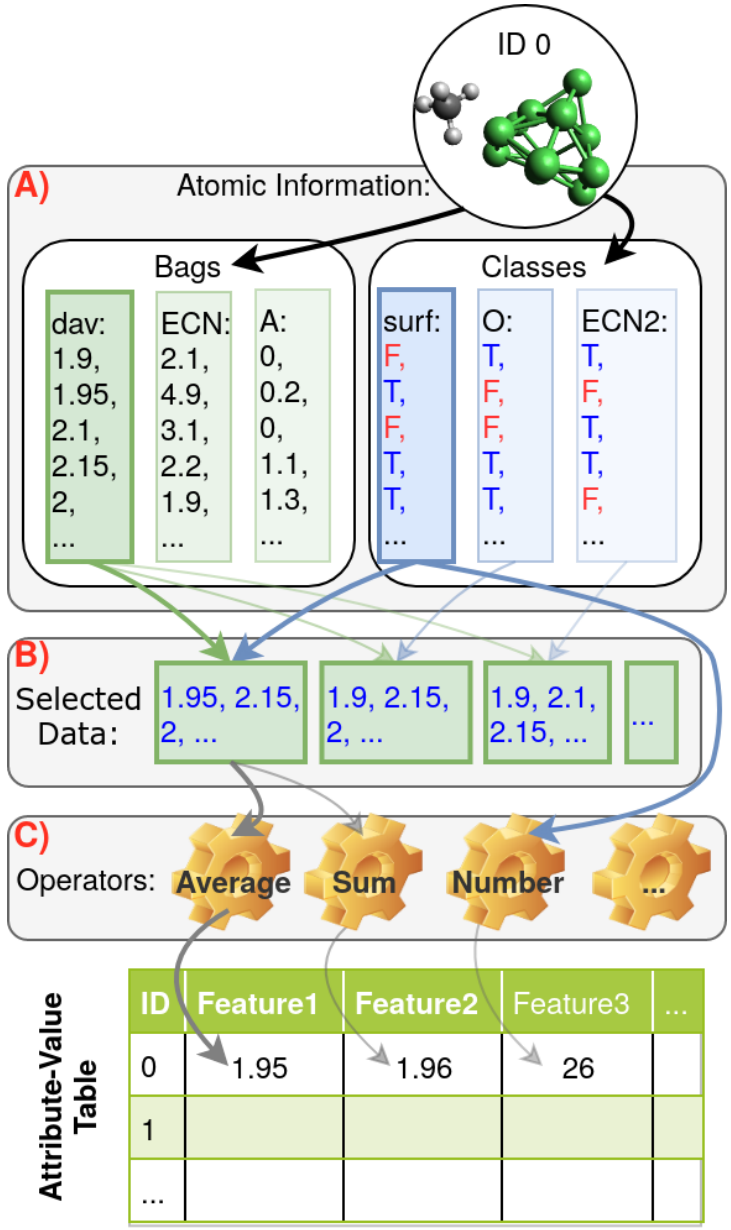
Class: A set of atoms with a similar characteristic.

Operator: A function that gets an array and returns a number.

Classes:
- Elements;
- Surface/Core;
- *ECN*-based.

Bags:
- *ECN*;
- d_{av};
- Charges;
- μ (PAW);
- A.

Operators:
- Average;
- Sum.



