CAMLC25

From quantum-chemical descriptors to clustering, an automated pipeline

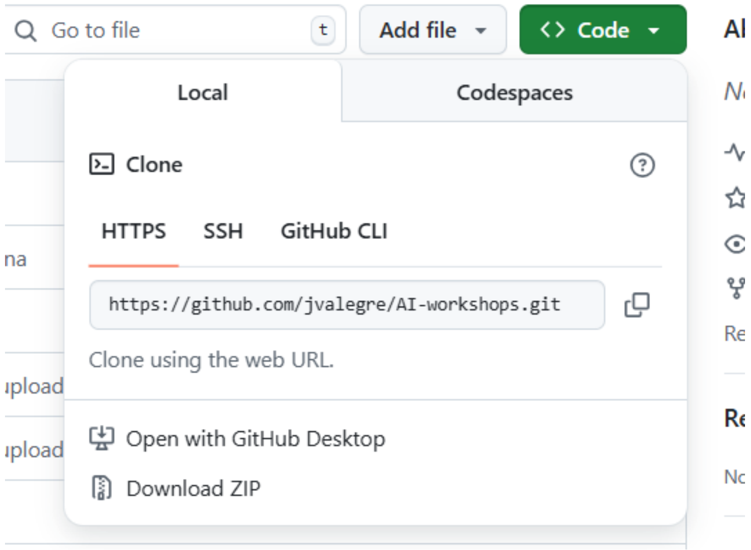
September 17, 2025

Dr. Susana García-Abellán

Set up

Before starting

1. https://github.com/sgabellan/CAMLC25_session6_QMdescp_cluster

2. 

https://github.com/jvalegre/AI-workshops.git

Clone using the web URL.

3. c/Users/your_user/ Documents/ML_course

Set up

For this session:

1. Open the Ubuntu terminal and type:

```
conda activate cheminf
pip install almos-kit
```

CAMLC25

Instituto de Síntesis Química y Catálisis Homogénea

From quantum-chemical descriptors to clustering, an automated pipeline

September 17, 2025

Dr. Susana García-Abellán

Case study – Automation of QM and clustering

1. Define the research framework and role of ML

2. Dataset preparation

3. Digital representation of molecules (featurization)

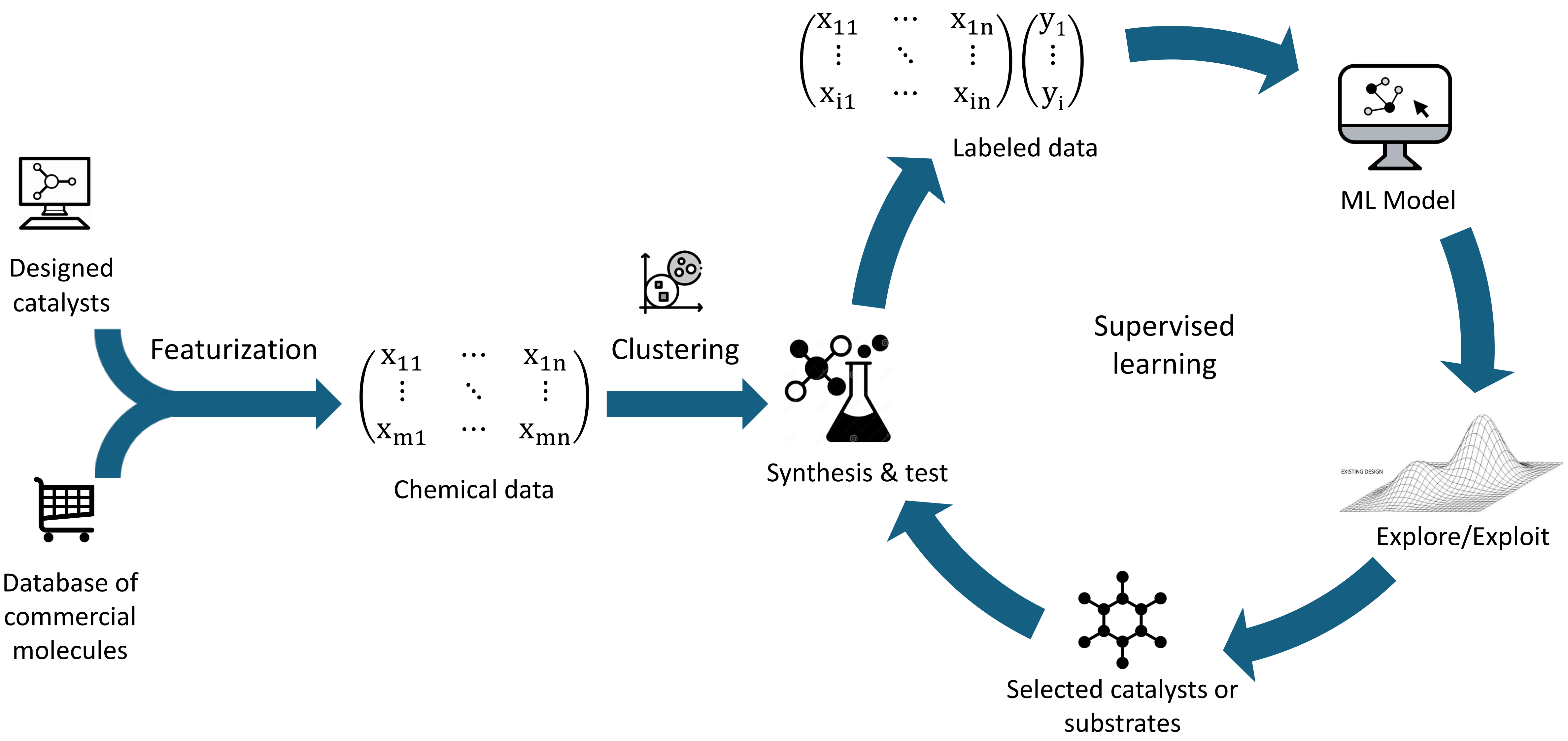
   AQME: automation of QM protocols

4. Unsupervised learning: clustering

   ALMOS: automation of clustering

5. Supervised learning: active learning

Workflow



Designed catalysts

Database of commercial molecules

Featurization

$$\begin{pmatrix} x_{11} & \cdots & x_{1n} \\ \vdots & \ddots & \vdots \\ x_{m1} & \cdots & x_{mn} \end{pmatrix}$$

Chemical data

Clustering

Synthesis & test

$$\begin{pmatrix} x_{11} & \cdots & x_{1n} \\ \vdots & \ddots & \vdots \\ x_{i1} & \cdots & x_{in} \end{pmatrix} \begin{pmatrix} y_1 \\ \vdots \\ y_i \end{pmatrix}$$

Labeled data

ML Model

Supervised learning

Explore/Exploit

Selected catalysts or substrates

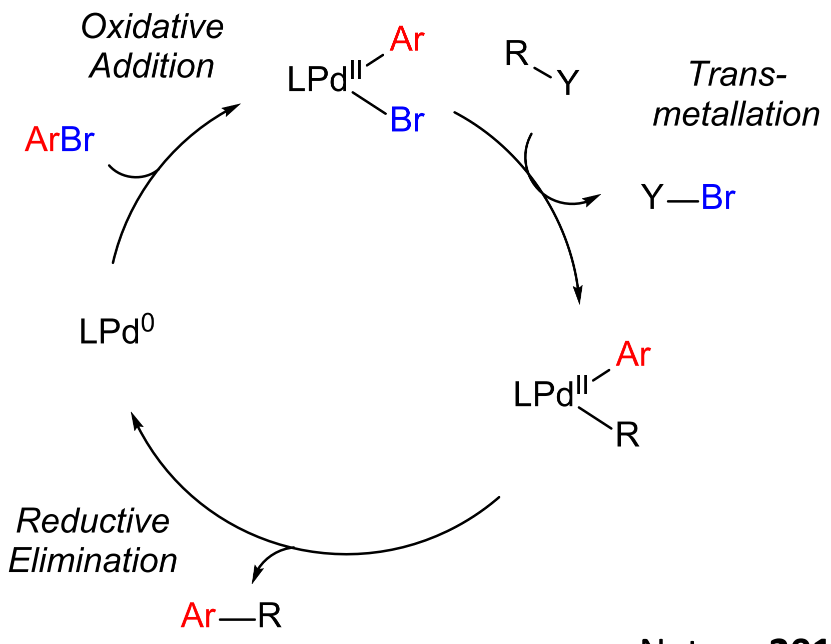
1. Research framework and role of ML

Case study: Cross-coupling of bromoaryl substrates catalysed by Pd-complexes

ML application options:

