Projet SCHISM

# AMÉLIORATION DE L'ANALYSE SAR VIA LA RÉDUCTION DES CARACTÉRISTIQUES PHARMACOPHORIQUES ET LA TRANSFORMATION DES CARACTÉRISTIQUES

Hajar Rehioui

GREYC, Normandie Univ., UNICAEN, CNRS – UMR 6072, 14000 Caen, France
hajar.rehioui-karine@unicaen.fr

Le projet SCHISM est financé par l'Union européenne dans le cadre du programme opérationnel FEDER/ FSE 2014-2020

13 Décembre 2021

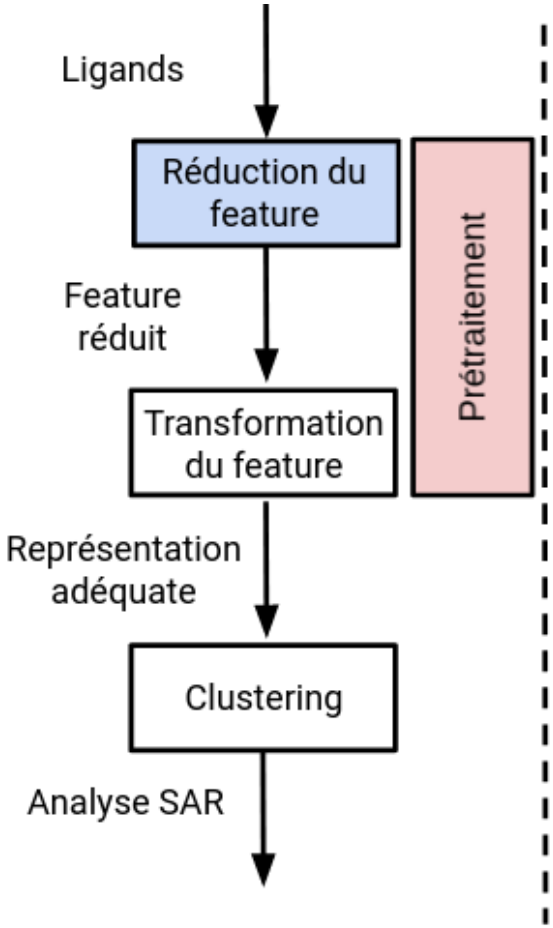
## Objectif général

- Analyse de la relation structure-activité (SAR) :
  - recherche d'une représentation appropriée.
  - regrouper les ligands par famille, découvrir les 'activity cliffs'.

## données

- 1485 ligands [1] testé sur la tyrosine kinase BCR-ABL, souvent retrouvée chez les patients atteints de leucémie myéloïde chronique.
- Chaque ligand est défini par 112048 pharmacophores (ordre variant de 3 à 7)
- Deux classes : 711 ligands de la classe inactive et 774 ligands de la classe active.
- Activité du ligand :
  - $K_i \leq 100nM \implies$ active
  - $K_i \geq 1000nM \implies$ inactive
  - sinon pas pris en compte

1. GAULTON, Anna, HERSEY, Anne, NOWOTKA, Michal, et al. The ChEMBL database in 2017. Nucleic acids research, 2017, vol. 45, no D1, p. D945-D954.




## Objectif : Nettoyer les données

- Éliminer la redondance.
- Conservez les informations pertinentes.

## Classes d'équivalence [2] (EC)

- EC : un groupe de pharmacophores qui apparaissent dans exactement les mêmes ligands.
- Les colonnes (pharmacophores) appartenant à la même EC sont réduites à une seule colonne représentative.

| | EC1 | | | EC2 | |
|---|---|---|---|---|---|
| Id | \|A\|R\|D\|H\| | \|A\|R\|R\|D\|H\| | \|A\|R\|R\|R\|D\|H\| | \|A\|A\|R\|D\| | \|A\|A\|R\|R\|D\| |
| ligand_1 | 1 | 1 | 1 | 0 | 0 |
| ligand_2 | 0 | 0 | 0 | 0 | 0 |
| ligand_3 | 1 | 1 | 1 | 1 | 1 |

2. FOURNIER-VIGER, Philippe, GUENICHE, Ted, ZIDA, Souleymane, et al. ERMiner : sequential rule mining using equivalence classes. In : International Symposium on Intelligent Data Analysis. Springer, Cham, 2014. p. 108-119.

