# Méthodes de classification supervisées dans l'évaluation de la relation structurale molécule-enzyme

## Contexte

Le présent article utilise des méthodes supervisées telles que la PCR, la PLS et la forêt aléatoire pour prédire l'effet (mesuré par la constante d'inhibition Ki) d'une molécule donnée sur l'enzyme DRD2.

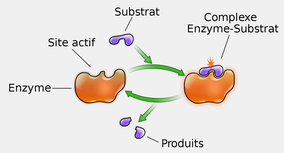



Figure 1. Complexe enzyme-substrat et molécule (produit)

L'objectif général de ce projet est donc de comparer leurs performances dans la prédiction de la variable à expliquer Ki. Nous travaillerons sur une base de données contenant plus de 4000 molécules et descripteurs.

## Modélisation

1. **Modèles linéaires**

   Ils sont au nombre de deux :

   - **Principal Components Regression** : Il s'agit d'une régression sur les facteurs de l'analyse en composantes principales des descripteurs. Les facteurs retenus seront les composantes qui maximisent la variabilités des observations projetées.
   - **Partial Least Squares Regression** : La régression PLS propose de construire également de nouvelles composantes comme des combinaisons linéaires des descripteurs. Comme PCR, les composantes sont calculées les unes après les autres et orthogonales entre elles. La principale différence avec PCR est que nous recherchons les composantes qui maximisent la colinéarité avec la variable à expliquer.

2. **Modèle non linéaire**

   **La forêt aléatoire** est un algorithme de machine learning qui va classer les variables explicatives en fonction de leurs liens avec la variable à expliquer. L'une des caractéristiques les plus importantes de l'algorithme de forêt aléatoire est qu'il peut gérer un ensemble de données contenant des variables continues comme dans le cas de la régression et des variables catégorielles comme dans le cas de la classification




## Résultats et Comparaison

A cette étape, la base de données est divisée en échantillons d'apprentissage et test qui représentent respectivement 80% et 20% du jeu de données initial. Nous avons donc pu obtenir la variance expliquée et l'erreur quadratique moyenne des 3 modèles :

Table 1. Comparaison

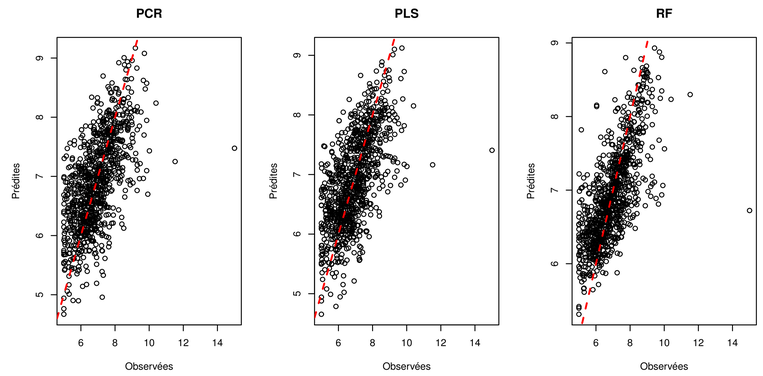
| | *PCR* | | *PLS* | | *RF* | |
|---|---|---|---|---|---|---|
| | $R^2$ | MSEP | $R^2$ | MSEP | $R^2$ | MSEP |
| Echantillon d'apprentissage | 0.519 | 0.529 | 0.55 | 0.494 | 0.497 | 0.553 |
| Echantillon Test | 0.386 | 0.756 | 0.384 | 0.758 | 0.504 | 0.63 |

**PCR vs PLS**

Nous remarquons dans un premier temps que le modèle PLS avec son coefficient de détermination **0,55** ajuste mieux que PCR avec **0,519**. Par contre, leurs pouvoirs prédictifs sont réduits sur l'échantillon de test et sont respectivement **0,384** et **0,386**. Le modèle PLS sera néanmoins choisi comme le meilleur des deux puisqu'il converge plus vite. En effet, **22** composantes ont été nécessaire à la construction du modèle PLS contre **370** pour PCR.

**PLS vs RF**

De l'autre côté, le modèle PLS a également un meilleur ajustement que la forêt aléatoire sur l'échantillon d'apprentissage. Néanmoins, le pouvoir prédictif sur l'échantillon de test est beaucoup plus intéressant pour le modèle non linéaire du fait de sa plus faible erreur quadratique **0,63** et variance expliquée de la variable cible de **0,504** comparée à celle du modèle PLS de **0,384**.

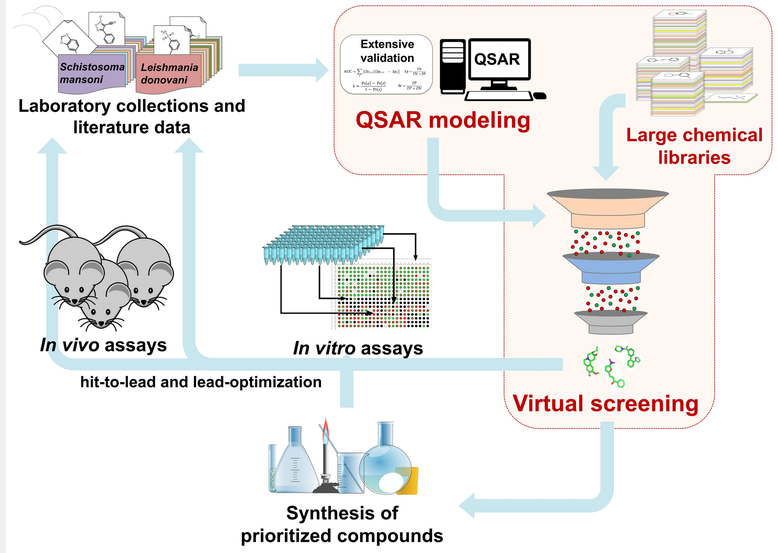



Il en ressort que la **forêt aléatoire** (RF) a le meilleur pouvoir prédictif.

## Conclusion et Perspectives

L'objectif de cette étude était d'appliquer des méthodes d'apprentissage supervisées afin de comparer leurs performances à prédire la constante d'inhibition Ki de l'enzyme DRD2. Nous avons donc eu recours à des modèles linéaires (PLSR et PCR) et non linéaire (Forêt aléatoire). La forêt aléatoire, plus robuste que les deux autres modèles s'est montrée plus efficace que les méthodes linéaires en terme de prédiction malgré leur faible coût. Cependant, son pouvoir prédictif dépassant à peine 50%, ce qui est faible pour bien prédire la constante d'inhibition. Ainsi, il reste à améliorer.

De ce fait, nous pouvons penser à d'autres méthodes plus efficaces pour améliorer la performance prédictive, comme les méthodes de deep learning. Il semble également qu'une sélection d'autres descripteurs plus pertinents soit nécessaire pour mieux ajuster les modèles.




**Étudiants :** AYEREBI Kattié - MENE ABANG Ramiro Obama - THIAM Papa Charles

Projet statistique 2ème année 2021-2022

**Tuteur :** NICOLIS Ioannis | **Coach** : JACKIE Aubree