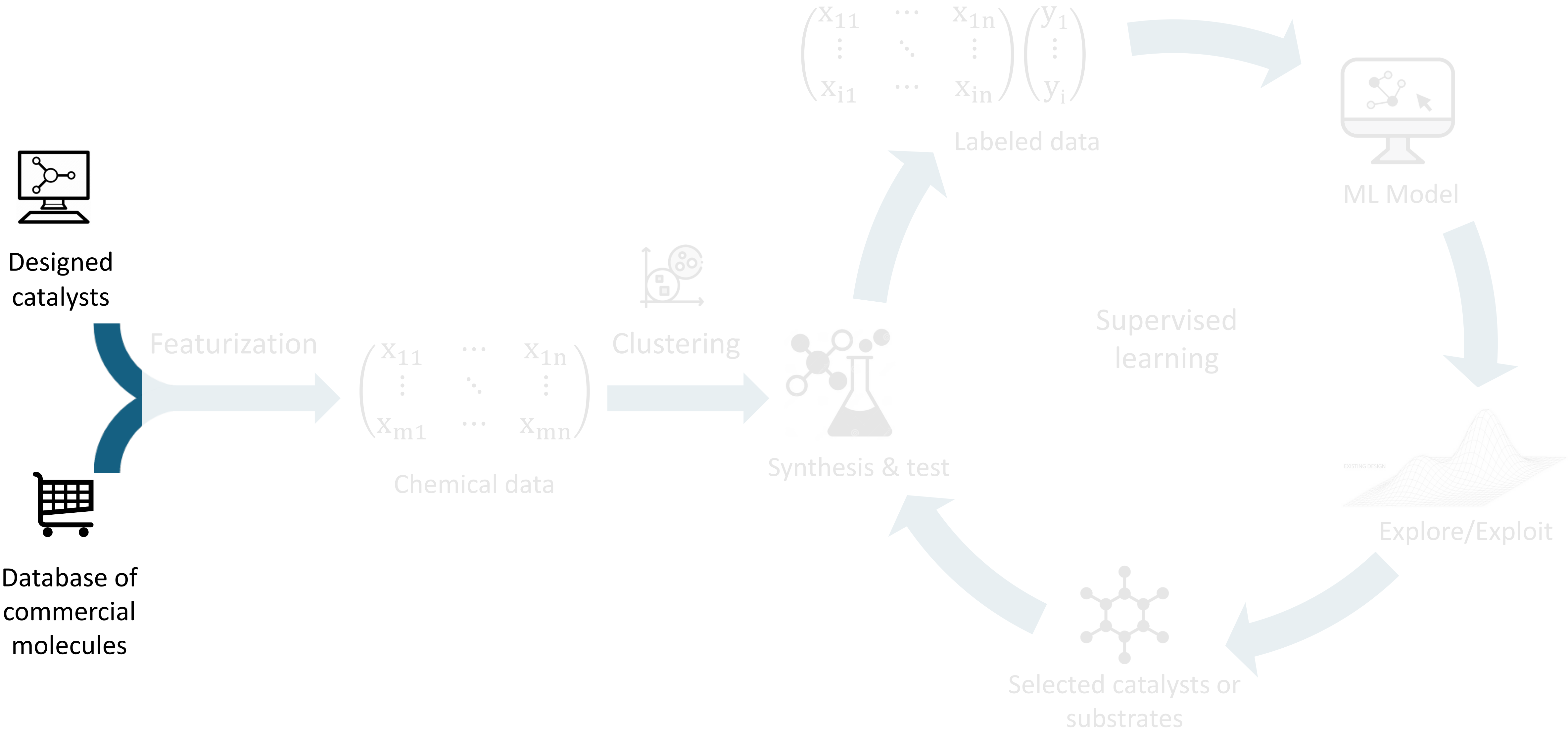
- Selection of catalyst
- Selection of substrates

General mechanism:



Nature **2015**, *524*, 454–457
Science **2018**, *362*, 670–674

2. Dataset preparation



Designed catalysts

Database of commercial molecules

Featurization

$$\begin{pmatrix} x_{11} & \cdots & x_{1n} \\ \vdots & \ddots & \vdots \\ x_{m1} & \cdots & x_{mn} \end{pmatrix}$$

Chemical data

Clustering

Synthesis & test

$$\begin{pmatrix} x_{11} & \cdots & x_{1n} \\ \vdots & \ddots & \vdots \\ x_{i1} & \cdots & x_{in} \end{pmatrix} \begin{pmatrix} y_1 \\ \vdots \\ y_i \end{pmatrix}$$

Labeled data

ML Model

Supervised learning

Explore/Exploit

Selected catalysts or substrates

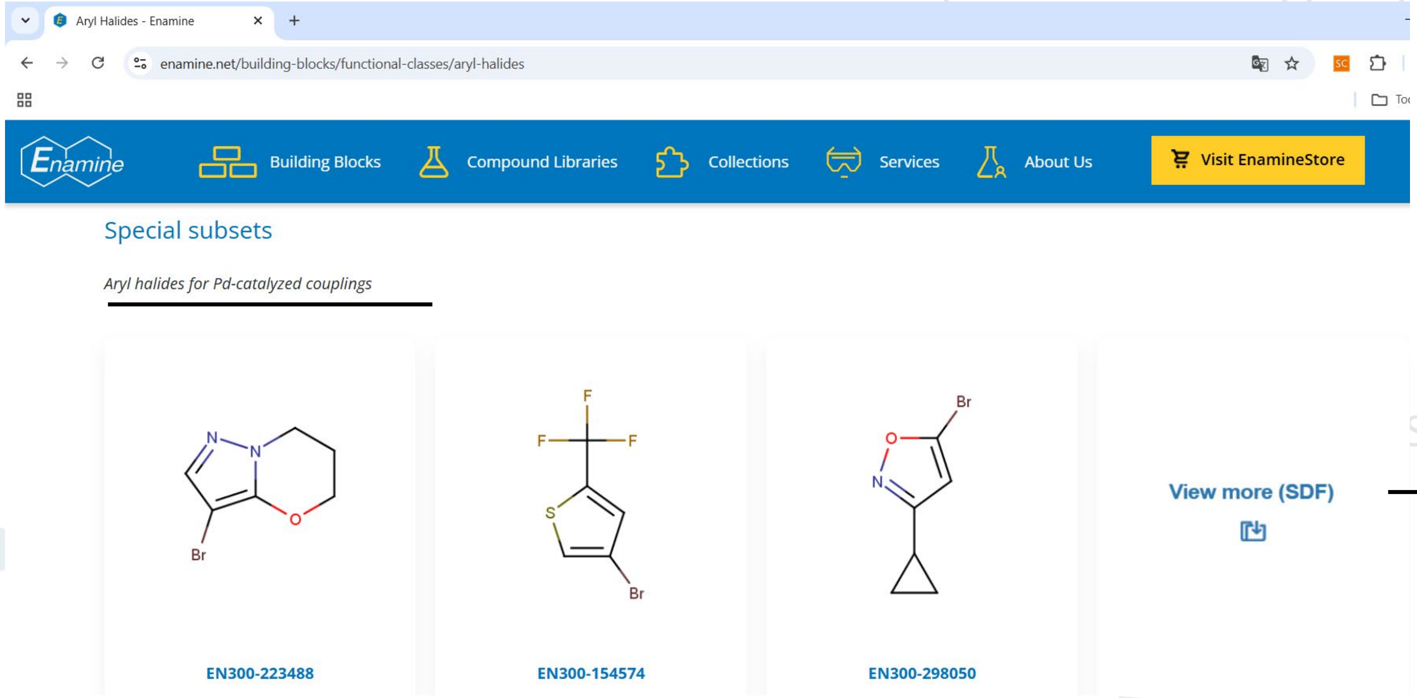
2. Dataset preparation

Instituto de Síntesis Química y Catálisis Homogénea

Designed catalysts

Database of commercial molecules

https://enamine.net/building-blocks/functional-classes/aryl-halides


Aryl Halides - Enamine

enamine.net/building-blocks/functional-classes/aryl-halides

Special subsets

Aryl halides for Pd-catalyzed couplings

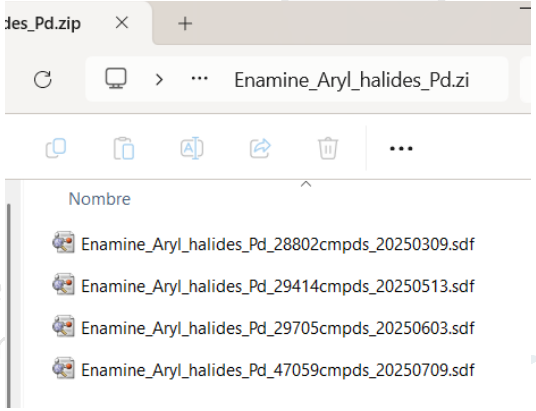
EN300-223488

EN300-154574

EN300-298050

View more (SDF)