- Réduction de 86.50%

| ligands | l'ancien features | le nouveau features | classes |
|---|---|---|---|
| 1485 | 112048 | **15129** | 2 |

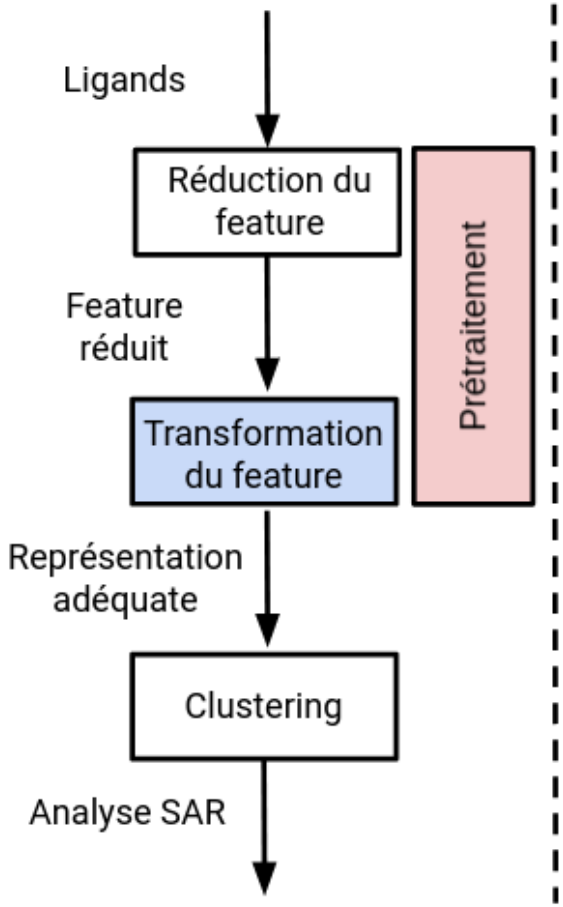
- Exemple des classes d'équivalence

| EC | number_of_ph | concept_score |
|---|---|---|
| \|A\|A\|R\|D\|H\|N\|-6279( | 5316 | 12 |
| \|A\|A\|A\|R\|R\|H\|H\|-853 | 3889 | 13 |
| \|A\|A\|A\|R\|D\|D\|H\|-874 | 3552 | 10 |
| \|A\|A\|A\|D\|H\|H\|-5283( | 3515 | 10 |
| \|A\|A\|A\|R\|R\|H\|-48532 | 1527 | 12 |
| \|A\|R\|R\|H\|-13429 | 1443 | 11 |
| \|A\|A\|R\|D\|-8476 | 1153 | 20 |
| \|A\|R\|R\|H\|P\|P\|-71742 | 1094 | 12 |
| \|A\|R\|R\|R\|D\|H\|-6714( | 1047 | 12 |
| \|A\|A\|A\|A\|R\|P\|-45904 | 1000 | 10 |

> outputs > {} Dictionary_classe_equivalence_ph4_bis.json >

```
        "|A|A|A|-131"
    ],
    "|A|A|A|-132": [
        "|A|A|A|-132"
    ],
    "|A|A|A|-133": [
        "|A|A|A|-133"
    ],
    "|A|A|A|-134": [
        "|A|A|A|-134",
        "|A|A|D|-567",
        "|A|A|D|-751",
        "|A|A|N|-1178",
        "|A|A|N|-1180",
```

- 11577 EC contiennent un seule pharmacophore, 3552 EC contiennent plus.