Finding Trends with Correlation Analysis

Correlation Coefficients: Pearson ρ, Spearman r_s, Kendall τ

$$\rho = \frac{cov(x, y)}{\sigma(x)\sigma(y)};$$

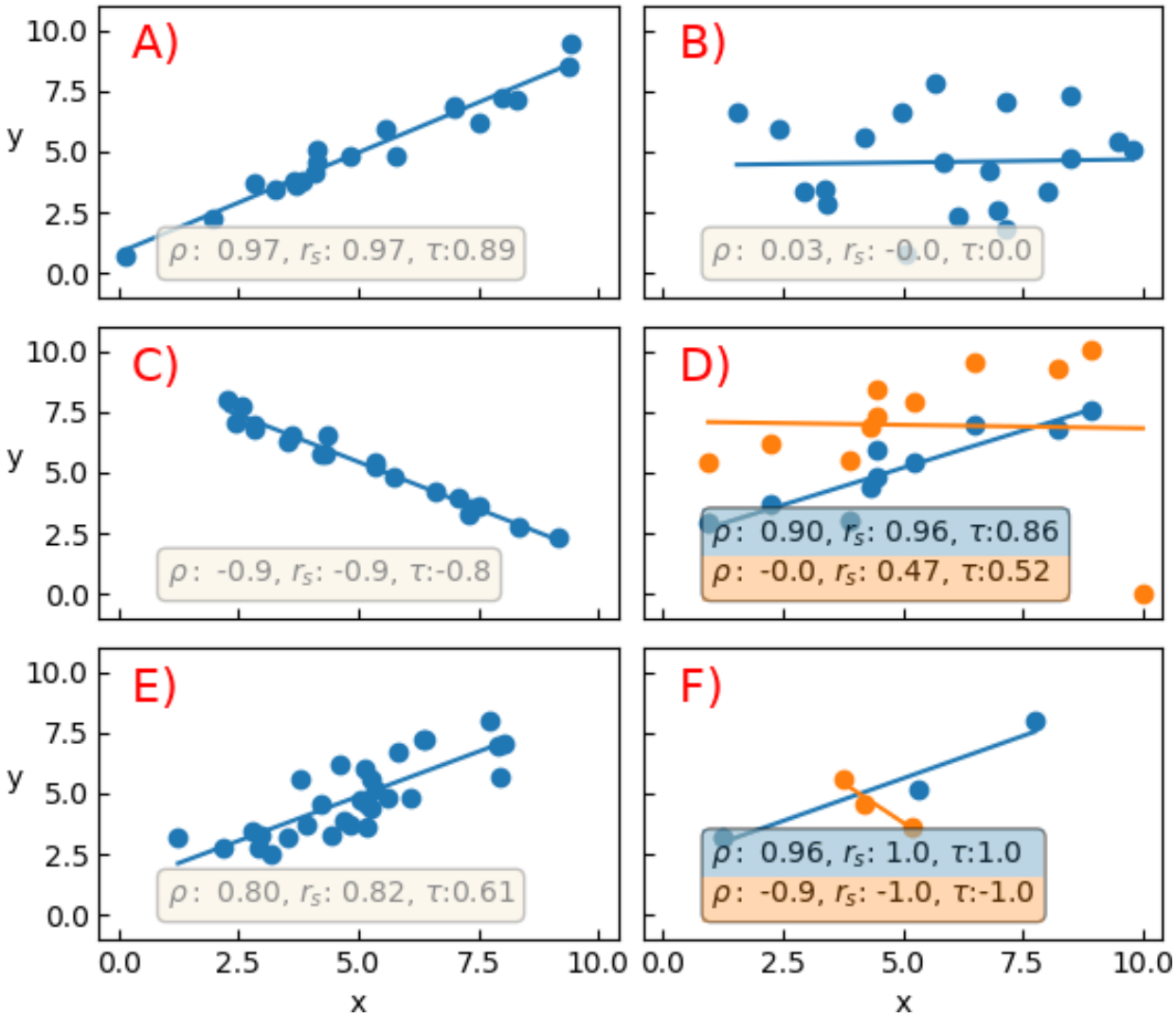
$$r_s = \frac{cov(r_x, r_y)}{\sigma(r_x)\sigma(r_y)};$$

(r_x, r_y: ranked x and y data)