SDF files


Enamine_Aryl_halides_Pd.zi

Nombre

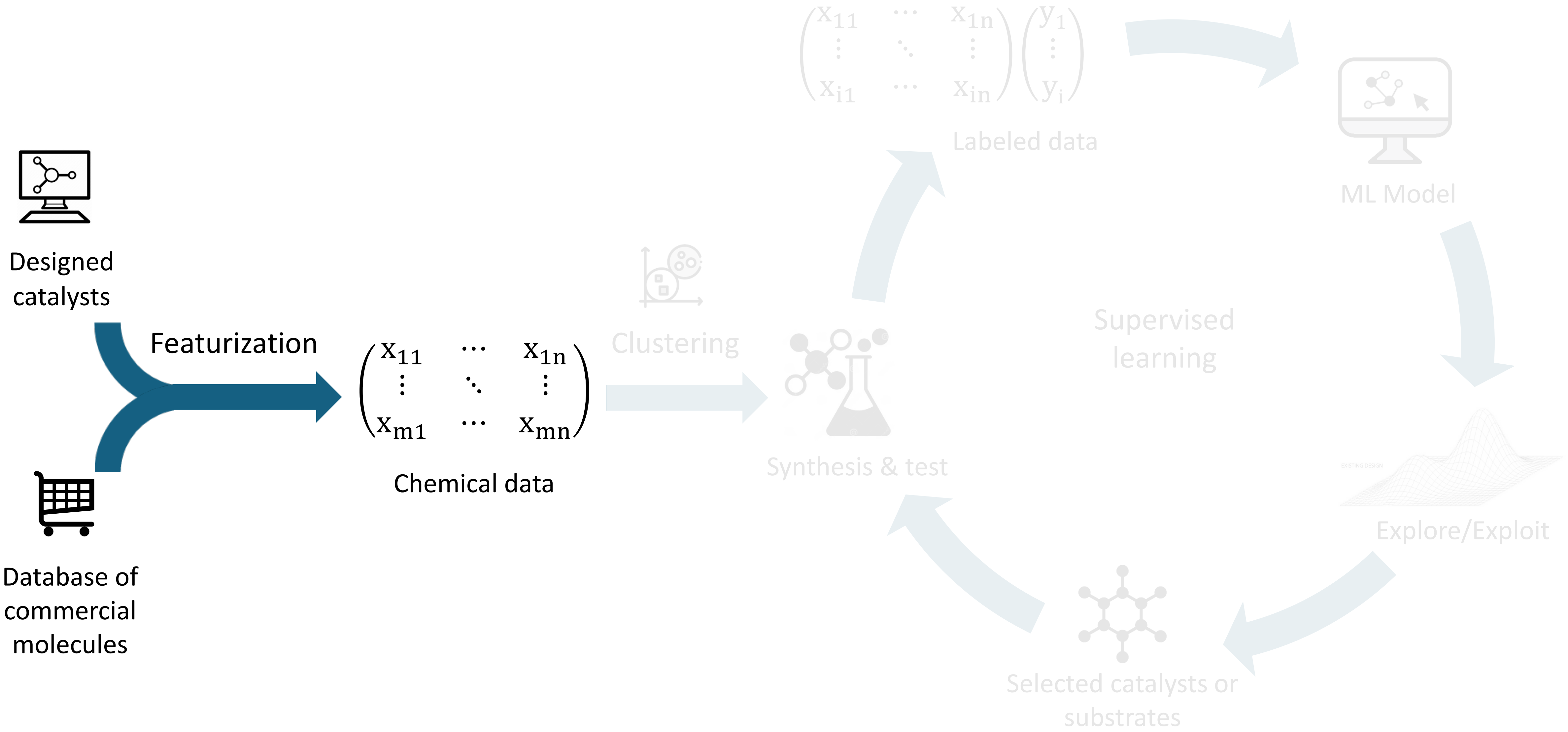
- Enamine_Aryl_halides_Pd_28802cmpds_20250309.sdf
- Enamine_Aryl_halides_Pd_29414cmpds_20250513.sdf
- Enamine_Aryl_halides_Pd_29705cmpds_20250603.sdf
- Enamine_Aryl_halides_Pd_47059cmpds_20250709.sdf

Apply filters (python, rdkit)

- one Br atom
- zero Cl or I atom
- no counterions present

≈ 25,000 molecules

3. Digital representation of molecules, featurization



Designed catalysts

Featurization

$$\begin{pmatrix} x_{11} & \cdots & x_{1n} \\ \vdots & \ddots & \vdots \\ x_{m1} & \cdots & x_{mn} \end{pmatrix}$$

Chemical data

Database of commercial molecules

Clustering

Synthesis & test

$$\begin{pmatrix} x_{11} & \cdots & x_{1n} \\ \vdots & \ddots & \vdots \\ x_{i1} & \cdots & x_{in} \end{pmatrix} \begin{pmatrix} y_1 \\ \vdots \\ y_i \end{pmatrix}$$

Labeled data

ML Model

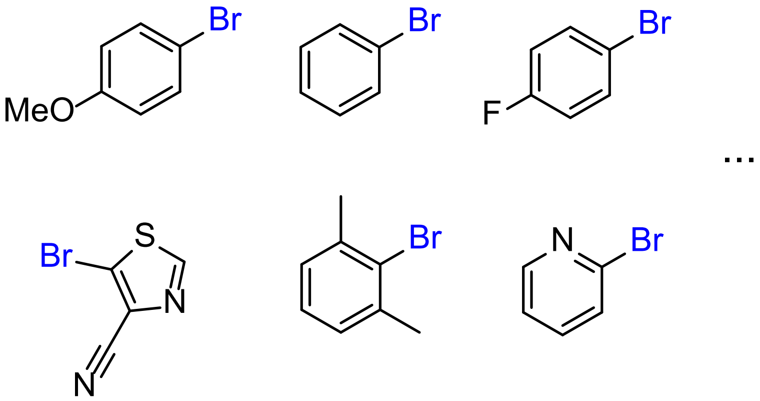
Supervised learning

Explore/Exploit

Selected catalysts or substrates

3. Digital representation of molecules, featurization

Dataset of bromoaryls



$$\begin{array}{c|ccc} & \text{descriptor 1} & \cdots & \text{descriptor n} \\ \hline \text{ArBr 1} & x_{11} & \cdots & x_{1n} \\ \cdots & \vdots & \ddots & \vdots \\ \text{ArBr m} & x_{m1} & \cdots & x_{mn} \end{array}$$

<u>Digitalization of molecules</u>

- Each molecule → *n*D vector
- m molecules → n × m matrix

3. Digital representation of molecules, featurization

- Descriptors are numerical values that capture different aspects of:

Entire molecule (**molecular descriptors**)

Specific atoms within it (**atomic descriptors**)

- Descriptors can be generated in different ways:

Experimental data (boiling point, solubilities, spectroscopic values…)

Cheminformatics tools like RDKit (molecular weight, number of H bond donors…)

Computational data (HOMO-LUMO gaps, charges, dipole moments…)

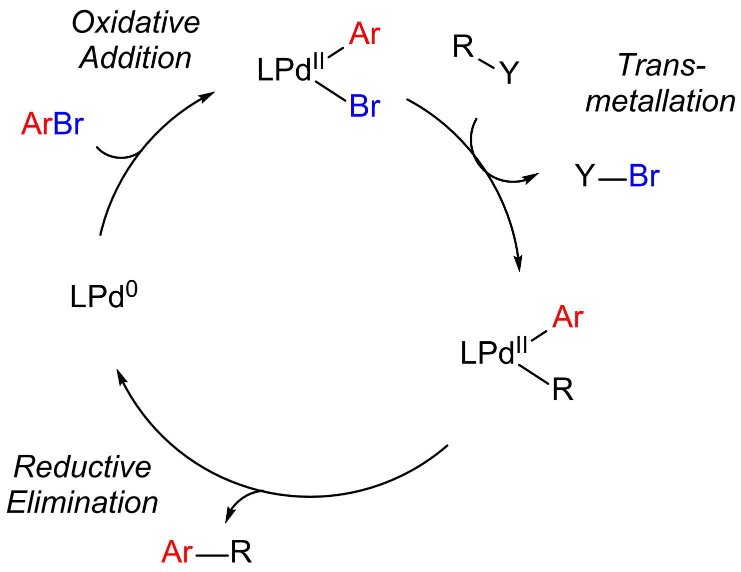
- These numbers can then be used as input for machine learning models.

