3rd Internal CINE-CMSC Workshop

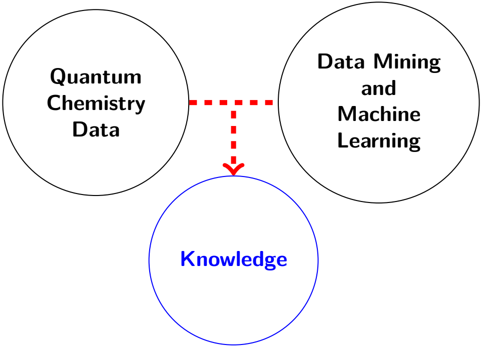
Data Mining Tools Applied to Quantum Chemistry Data

Speaker:
Johnatan Mucelini

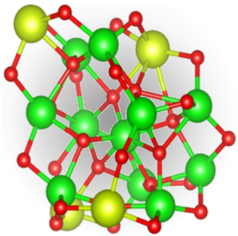
Supervisor:
Prof. Dr. Juarez L. F. Da Silva

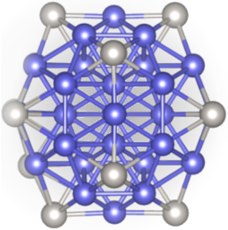
QC Collaborators:
Paulo de Carvalho Dias Mendes,
Priscilla Felício-Sousa,
Karla F. Andriani.

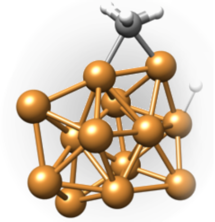
DM/ML Collaborators:
Marcos G. Quiles,
Ronaldo C. Prati.



Introduction - Materials







$Zr_qCe_{15\text{-}q}O_{30}$

•Employed in TWC and candidate for others process.
•Explore material in nanoparticles structures.

$Pt_lTM_{55\text{-}l}$

•Pt-based Catalysts are wildly employed.
•Explore this alloys structural preferences.

$CH_m/xH/TM_{13}$

•CH_4 dehydrogenation candidates.
•Reaction intermediates study.

Motivation:

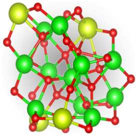
- Develop the chem/comp overlap;
- Develop material science area.

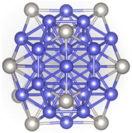
Objectives:

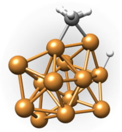
- Find new patterns;
- Initial Studies;
- Tools Development.

Data Mining - Challenge

Calculations Sets:







	$Zr_qCe_{15\text{-}q}O_{30}$	$Pt_lTM_{55\text{-}l}$	$CH_m/xH/TM_{13}$
n	1600	330	700
vars	q (16)	l (2)	(m, x) (9)
		TM (9)	TM (4)

Initial Data:

Molecular info (structured data):
- Energy [-42.0 eV];
- HOMO [-4.123 eV];

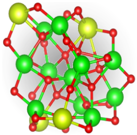
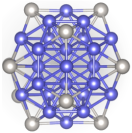
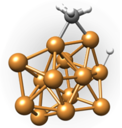
Atomic info (attribute-vector):
- Exposed to vacuum [True, False, ...];
- Atomic Charges [0.7, -0.8, ...].

DM/ML Input:

attribute-value table	feature0	feature1	...
sample0	data00	data01	...
sample1	data10	data11	...
...	...	...	...

Data Mining - Challenge

Calculations Sets:

	$Zr_qCe_{15\text{-}q}O_{30}$	$Pt_l/TM_{55\text{-}l}$	$CH_m/xH/TM_{13}$
n	1600	330	700
vars	q (16)	l (2)	(m, x) (9)
		TM (9)	TM (4)

Quandarium (python):

- Find Calc.
- Extract Info.
- Geometical Analysis:
- ECN/d_{av}
- Find surf. atoms
- Connections Analysis

It operate recursively!

Initial Data:

Molecular info (structured data):
- Energy [-42.0 eV];
- HOMO [-4.123 eV];

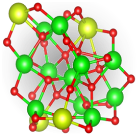
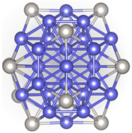
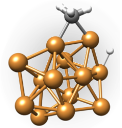
Atomic info (attribute-vector):
- Exposed to vacuum [True, False, ...];
- Atomic Charges [0.7, -0.8, ...].

DM/ML Input:

attribute-value table	feature0	feature1	...
sample0	data00	data01	...
sample1	data10	data11	...
...	...	...	...