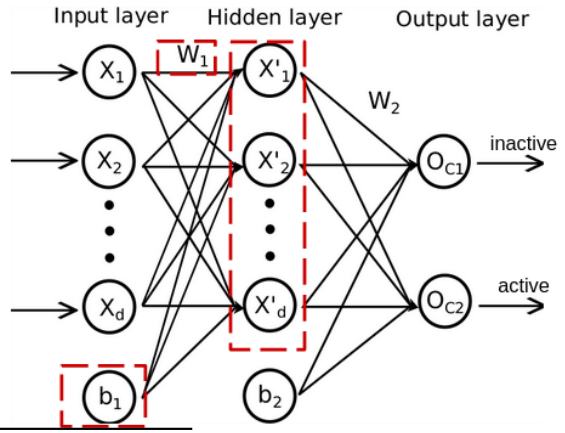



## Objectif

- Séparer les ligands actifs des inactifs
- Trouver une transformation adéquate

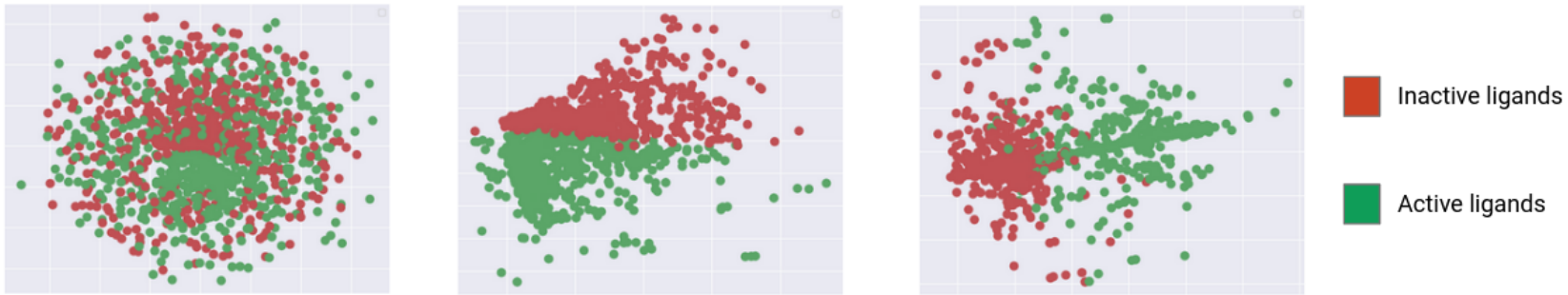
## Transformation supervisée

- Transformer les caractéristiques des ligands en fonction de leur activité biologique.
- Utiliser un réseau de neurones [3] qui minimise la fonction d'optimisation "categorical_crossentropy".




$$x'_{ij} = \sum_{j'=1}^{d} (x_{ij'} \times w_{j'i}) + b$$

3. BEBIS, George et GEORGIOPOULOS, Michael. Feed-forward neural networks. IEEE Potentials, 1994, vol. 13, no 4, p. 27-31.




Originale | Transformée | Transformée avec régularisation

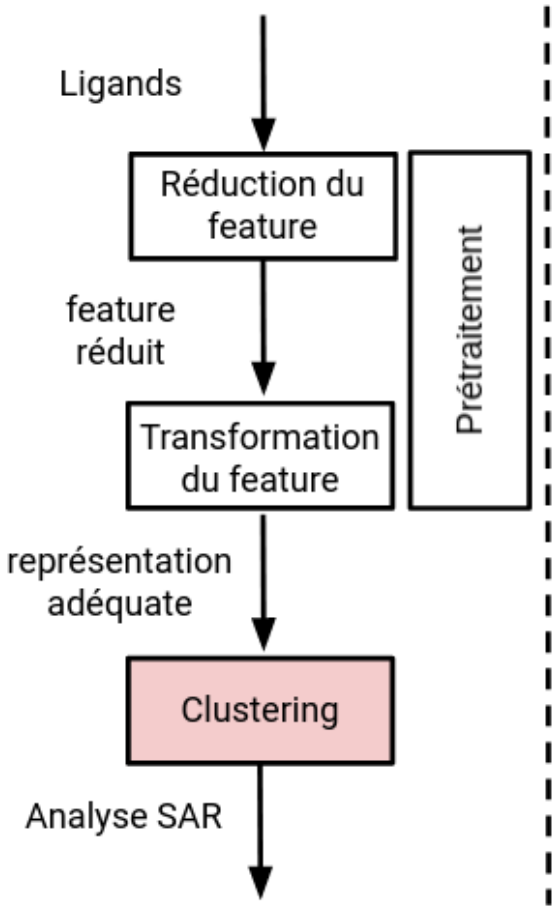
## Visualisation de la projection 2D

Les 15129 pharmacophores sont projetées par la méthode 'Multidimensional scaling'[4](MDS) method.

## Régularisation du réseau de neurones

- Le Dropout : technique qui empêche le surapprentissage en supprimant temporairement les neurones (dans les couches d'entrée ou cachées)
- Régularisation L2 : technique qui empêche le surapprentissage en ajoutant $Loss_{L2} = \sum_{i=1}^{n} w_i^2$ à la fonction loss du modèle d'apprentissage.