3. Digital representation of molecules, featurization

Important aspects of featurization

- Performance of ML models are strongly influenced by the relevance of the input features.
- Catalytic activity and selectivity frequently depend on the specific local environment around reactive sites.
- Chemical intuition can play a central role in selecting which descriptors to generate.

Often the **late-limiting step**



Substrate effects (aryl bromides in cross-coupling)

Electronic effects

Electron-withdrawing groups (–NO_2, –CF_3, –CN, etc.): Make the aryl bromide more reactive toward oxidative addition.

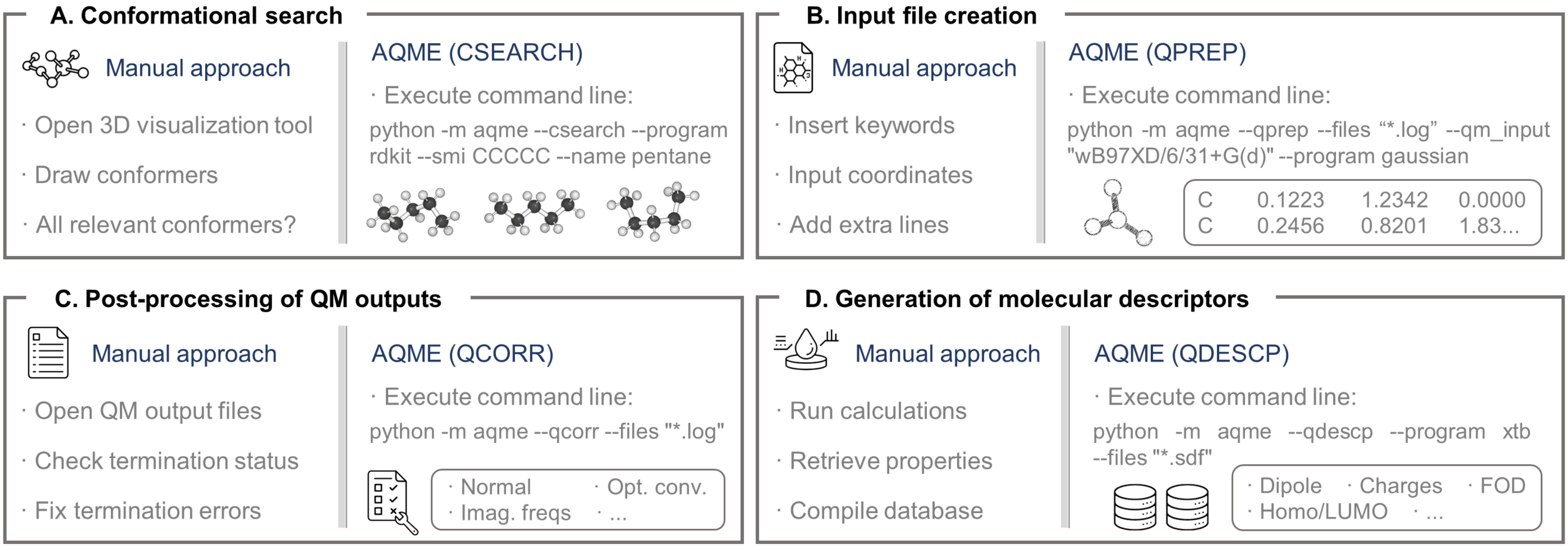
Partial charge

Steric effects

Ortho-substitution (bulky groups close to the bromine): Hinders access of the metal center to the C–Br bond.

Buried volume

AQME: automation of QM protocols



A. Conformational search

Manual approach

· Open 3D visualization tool

· Draw conformers

· All relevant conformers?

AQME (CSEARCH)

· Execute command line:

```
python -m aqme --csearch --program
rdkit --smi CCCCC --name pentane
```

B. Input file creation

Manual approach

· Insert keywords

· Input coordinates

· Add extra lines

AQME (QPREP)

· Execute command line:

```
python -m aqme --qprep --files “*.log” --qm_input
"wB97XD/6/31+G(d)" --program gaussian
```

```
C    0.1223    1.2342    0.0000
C    0.2456    0.8201    1.83...
```

C. Post-processing of QM outputs

Manual approach

· Open QM output files

· Check termination status

· Fix termination errors

AQME (QCORR)

· Execute command line:

```
python -m aqme --qcorr --files "*.log"
```

· Normal · Opt. conv.
· Imag. freqs · ...

D. Generation of molecular descriptors

Manual approach

· Run calculations

· Retrieve properties

· Compile database

AQME (QDESCP)

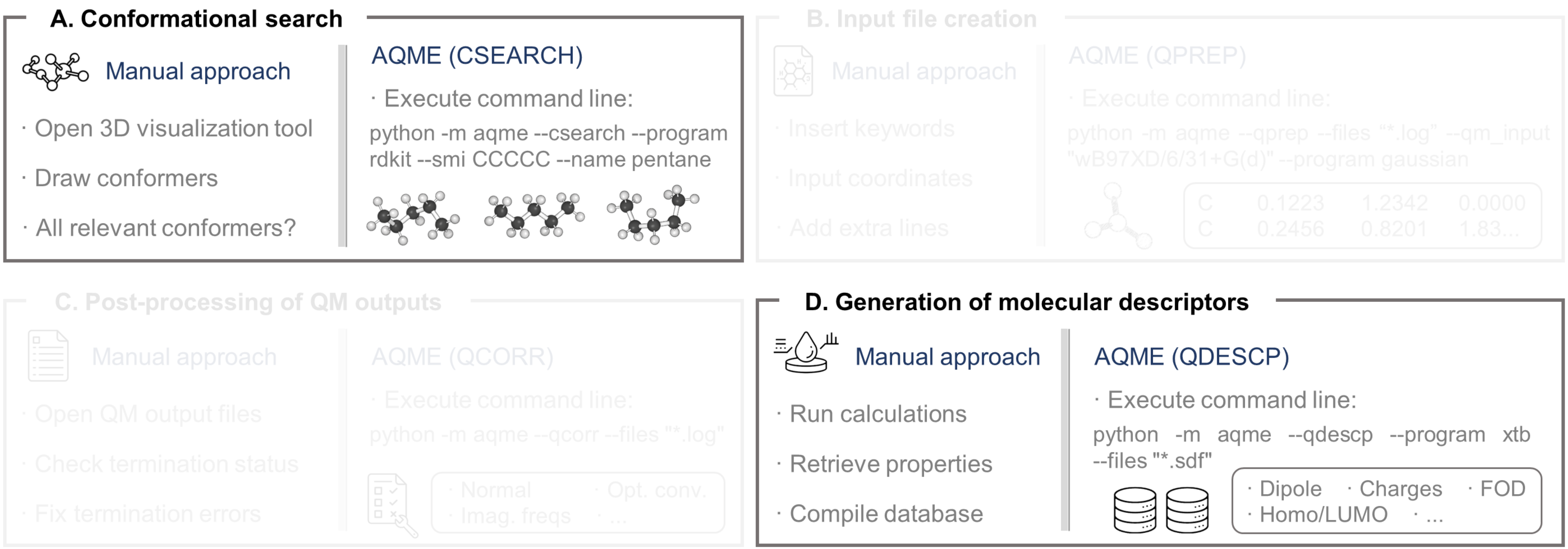
· Execute command line:

```
python -m aqme --qdescp --program xtb
--files "*.sdf"
```

· Dipole · Charges · FOD
· Homo/LUMO · ...

Alegre-Requena, Paton & col., *WIREs. Comput. Mol. Sci.* **2023**, e1663

AQME: automation of QM protocols



A. Conformational search

Manual approach

- Open 3D visualization tool
- Draw conformers
- All relevant conformers?

AQME (CSEARCH)

- Execute command line:

```
python -m aqme --csearch --program
rdkit --smi CCCCC --name pentane
```

B. Input file creation

Manual approach

- Insert keywords
- Input coordinates
- Add extra lines

AQME (QPREP)

- Execute command line:

```
python -m aqme --qprep --files “*.log” --qm_input
"wB97XD/6/31+G(d)" --program gaussian
```

```
C     0.1223     1.2342     0.0000
C     0.2456     0.8201     1.83...
```

C. Post-processing of QM outputs

Manual approach

- Open QM output files
- Check termination status
- Fix termination errors

AQME (QCORR)

- Execute command line:

```
python -m aqme --qcorr --files "*.log"
```

- Normal · Opt. conv.
- Imag. freqs · ...

D. Generation of molecular descriptors

Manual approach

- Run calculations
- Retrieve properties
- Compile database

AQME (QDESCP)

- Execute command line:

```
python -m aqme --qdescp --program xtb
--files "*.sdf"
```

- Dipole · Charges · FOD
- Homo/LUMO · ...