Data Mining - Challenge

Calculations Sets:

	$Zr_qCe_{15\text{-}q}O_{30}$	$Pt_l/TM_{55\text{-}l}$	$CH_m/xH/TM_{13}$
n	1600	330	700
vars	q (16)	l (2)	(m, x) (9)
		TM (9)	TM (4)

Quandarium (python):

- Find Calc.
- Extract Info.
- Geometical Analysis:
- Find surf. atoms
- Connections Analysis
- ECN/d_{av}

It operate recursively!

Initial Data:

Molecular info (structured data):

- Energy [-42.0 eV];
- HOMO [-4.123 eV];

Atomic info (attribute-vector):

- Exposed to vacuum [True, False, ...];
- Atomic Charges [0.7, -0.8, ...].

DM/ML Input:

attribute-value table	feature0	feature1	...
sample0	data00	data01	...
sample1	data10	data11	...
...	...	...	...

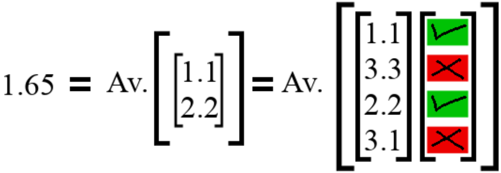
How to get molecular data from atomic data?

Data Mining

How to get molecular data from atomic data?

Take **molecular data** from **operator** over a **bag** (properties), for all or a **class** of atoms.

$$Molecular\_Data = \underbrace{OPERATOR}_{\text{average}}\left[\underbrace{BAG}_{\text{ECN}}\left[\underbrace{CLASS}_{\text{Pt}}\right]\right]$$



Operator: Operates over one or more arguments, and return a number (Ex.: sum).

Classes: Set of atoms that meet a condition!

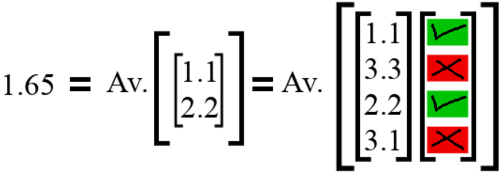
- Ex.: O, O exposed to the vacuum, O exposed to the vacuum with $1 < ECN < 2$, ...

Data Mining

How to get molecular data from atomic data?

Take **molecular data** from **operator** over a **bag** (properties), for all or a **class** of atoms.

$$Molecular\_Data = \underbrace{OPERATOR}_{\text{average}} \left[\underbrace{BAG}_{\text{ECN}} \left[\underbrace{CLASS}_{\text{Pt}} \right] \right]$$



Operator: Operates over one or more arguments, and return a number (Ex.: sum).

Classes: Set of atoms that meet a condition!

- Ex.: O, O exposed to the vacuum, O exposed to the vacuum with $1 < ECN < 2$, ...

Quandarium (python):

Flexible data manipulation:

- bag $\rightarrow$ class
- class1 + class2 $\rightarrow$ class3
- class + bag $\rightarrow$ class
- class $\rightarrow$ bag

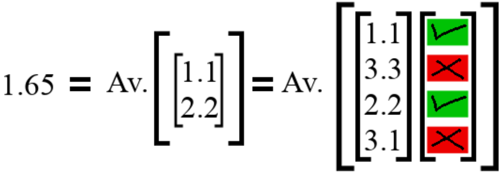
It operate recursively!

Data Mining

How to get molecular data from atomic data?

Take **molecular data** from **operator** over a **bag** (properties), for all or a **class** of atoms.

$$Molecular\_Data = \underbrace{OPERATOR}_{\text{average}}\left[\underbrace{BAG}_{\text{ECN}}\left[\underbrace{CLASS}_{\text{Pt}}\right]\right]$$



Operator: Operates over one or more arguments, and return a number (Ex.: sum).

Classes: Set of atoms that meet a condition!

- Ex.: O, O exposed to the vacuum, O exposed to the vacuum with $1 < ECN < 2$, ...

Quandarium (python):

Flexible data manipulation:

- bag → class
- class1 + class2 → class3
- class + bag → class
- class → bag

It operate recursively!

Thus we access many molecular data (attribute-value table). Lets **analyse** the data!

Correlation Analysis

Pearson

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

- Non-ranked data;
- Sensitive to outliers.

Spearman

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

- Ranked data;
- Robust to outliers.