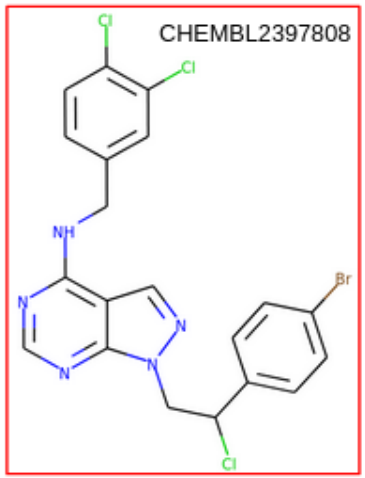
4. COX, Michael AA et COX, Trevor F. Multidimensional scaling. In : Handbook of data visualization. Springer, Berlin, Heidelberg, 2008. p. 315-347.

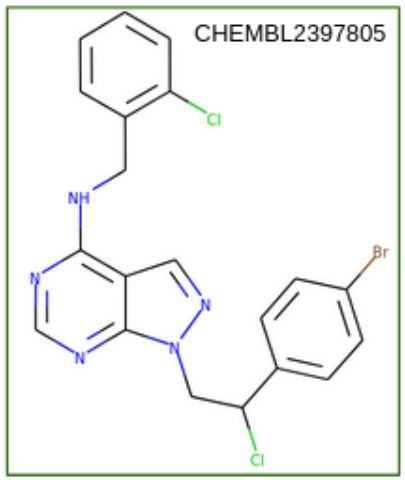



**Objectif : grouper les structures similaires ensembles**

trouver des familles significatifs de ligands

- cluster homogène : familles d'activités biologiques.
- cluster non homogène : présence probable 'd'activity cliffs' [5].



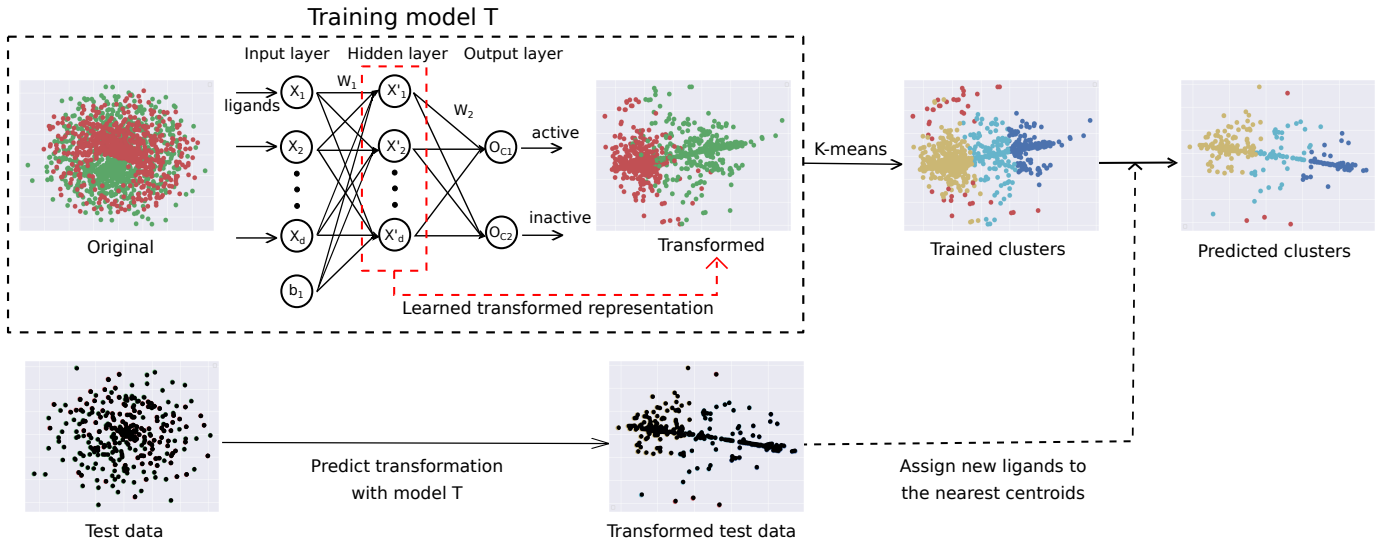

5. STUMPFE, Dagmar, HU, Huabin, et BAJORATH, Jürgen. Evolving concept of activity cliffs. ACS omega, 2019, vol. 4, no 11, p. 14360-14368.