Alegre-Requena, Paton & col., *WIREs. Comput. Mol. Sci.* **2023**, e1663

AQME: automation of QM protocols

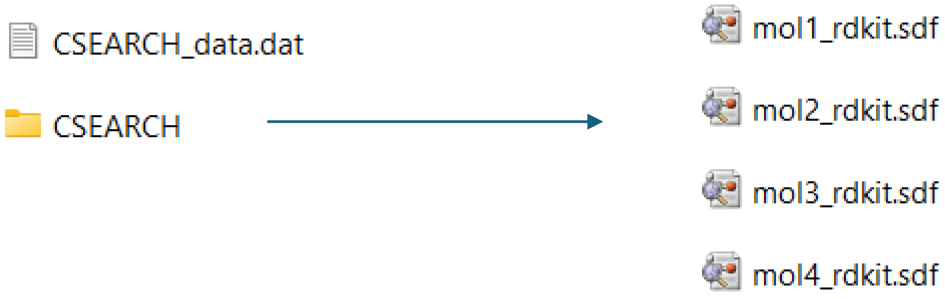
CSEARCH **Generate 3D geometries and search for conformers of molecules**

Input: a SMILE (if you only want one molecule)

CSV file with SMILES and code_name

other files: .sdf, .cdx, .csv, .com, .gjf,
.mol, .mol2, .xyz, .txt, .yaml, .yml, .rtf

Output: a SDF file for each molecule



CSEARCH_data.dat

CSEARCH → mol1_rdkit.sdf, mol2_rdkit.sdf, mol3_rdkit.sdf, mol4_rdkit.sdf

sample (default = 25): maximum number of final conformers generated.

It removes duplicates using energy and structural similarity filters.

If there are > 25, it performs clustering using the molecule's dihedral angles to select the 25 most different ones .

program (default = rdkit). crest if you want more exhaustive computational sampling.

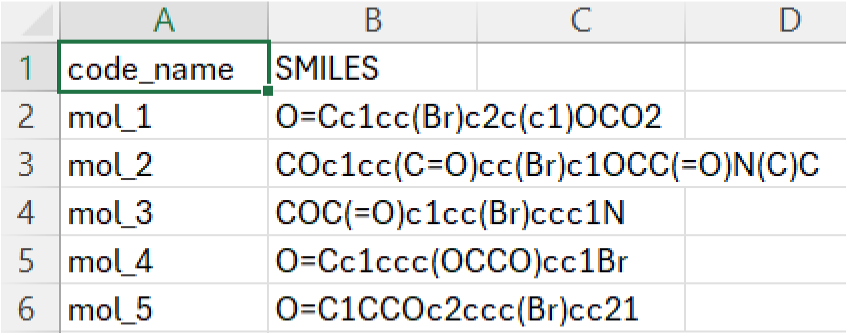
- Go to the **case_study** folder (ubuntu terminal) and open **CSEARCH.ipynb** typing **code .**

https://aqme.readthedocs.io/en/latest/API/aqme.csearch.html

AQME: automation of QM protocols

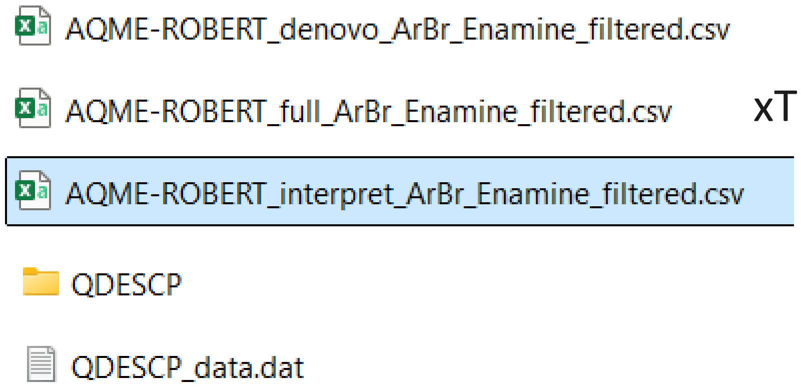
QDESCP **Generate descriptors from semi-empirical QM (xTB), RDKit and MORFEUS**

Input: CSV file with SMILES and code_name:



	A	B	C	D
1	code_name	SMILES		
2	mol_1	O=Cc1cc(Br)c2c(c1)OCO2		
3	mol_2	COc1cc(C=O)cc(Br)c1OCC(=O)N(C)C		
4	mol_3	COC(=O)c1cc(Br)ccc1N		
5	mol_4	O=Cc1ccc(OCCO)cc1Br		
6	mol_5	O=C1CCOc2ccc(Br)cc21		

Output: 3 CSV files containing different numbers of descriptors (**3 levels**):



AQME-ROBERT_denovo_ArBr_Enamine_filtered.csv xTB + MORFEUS

AQME-ROBERT_full_ArBr_Enamine_filtered.csv xTB + MORFEUS + RDKit

AQME-ROBERT_interpret_ArBr_Enamine_filtered.csv xTB + MORFEUS

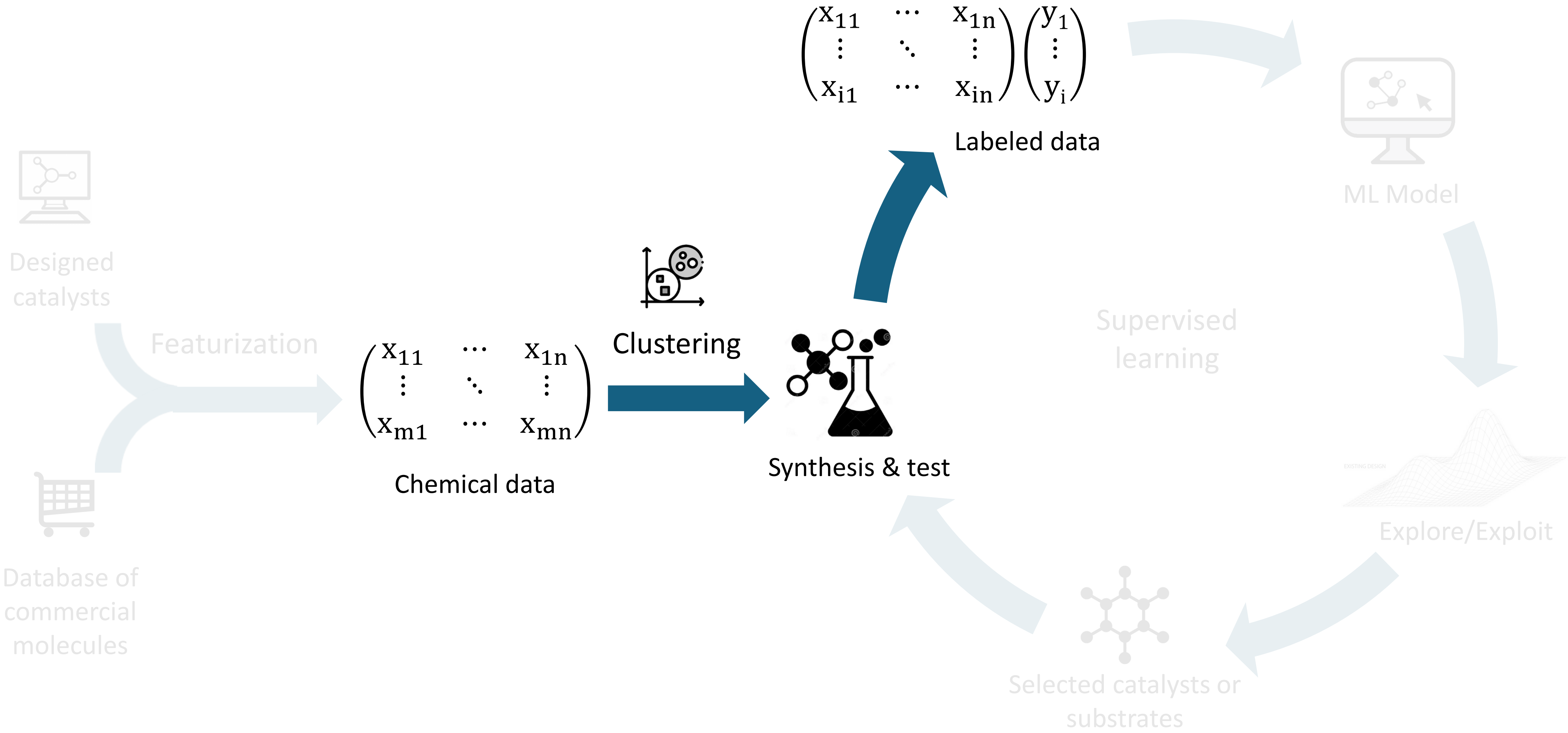
QDESCP

QDESCP_data.dat

- Go to the **descriptors_result** folder and check the CSV files generated for the 25,000 molecules
- Go to the **case_study** folder (ubuntu terminal) and open **QDESCP.ipynb** typing **code .**

https://aqme.readthedocs.io/en/latest/API/aqme.qdescp.html

4. Unsupervised learning: clustering



Designed catalysts

Database of commercial molecules

Featurization

$$\begin{pmatrix} x_{11} & \cdots & x_{1n} \\ \vdots & \ddots & \vdots \\ x_{m1} & \cdots & x_{mn} \end{pmatrix}$$