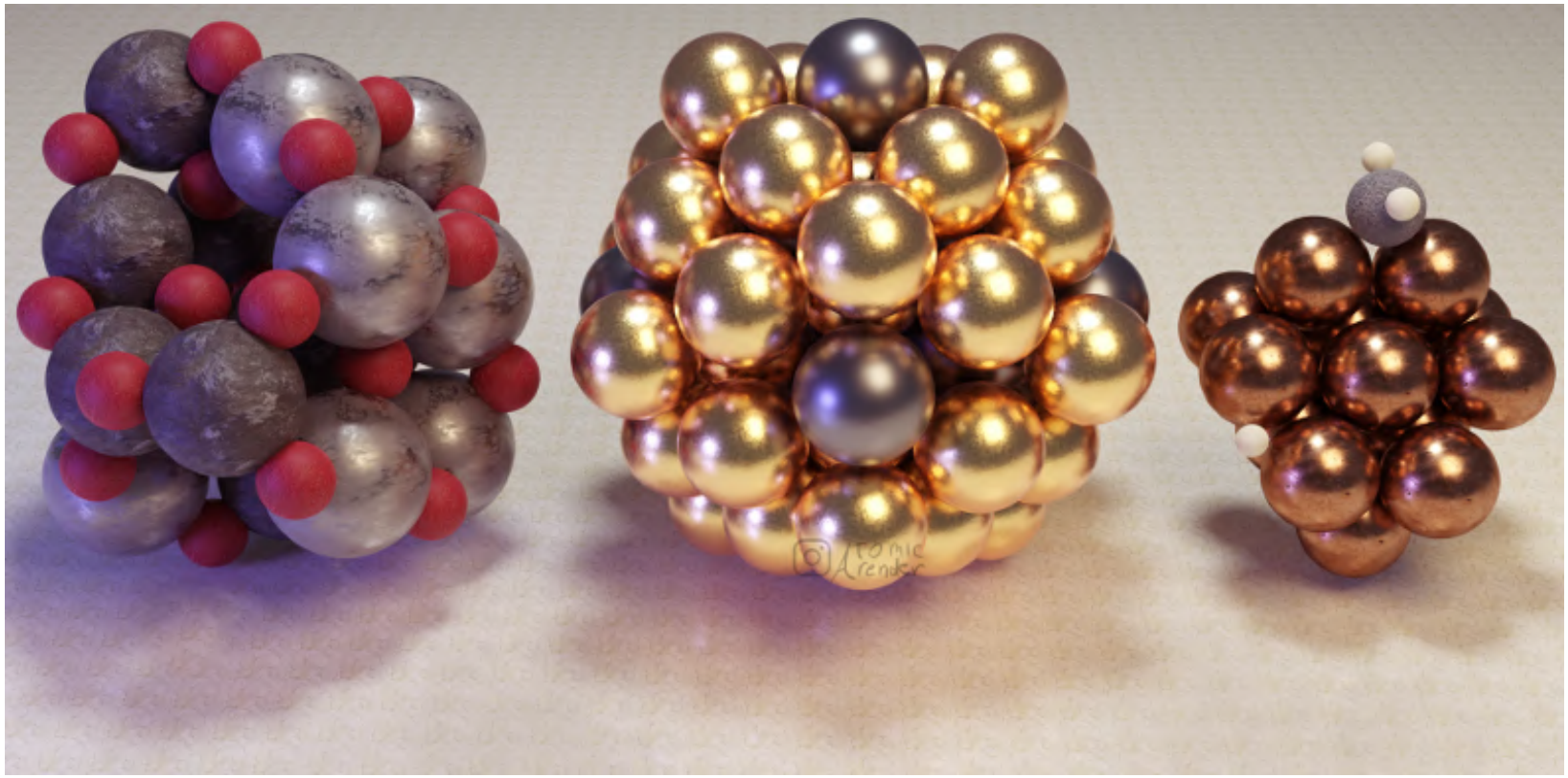
$$\tau = \sum_{i>j} \frac{\text{sign}(x_i - x_j)\text{sign}(y_i - y_j)}{n(n-1)/2}$$

Outliers: r_s and τ better than ρ

Correlation Significance: Bootstrap approaches (null and alternative hypothesis)