## Objectif

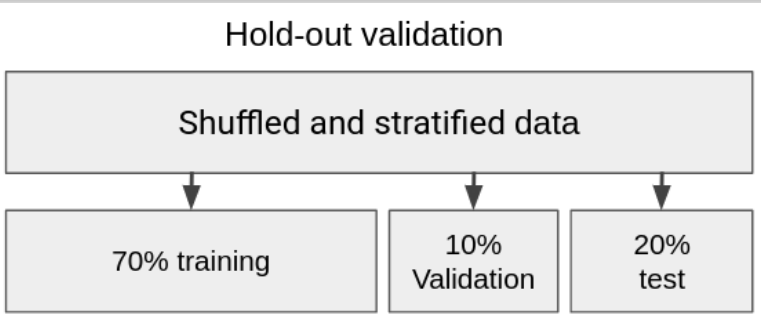
- Prédire le cluster (famille) d'un nouveau ligand non étiqueté.

## Principe

- Appliquer l'une des méthodes de clustering (K-means) sur l'ensemble d'apprentissage après sa transformation
- Attribuez à chaque cluster construit à la première étape les données test qui lui correspondent (les plus similaires)



# Evaluation




## Evaluation by normalized mutual information [6] (NMI)

- quality measure which compares the resulting clusters or classes with the ground truth.
- The results vary between 0 (no mutual information) and 1 (perfect correlation)

## Evaluation by Silhouette [7]

- measure calculated using the mean intra-cluster distance and the mean distance to the nearest cluster.
- The best value is 1 and the worst is -1. Values close to 0 indicate that the clusters overlap. Negative values usually indicate that a sample was assigned to the wrong cluster.

---

6. ESTEVEZ, Pablo A., TESMER, Michel, PEREZ, Claudio A., et al. Normalized mutual information feature selection. IEEE Transactions on neural networks, 2009, vol. 20, no 2, p. 189-201.

7. ROUSSEEUW, Peter J. Silhouettes : a graphical aid to the interpretation and validation of cluster analysis. Journal of computational and applied mathematics, 1987, vol. 20, p. 53-65.

# Choice of the number of clusters K

| | data | NMI | Silhouette | misclassified ligands |
|---|---|---|---|---|
| k=2 | Original | 0.287 | 0.056 | 89 |
| | Transformed | 0.425 | 0.534 | 53 |
| k=3 | Original | 0.277 | 0.068 | 86 |
| | Transformed | 0.399 | 0.552 | 53 |
| k=4 | Original | 0.288 | 0.081 | 80 |
| | Transformed | 0.438 | 0.443 | 33 |
| k=5 | Original | 0.245 | 0.092 | 79 |
| | Transformed | 0.406 | 0.441 | 35 |
| k=6 | Original | 0.246 | 0.099 | 75 |
| | Transformed | 0.415 | 0.440 | 34 |
| k=7 | Original | 0.240 | 0.090 | 78 |
| | Transformed | 0.374 | 0.404 | 38 |
| k=8 | Original | 0.230 | -0.109 | 72 |
| | Transformed | 0.318 | 0.380 | 47 |
| k=9 | Original | 0.222 | 0.109 | 74 |
| | Transformed | 0.339 | 0.386 | 40 |

- Résultats en cours d'analyse.

## Formes multiple d'interactivité

- Interactivité par navigation et exploration du réseau (Network Navigation) : trés répondu dans le domaine chemo-informatique [8] (ex :Cytoscape).
- Interactivité dans les méthodes de clustering [9]
  - Interaction par changement de paramétrages
  - Interaction avec les résultats du clustering en fournissant des requêtes concernant les erreurs et des conseils pour une solution améliorée
  - Interaction en interrogeant explicitement un expert de fournir plus d'éclaircissement concernant les données initiales

8. WOLLENHAUPT, Sabrina et BAUMANN, Knut. inSARa : intuitive and interactive SAR interpretation by reduced graphs and hierarchical MCS-based network navigation. Journal of chemical information and modeling, 2014, vol. 54, no 6, p. 1578-1595.