Chemical data

Clustering

Synthesis & test

$$\begin{pmatrix} x_{11} & \cdots & x_{1n} \\ \vdots & \ddots & \vdots \\ x_{i1} & \cdots & x_{in} \end{pmatrix} \begin{pmatrix} y_1 \\ \vdots \\ y_i \end{pmatrix}$$

Labeled data

ML Model

Supervised learning

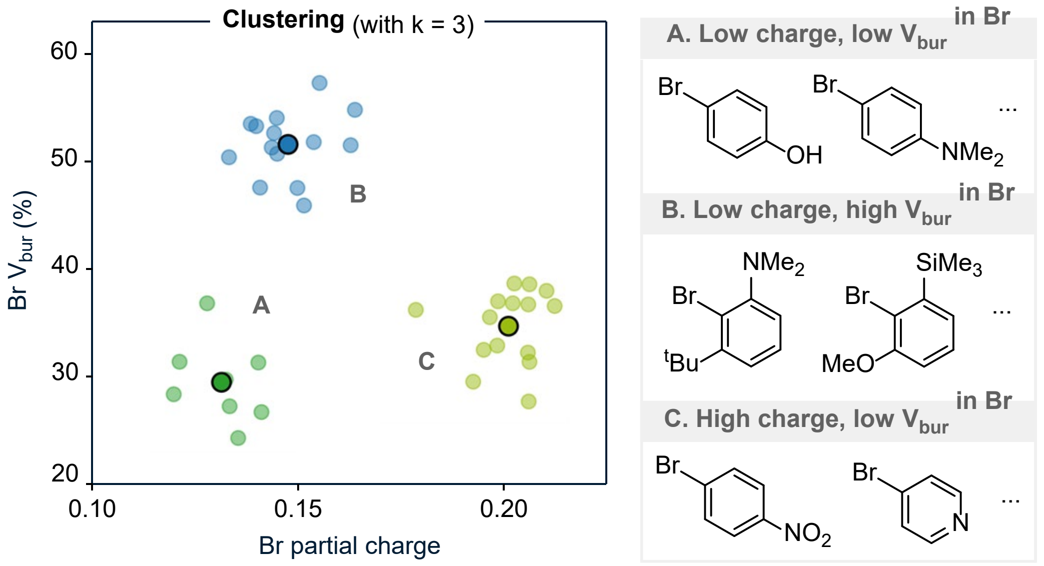
Explore/Exploit

Selected catalysts or substrates

4. Unsupervised learning: clustering

Clustering

- Group similar molecules together and select 1 per group.
- To make the most **efficient selection** of **initial data**.
- It allows to build more general and reliable models.
- k-means, HDBSCAN, UMAP, and t-SNE.



Chemical space

- Each molecule can be thought of as a point in this space, defined by their descriptors.
- Helps to understand **where our data is located** and **which regions covers**.
- If we only train models on a small or narrow part of that space, the model might work well there, but fail when applied to new, unseen molecules from other regions.

4. Unsupervised learning: clustering

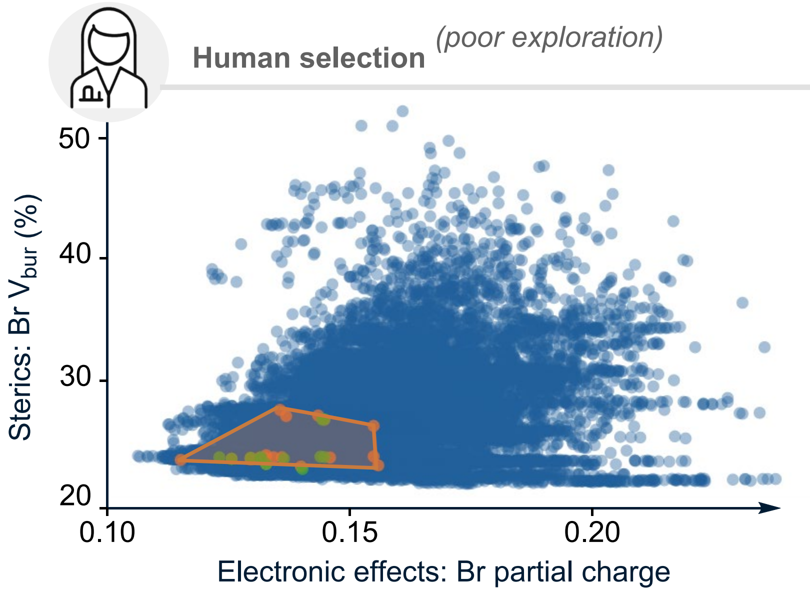
Study case I

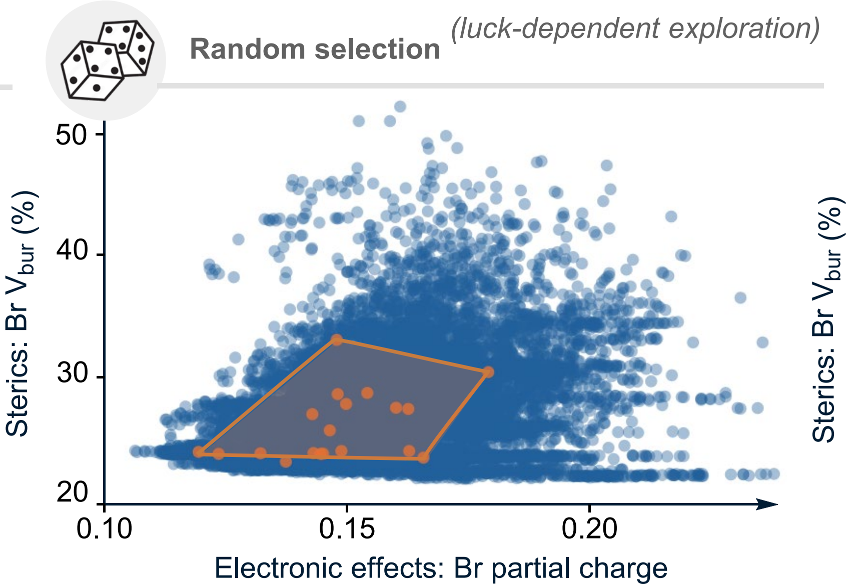
25,000 substrates

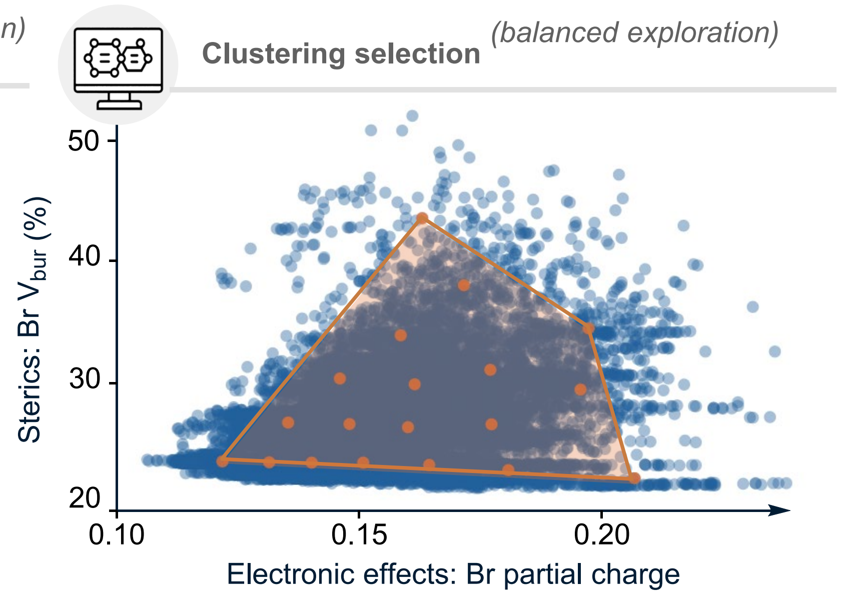
↓

19 clusters

- Experimentally, we can't test those 25,0000 substrates.
- We choose 19, trying to make them as heterogeneous as possible with respect to their descriptors.