Materials Investigated - QC datasets



$Zr_nCe_{15-n}O_{30}$

- 1646 samples.
- $0 \leq n \leq 15$
- Target: Relative Energy

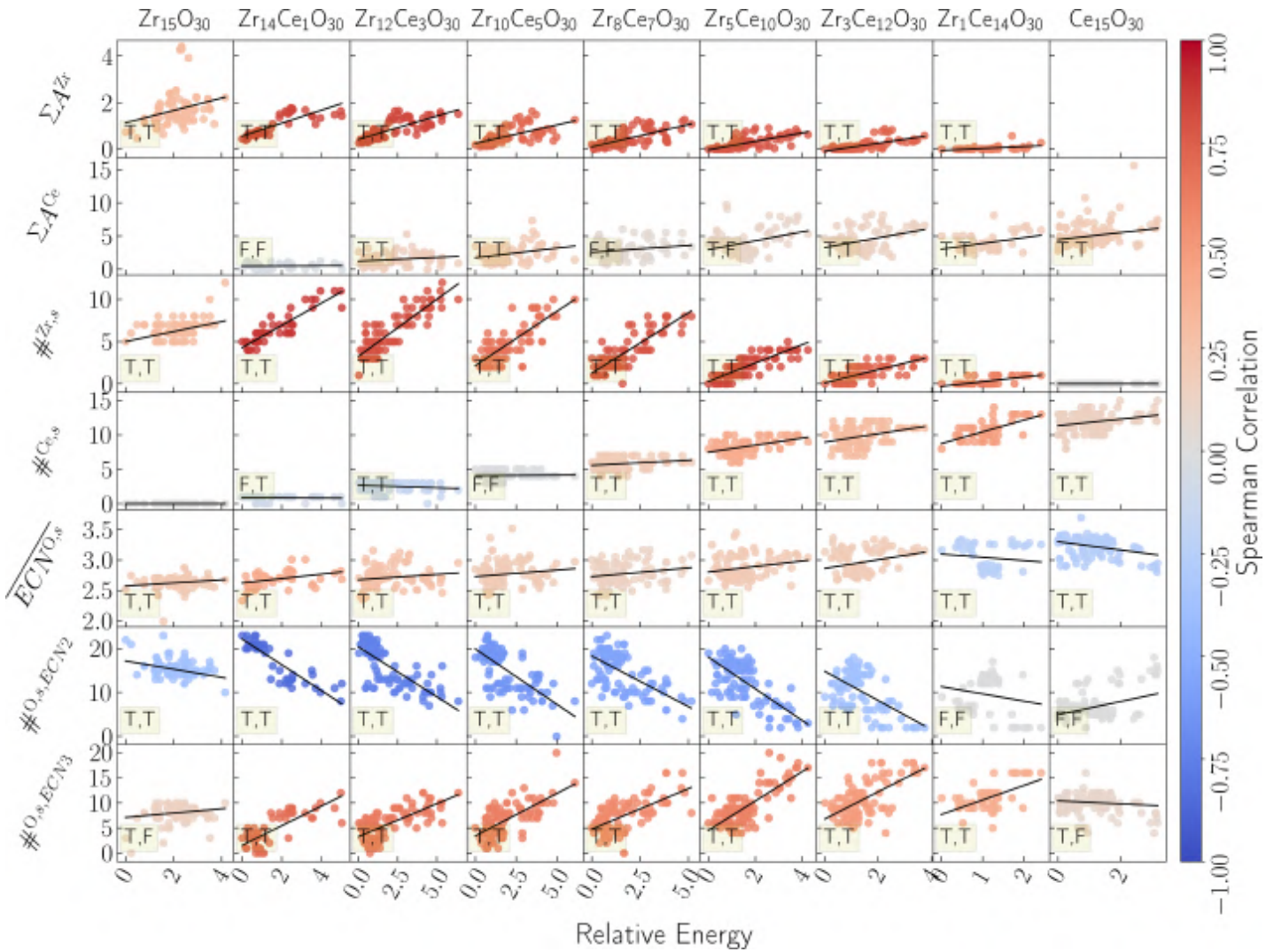
Pt_nTM_{55-n}

- 330 samples;
- TM = Fe, Co, Ni, Cu, Ru, Rh, Pd, Ag, Os, Ir, Au; $n = 13, 42$
- Target: Excess Energy

$(CH_n+mH)/TM_{13}$

- 770 samples.
- TM = Fe, Co, Ni, Cu. $0 \leq n \leq 4$. $m = 0, 4 - n$
- Target: Adsorption Energy