9. BAE, Juhee, HELLDIN, Tove, RIVEIRO, Maria, et al. Interactive clustering : a comprehensive review. ACM Computing Surveys (CSUR), 2020, vol. 53, no 1, p. 1-39.

## Interface interactive [10]

- Base de données : 3000 items de GWAS Catalog
- Interaction avec des biologistes
- Métriques d'évaluation : Silhouette Coefficient, Davies-Bouldin index
- Méthodes de clustering : K-Means, DBScan, Agglomerative Clustering, and Spectral Clustering
- Détails techniques : Python, JavaScript, D3.js


10. DAS, Subhajit, SAKET, Bahador, KWON, Bum Chul, et al. Geono-cluster : Interactive visual cluster analysis for biologists. IEEE Transactions on Visualization and Computer Graphics, 2020.

## A explorer

- Smile ou l'image de chaque ligand (affichage en cliquant sur un point)
- Étiqueter une partie des ligands (pseudo-labeled data) puis procéder à la transformation d'une manière semi-supervisée
- Permettre à l'expert de guider la transformation
- Afficher l'analyse du feature importance
- Permettre de pondérer les features en se basant sur les résultats du feature importance

## Définition

- La découverte polypharmacologique [11] fait référence à la conception d'une molécule médicamenteuse unique qui interagit simultanément avec plusieurs cibles au sein d'un réseau moléculaire lié à une maladie pour obtenir les effets thérapeutiques souhaités.
- Plus simple : des ligands qui interagissent avec plus d'une seule cible

## A explorer

effectuer des transformations en prenant en considération les différentes cibles

11. YANG, Xin, WANG, Yifei, BYRNE, Ryan, et al. Concepts of artificial intelligence for computer-assisted drug discovery. Chemical reviews, 2019, vol. 119, no 18, p. 10520-10594.

Projet SCHISM

# AMÉLIORATION DE L'ANALYSE SAR VIA LA RÉDUCTION DES CARACTÉRISTIQUES PHARMACOPHORIQUES ET LA TRANSFORMATION DES CARACTÉRISTIQUES

Hajar Rehioui

GREYC, Normandie Univ., UNICAEN, CNRS – UMR 6072, 14000 Caen, France
hajar.rehioui-karine@unicaen.fr

Our work is part of the European project "SCHISM", funded by the European Union within the framework of the Operational Programme ERDF/ESF 2014-2020

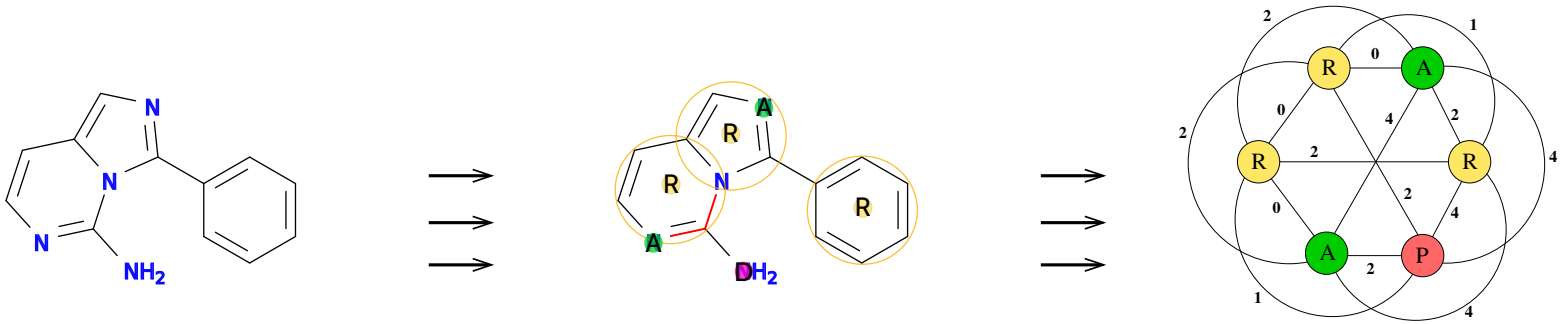
13 Décembre 2021

# appendix

# Structure pharmacophorique

## Charactéristiques pharmacophoriques [12]

Hydrogen Bond Acceptor, Hydrogen Bond Donor, aromatic Ring, Hydrophobic area, Positively ionizable group, Negatively ionizable group