Correlation Interpretation:

- $-1 \geq r \geq 1$
- If X increase, Y is **expected** to go:

  downward

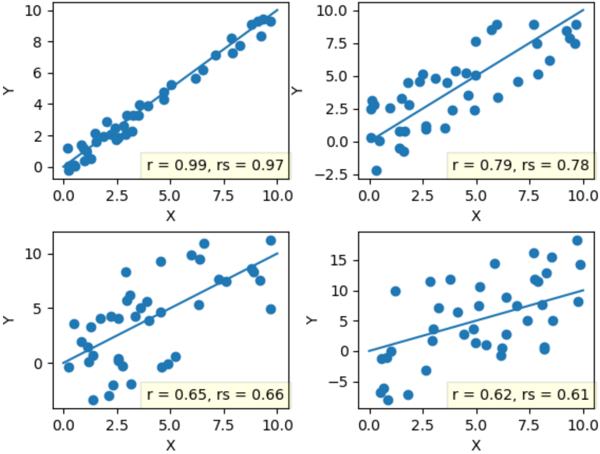
  $r < 0 < r$

  upward

- How much expected?

  "How strong correlated?"

  "As large as was $|r|$."



Correlation Analysis

Pearson

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

- Non-ranked data;
- Sensitive to outliers.

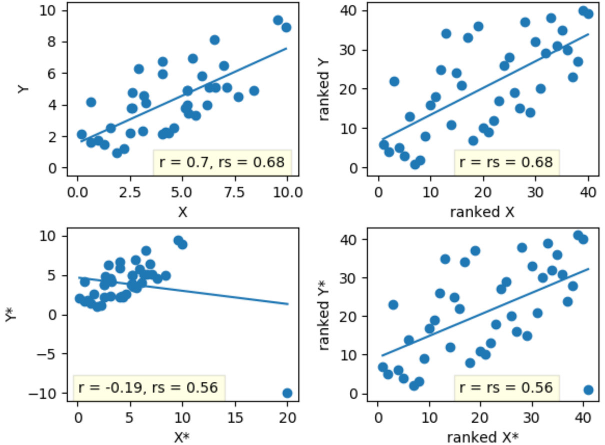
Spearman

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

- Ranked data;
- Robust to outliers.

An example of outlier effect:

- r reduced 0.89;
- r_s reduced 0.12;



Correlation Analysis

Pearson

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

- Non-ranked data;
- Sensitive to outliers.

Spearman

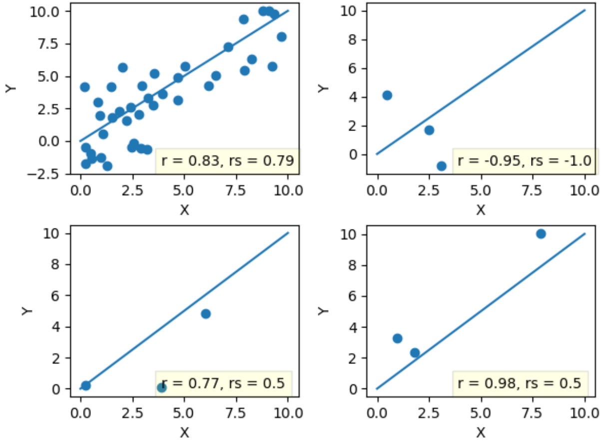
$$r_s = \frac{cov(r_x, r_y)}{\sigma(r_x)\sigma(r_y)}$$

- Ranked data;
- Robust to outliers.

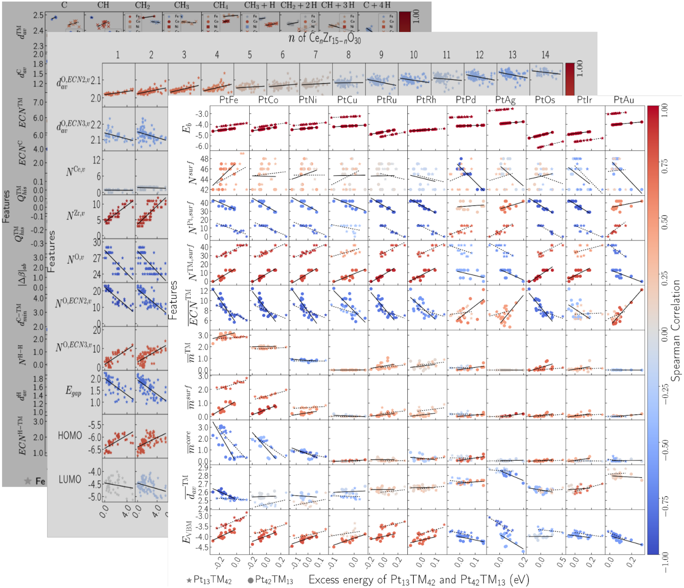
- Should I trust large correlation?

  Depend on data **size** and **distribution**.

  Good practice: Hypothesis test (Bootstrap) and pvalue.