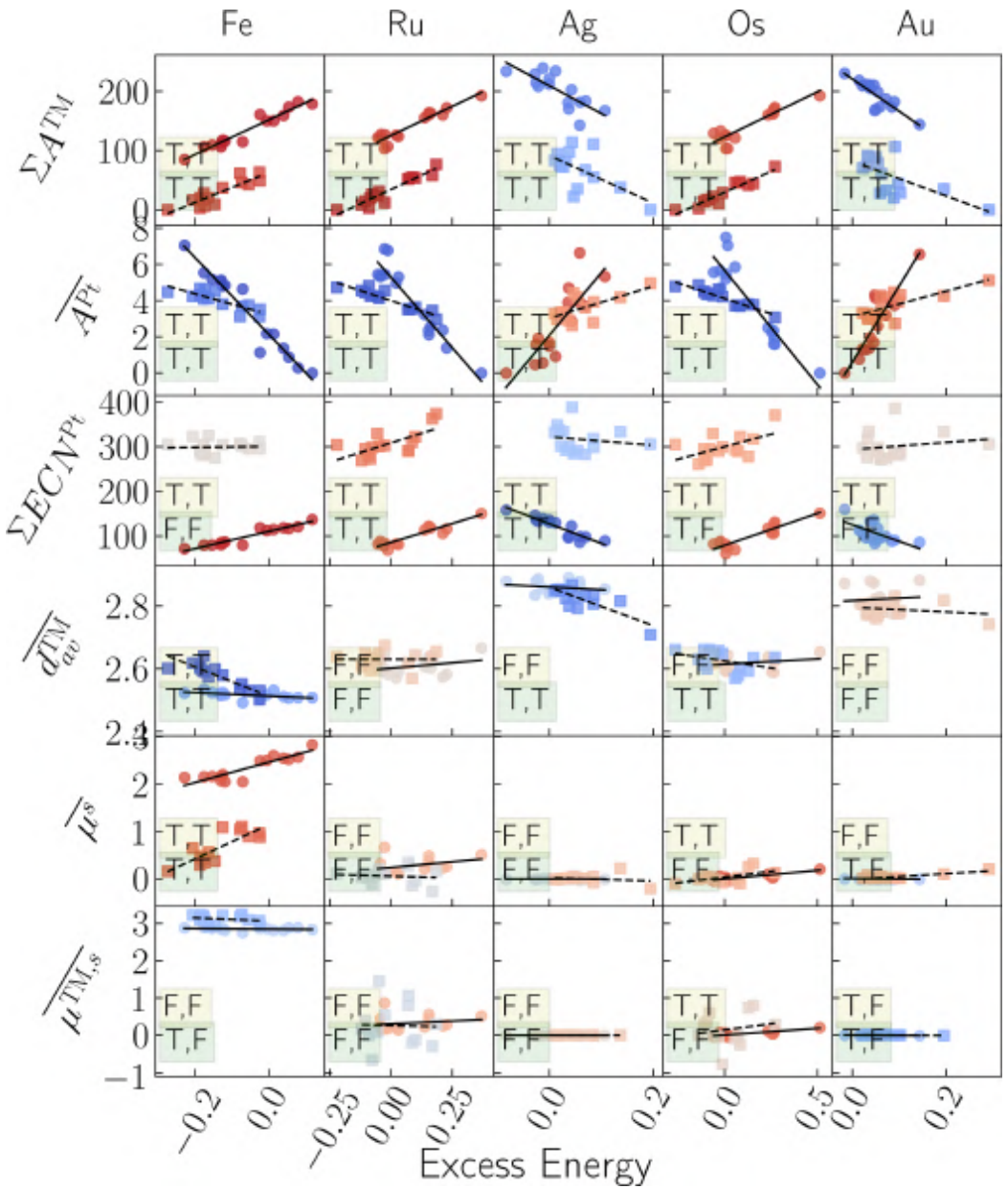
Correlation in $Zr_nCe_{15-n}O_{30}$ dataset



Article: *Phys. Chem. Chem. Phys.*, **2019**, 21, 26637-26646.

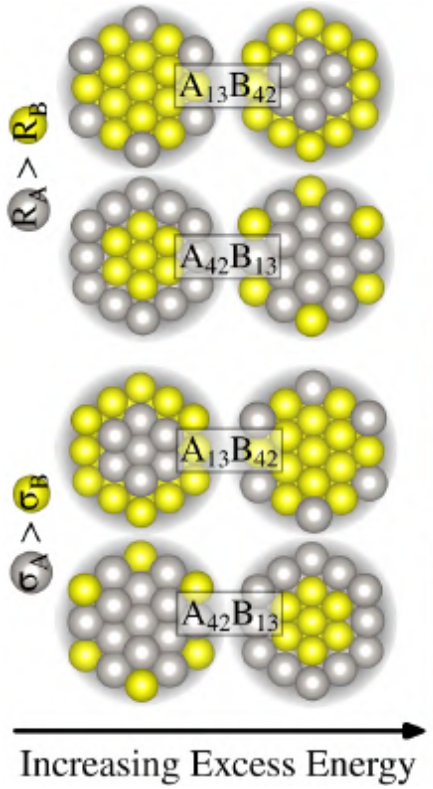
- Many correlations are significant;
- Both Ce *vs* Zr prefer core sites;
- Zr trend is stronger;
- Trends differ for O wich ECN=2, 3;
- Structures with Zr prefer surface O with ECN=2.