Formule squelettique

Occurrences des caractéristiques

Pharmacophore graph

## Pharmacophore graph of a molecule

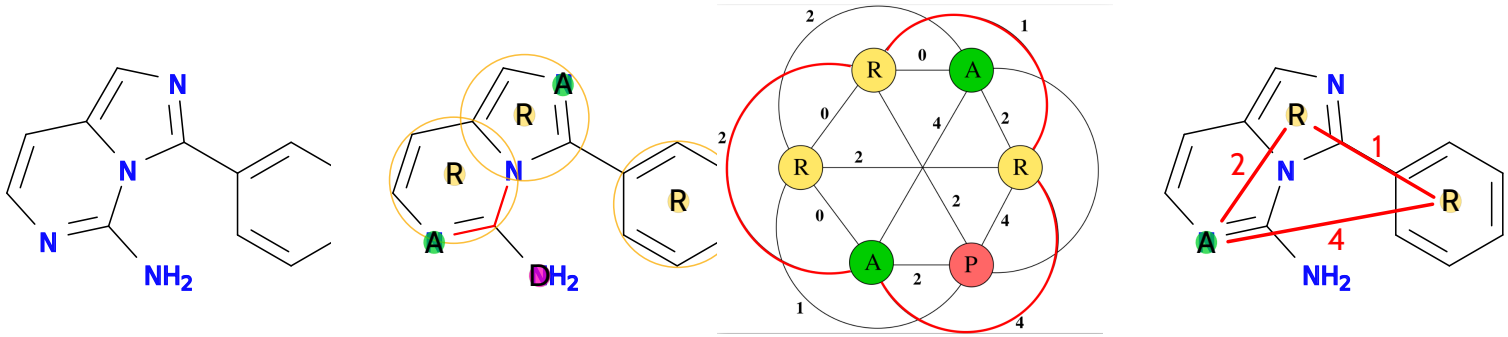
- A vertex : an occurrence of a feature
- An edge : the distance between two features

12. OpenBabel, N. M. O'Boyle, M. Banck, C. A. James, C. Morley, T. Vandermeersch, G. R. Hutchison, J. Cheminformatics 2011, 3, 33.

# Adopted Pharmacophoric vision

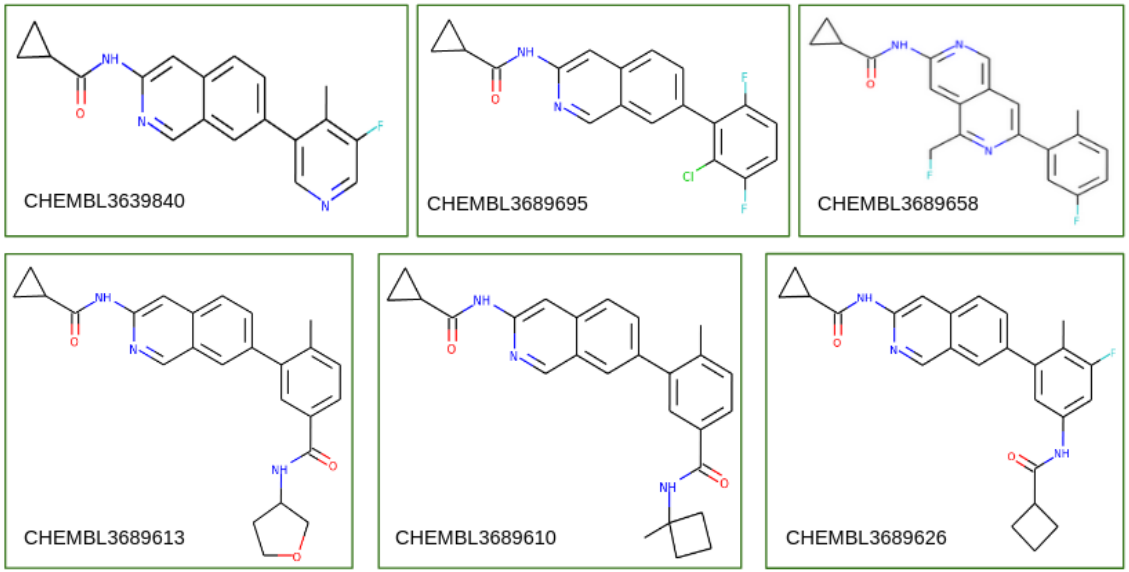
## pharmacophore

- fragment of the overall pharmacophore graph of the ligand responsible of its biological activity (active or inactive).
- sufficiently present (e.g., appearing in at least 10 ligands)