Data Representation



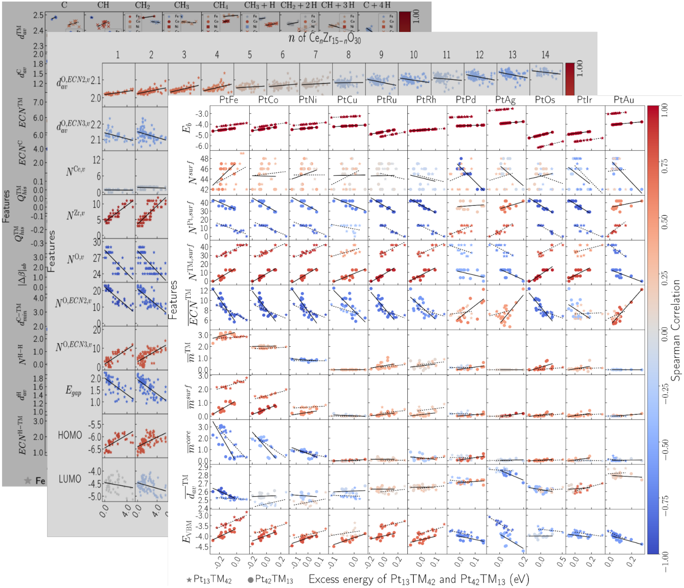
Scatter-plot Matrix:

- Rows: Features;
- Columns: Dataset part.

Cell:

- Scatter-plot;
- Y axis: Feature;
- X axis: Energy;
- Correlation: Colors;
- Linear Model.

Data Representation



Scatter-plot Matrix:

- Rows: Features;
- Columns: Dataset part.

Cell:

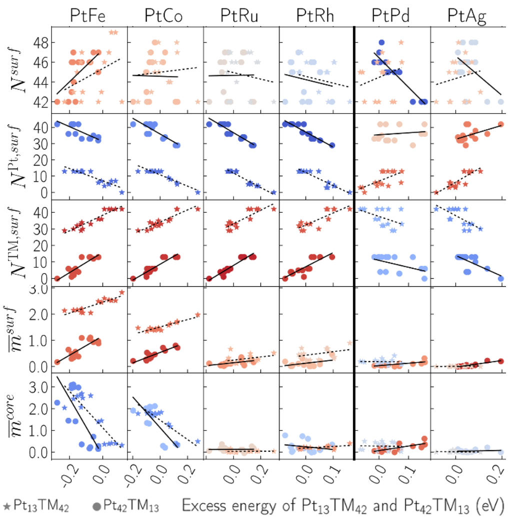
- Scatter-plot;
- Y axis: Feature;
- X axis: Energy;
- Correlation: Colors;
- Linear Model.

Quandarium:

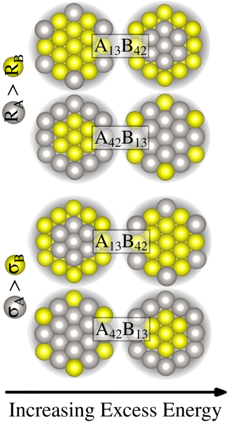
- Scatternplot
- bag histograms

Results

$Pt_{13}TM_{42}$ and $Pt_{42}TM_{13}$



- High correlation! Few samples.
- TM *vs* Pt sites preference (N, ECN);
- Influence in other properties (m, d_{av});



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

σ: Surface Energy

χ: 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

Conclusion

Quantum Chemistry Data Mining with Correlations:

Analysis Benefits:

- Easy access to useful **chemistry information**;
- **Quantitative trends** analysis;
- Very **little explored**.

Analysis Limitation:

- Small **dataset size**;
- Many **variables** in the study.

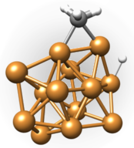
Analysis Applicability:

- Can be applied to **any material**;
- Require small **programming skills**;
- Require some **statistical concepts**.

Perspectives 2019 - 2020

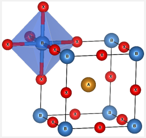
Complete the data mining works in progress!

Molecules over TM



- $CH_n/H_m/TM_{13}$ Dataset (previously presented).

Perovskites



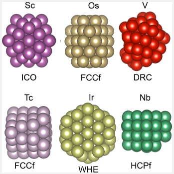
- Solid State Feature Extraction.

Article: Nanocluster DM Analysis

- **DM** and **correlation analysis**;
- For **chemists**;
- 5 QC datasets (3+2);
- Release **Quandarium**.

TM Nanoclusters and Alloys Energy Regression:

TM nanoclusters and alloys:



- Employ several previous QTNano studies (TM_{13}, TM_{55});
- Methodological normalisation;
- Algorithms: MLP, random-forest, kernel regression...

Acknowledgements

Thanks for your Attention!