Correlation in Pt_nTM_{55-n} dataset



Article: *J. Phys. Chem.*, **2020**, 124, 1, 1158-1164.

- Many correlations are significant;
- TM *vs* Pt sites preference (A,ECN,d_{av});
- Influence in μ.



R: Radius = $d_{av}^{bulk}/2$

σ: Surface Energy

x: Electronegativity

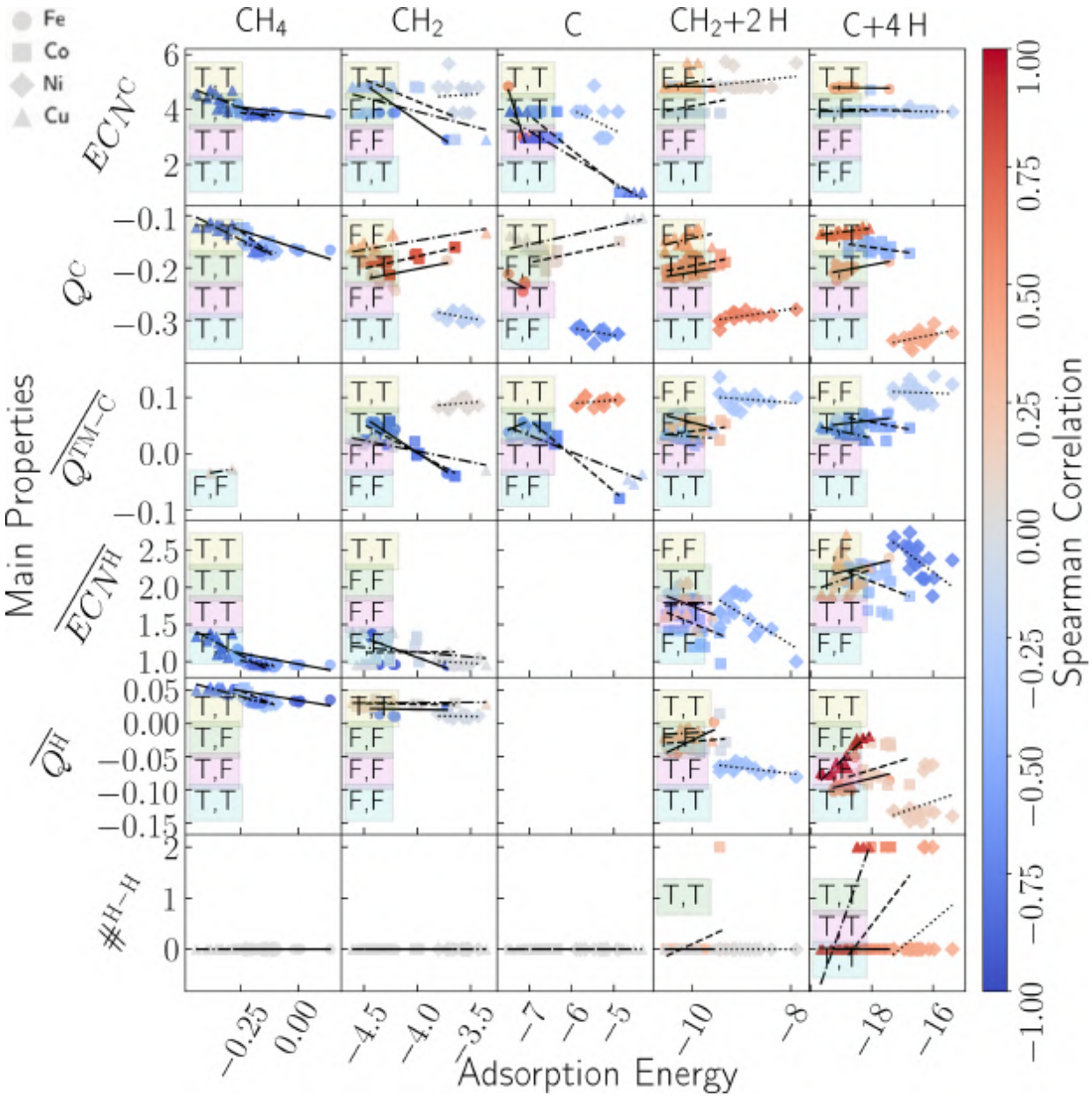
R(Å)