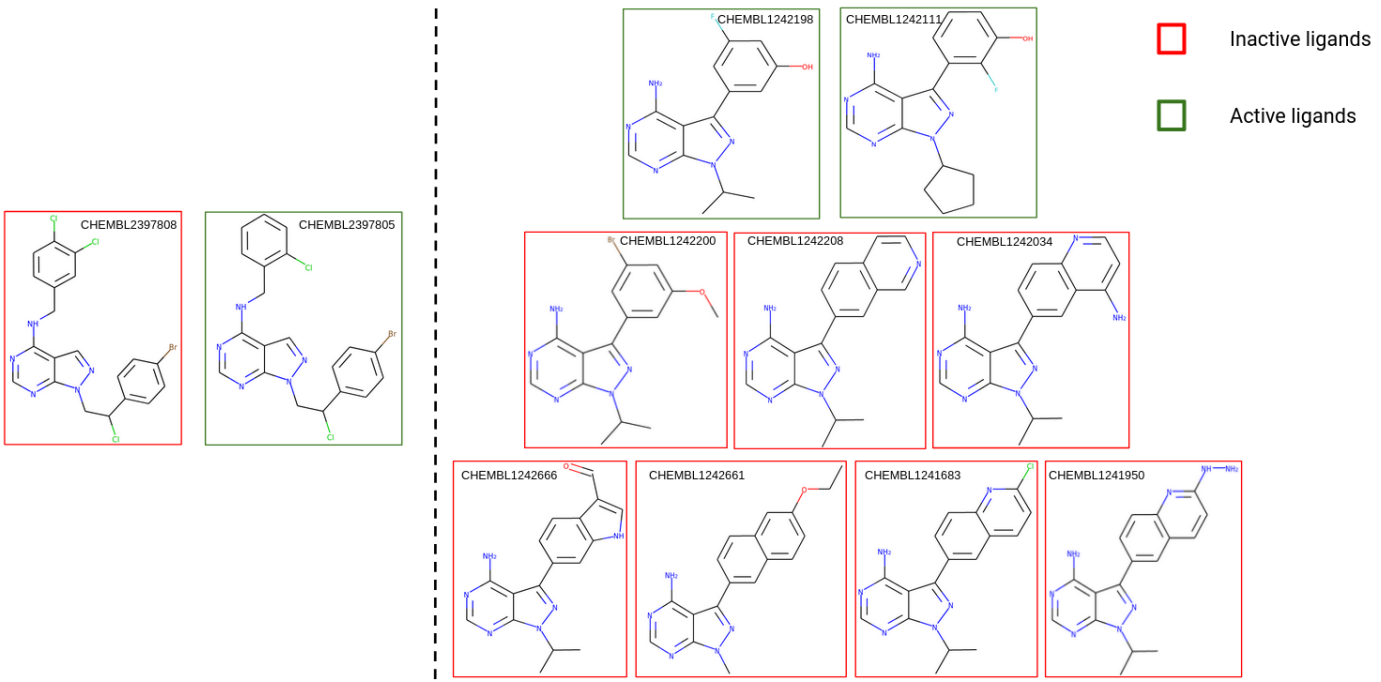
# SAR analysis

| test data | Original | | | | | | | | | Transformed | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Inactive (142) | 0 | 111 | 1 | 1 | 1 | 1 | 5 | 22 | 0 | 21 | 6 | 0 | 64 | 0 | 30 | 0 | 13 | 8 |
| Active (155) | 60 | 52 | 4 | 4 | 1 | 11 | 1 | 17 | 5 | 14 | 41 | 48 | 7 | 10 | 8 | 6 | 18 | 3 |



# SAR analysis

| test data | Original | | | | | | | | | Transformed | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Inactive (142) | 0 | 111 | 1 | 1 | 1 | 1 | 5 | 22 | 0 | 21 | 6 | 0 | 64 | 0 | 30 | 0 | 13 | 8 |
| Active (155) | 60 | 52 | 4 | 4 | 1 | 11 | 1 | 17 | 5 | 14 | 41 | 48 | 7 | 10 | 8 | 6 | 18 | 3 |




A cluster with two activity cliffs