Human selection *(poor exploration)*

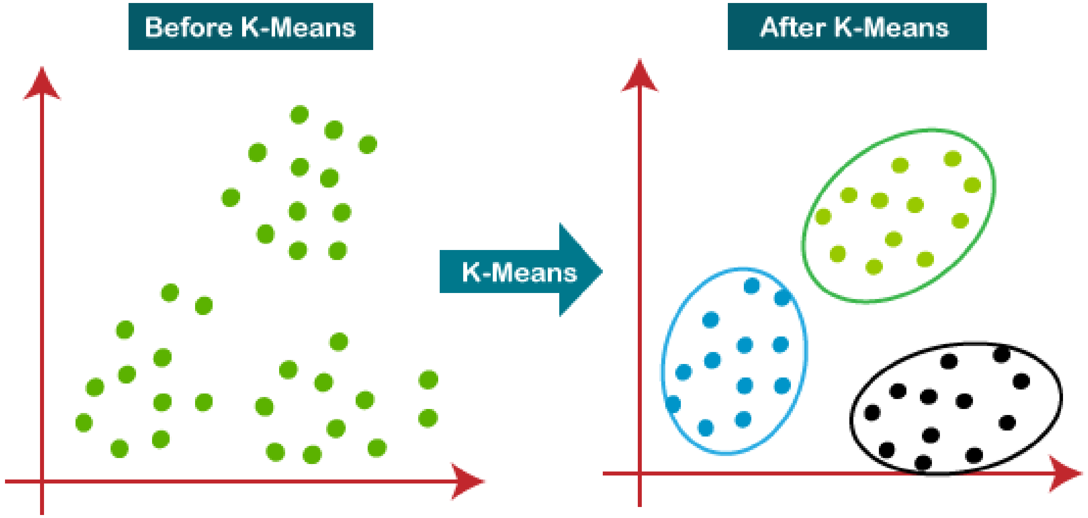

Random selection *(luck-dependent exploration)*


Clustering selection *(balanced exploration)*

4. Unsupervised learning: clustering

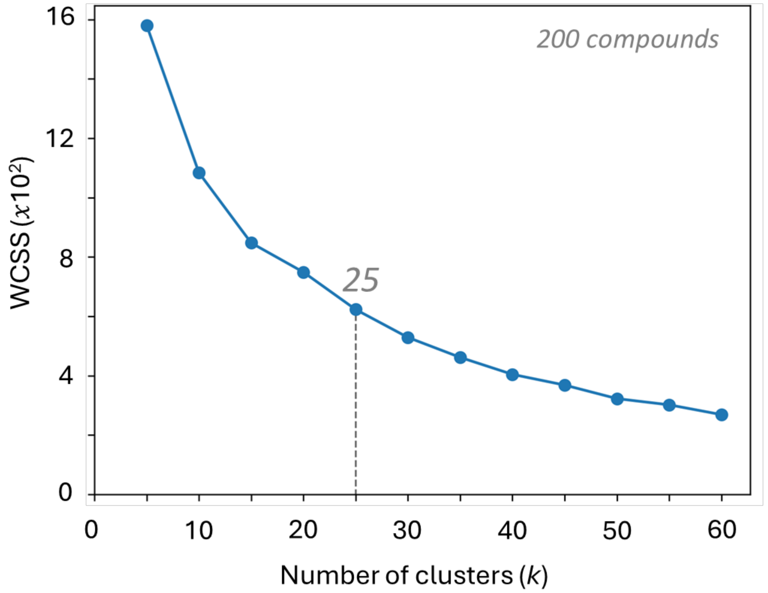
K-means clustering

1. Choose the number of clusters (*k)*.
2. Initialize *k* centroids (randomly).
3. Assign each points to the nearest centroid.
4. Update centroids as the mean of assigned points.
5. Repeat until centroids stabilize (convergence).



Choosing k: the elbow method (as guideline)

Approach to decide the number of clusters: Plot explained variance (or inertia) vs. k and look for the "elbow" point, where slope changes.



Notes

Quality depends on chemical descriptors chosen.

Example in 2D for visualization, but clustering can work in N dimensions.

4. Unsupervised learning: clustering

Principal Component Analysis (PCA)

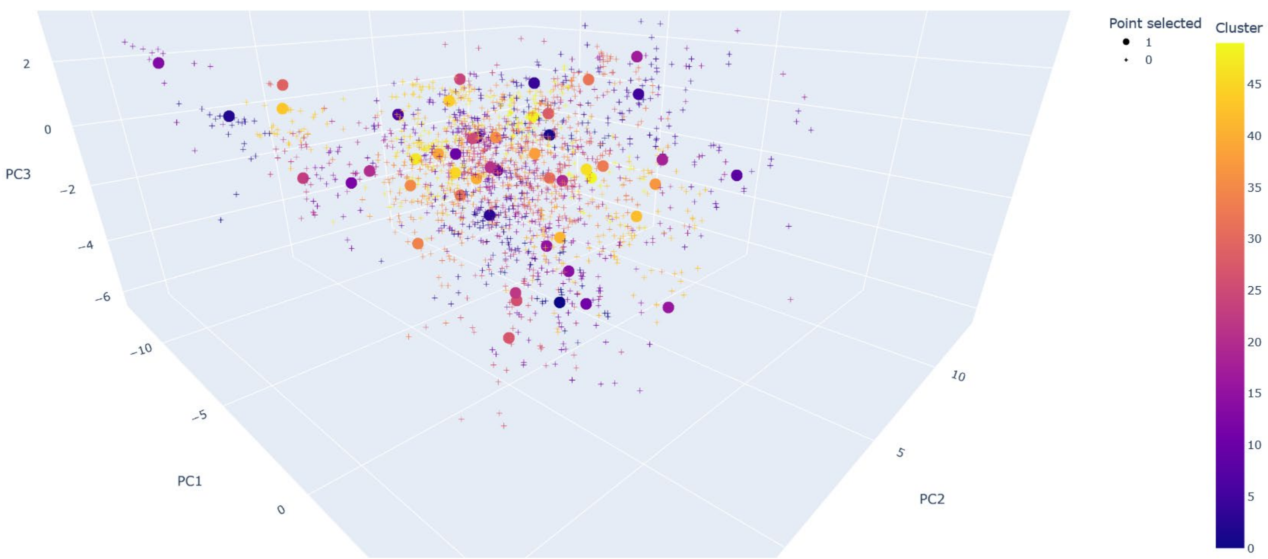
- A linear method that transforms data into a new coordinate system, maximizing variance along principal components.
- Reduction of dimensionality (variables), losing as little information (variance) as possible.
- Each dimension or principal component generated by PCA will be a **linear combination of the original variables**, and they will also be independent or uncorrelated with each other.

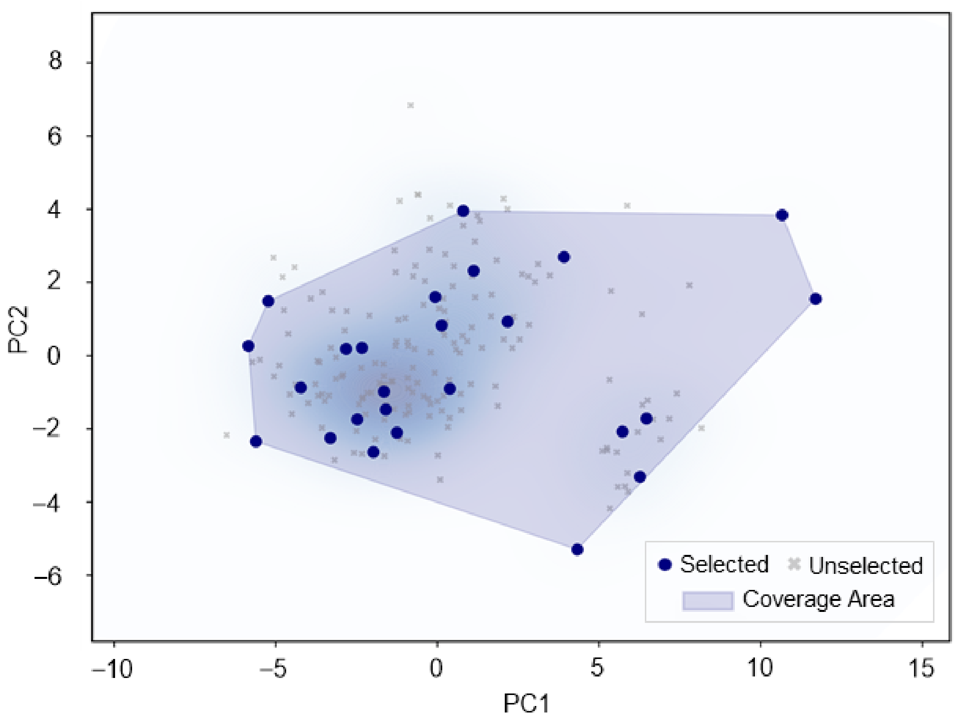
$$\begin{array}{cccc} & \text{descriptor 1} & \dots & \text{descriptor n} \\ \text{ArBr 1} & x_{11} & \cdots & x_{1n} \\ \dots & \vdots & \ddots & \vdots \\ \text{ArBr m} & x_{m1} & \cdots & x_{mn} \end{array} \xrightarrow{\text{3D PCA}} \begin{array}{cccc} & \text{PC1} & \text{PC2} & \text{PC3} \\ \text{ArBr 1} & x_{11} & x_{12} & x_{13} \\ \dots & \vdots & \ddots & \vdots \\ \text{ArBr m} & x_{m1} & x_{m2} & x_{m3} \end{array}$$

4. Unsupervised learning: clustering

Evaluation clustering with PCA

- Reduce chemical space to 2D/3D for visualization.
- Meaningful only if PCA explains ≥ 60-70% of variance.
- Visual inspection can help to assess cluster formation and coverage.