Fe 1.26	Co 1.24	Ni 1.24	Cu 1.28
Ru 1.34	Rh 1.35	Pd 1.39	Ag 1.47
Os 1.36	Ir 1.37	**Pt 1.40**	Au 1.47

σ (eV/atom)

Fe 0.88	Co 0.71	Ni 0.65	Cu 0.47
Ru 1.05	Rh 0.81	Pd 0.56	Ag 0.33
Os 1.21	Ir 0.90	**Pt 0.64**	Au 0.32

Correlation in $(CH_n+mH)/TM_{13}$ dataset



Article: *Fuel*, **2020**, 275, 117790.

- Dehydrogeneted CH_n do more C−TM;
- Charge: TM→C,H (H co-adsorption increase);

- Trends with small significance;
- Data distribution problems;
- Trends change irregularly along with the systems;

- Different built of structures;
- Adsorption Energy sensitively.

Implementation - Quandarium

Fully recursive! (partially parallelized)

- Find Calculation → Extract Info → Molecular Analysis → Featurization → Plots
- Find Calculation → Extract Info → Molecular Analysis → Save Data
- Read Data → Featurization → Save Data
- Read Data → Plots

Find Calculation:

- Search calculations folders.

Extract Info:

•Energy; •State Energies;
•Positions; •Chemical Species;
•Charges, •....

Molecular Analysis:

•ECN; •d_{av}; •Connectivy;
•Surf_or_core; •Site_geometry;
•Number_of_conections.

Featurization:

- bag ↔ class
- class1, class2 → class3
- bag1 → bag2
- bag1, bag2, ... → bag3
- bag[class] → molecular_data
- bag → molecular_data
- class → molecular_data

Plots:

- Scatternplot with correlations;
- Bag histograms;

Other features:

- Python;
- Based on Pandas;
- Fast load/save data (json files);

Conclusions and Perspectives

- A New Data Mining Framework;
  - Easy to employ;
  - Quantitative trends;
  - Any material;
  - Very little explored;
- Contributions for Three Works;
  - $Zr_nCe_{15-n}O_{30}$;
  - Pt_nTM_{55-n};
  - $(CH_n+mH)/TM_{13}$;
- Quandarium (Implementation);
  - Data Extraction;
  - Molecular Analysis;
  - Featurization Process;
  - Correlation and Bootstrap.

Conclusions and Perspectives

- A New Data Mining Framework;
  - Easy to employ;
  - Quantitative trends;
  - Any material;
  - Very little explored;
- Contributions for Three Works;
  - $Zr_nCe_{15-n}O_{30}$;
  - Pt_nTM_{55-n};
  - $(CH_n+mH)/TM_{13}$;
- Quandarium (Implementation);
  - Data Extraction;
  - Molecular Analysis;
  - Featurization Process;
  - Correlation and Bootstrap.

Perspectives:

- Article (methodology);
- Release Quandarium.

Conclusions and Perspectives

- A New Data Mining Framework;
  - Easy to employ;
  - Quantitative trends;
  - Any material;
  - Very little explored;
- Contributions for Three Works;
  - $Zr_nCe_{15-n}O_{30}$;
  - Pt_nTM_{55-n};
  - $(CH_n+mH)/TM_{13}$;
- Quandarium (Implementation);
  - Data Extraction;
  - Molecular Analysis;
  - Featurization Process;
  - Featurization Process;
  - Correlation and Bootstrap.

Perspectives:

- Article (methodology);
- Release Quandarium.

Acknowledgments

CINE Center for Innovation on New Energies
QTNano
IQSC
CNPq
CAPES
FAPESP
UNIFESP
UFABC

Thanks for your Attention!