72.7% explained variability: PC1 53.3%, PC2 19.4%

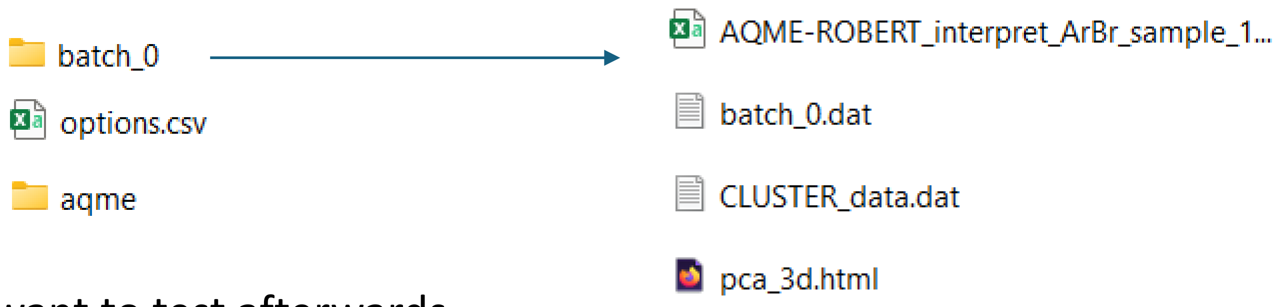
ALMOS: automation of clustering

CLUSTER Group similar molecules together and select 1 per group using *k*-means

Input: CSV file with name and descriptors

CSV file with SMILES and code_name

Output: batch_0 selection and PCA representation



batch_0

options.csv

aqme

AQME-ROBERT_interpret_ArBr_sample_1...

batch_0.dat

CLUSTER_data.dat

pca_3d.html

name: name of your molecules (i.e. code_name)

n_clusters: number of representative molecules you want to test afterwards

ignore: list with the columns that aren’t the n descriptors [‘…’, ‘…’] (list of strings)

aqme: to generate also the descriptors using AQME

remove name

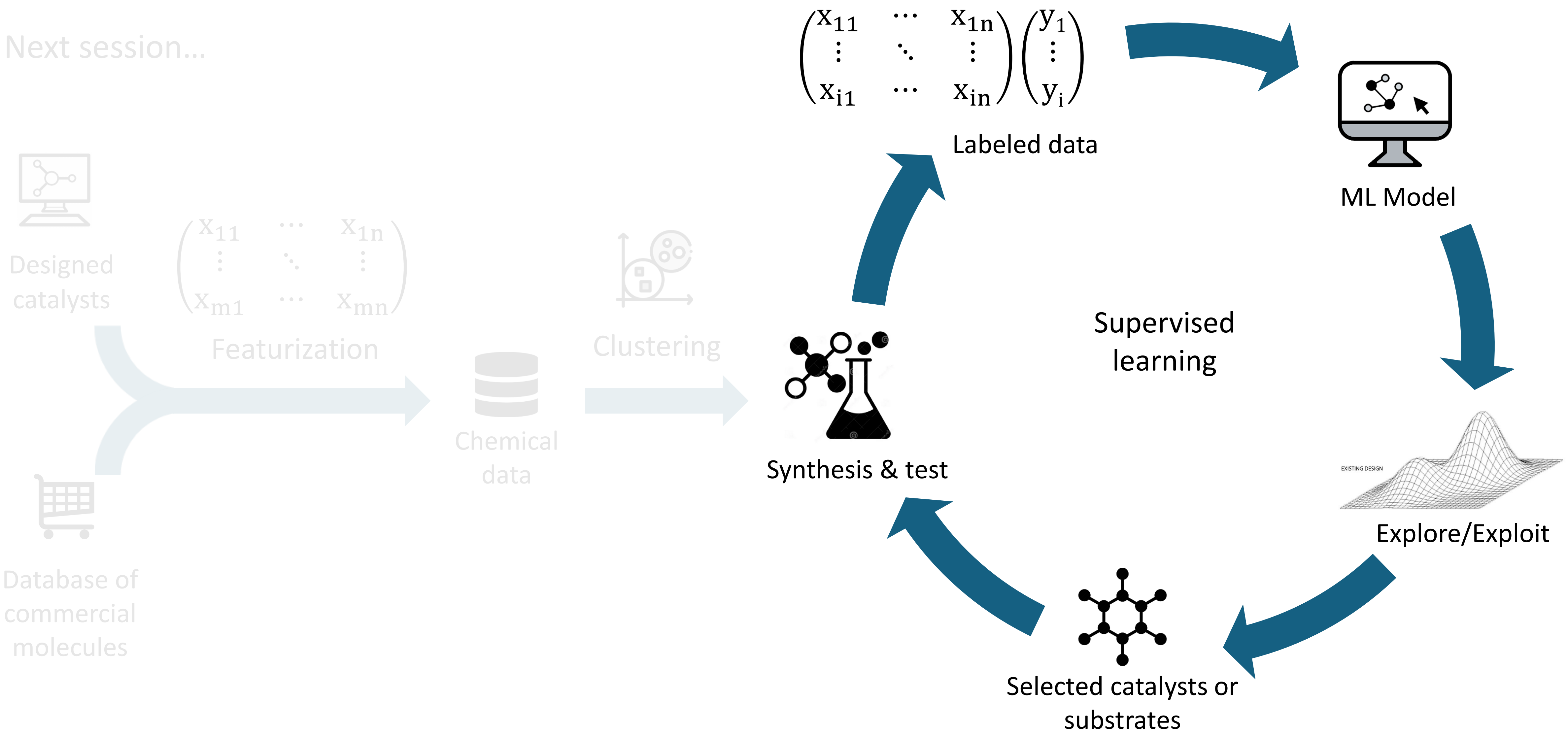
aqme_keywords: for atomic descriptors or another AQME specification

➢ Go to the **case_study** folder (ubuntu terminal) and open **CLUSTER.ipynb** typing **code .**

https://github.com/MiguelMartzFdez/almos https://almos.readthedocs.io/en/latest/

5. Supervised learning: active learning

Next session…


Designed catalysts
$\begin{pmatrix} x_{11} & \cdots & x_{1n} \\ \vdots & \ddots & \vdots \\ x_{m1} & \cdots & x_{mn} \end{pmatrix}$
Featurization
Chemical data
Clustering
Database of commercial molecules
$\begin{pmatrix} x_{11} & \cdots & x_{1n} \\ \vdots & \ddots & \vdots \\ x_{i1} & \cdots & x_{in} \end{pmatrix} \begin{pmatrix} y_1 \\ \vdots \\ y_i \end{pmatrix}$
Labeled data
ML Model
Supervised learning
EXISTING DESIGN
Explore/Exploit
Selected catalysts or substrates
Synthesis & test

Exercises

Go to the **exercises** folder (ubuntu terminal) and open **exercises.ipynb** typing **code .** :

Step 1: Use AQME QDESCP (which includes CSEARCH) to generate descriptors from the file alkynes.csv.
If the alkyne is relevant to the reaction under study, would you add atomic descriptors? Which ones?

Step 2: Performs clustering using ALMOS CLUSTER and the CSV files from alkyne_32 folder

- *Explore the 3 levels of descriptors (full, interpret, denovo)*
- *Explore different number of clusters, analysing the PCA representation*
- *Explore the result of apply the elbow method*

Instituto de Síntesis Química y Catálisis Homogénea

From quantum-chemical descriptors to clustering, an automated pipeline

September 17, 2025

Dr. Susana García-Abellán