

M2 AMIS | Modélisation moléculaire | UVSQ

Machine Learning

In chemoinformatique

And drug discovery

Auteurs :

1. Yu-Chen Lo
2. Stefano E Rensi
3. Wen Torng
4. Russ B. Altman

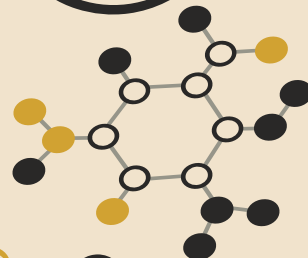
Publié :

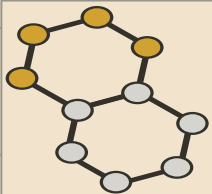
Revue :
"Drug Discovery Today"
Année :
2018

Présenté :

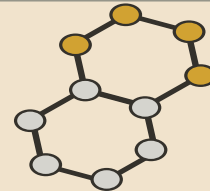
Par :
Sarah
OUHOCINE

Professeur : Sandrine VIAL

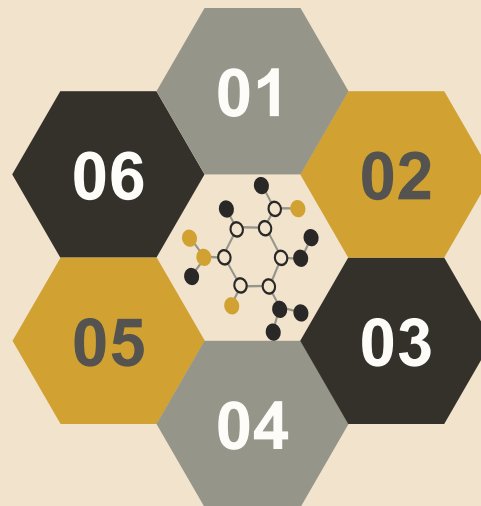




Sommaire

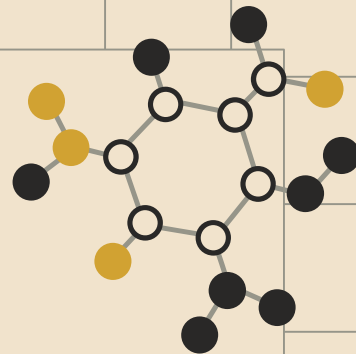


01	Introduction
02	Chimio-informatique
03	Préparation et extraction de données
04	Modèles de Machine Learning supervisé
05	Modèles de Machine Learning non supervisé
06	Conclusion



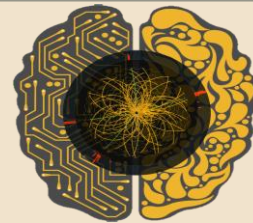
01

Introduction





Introduction

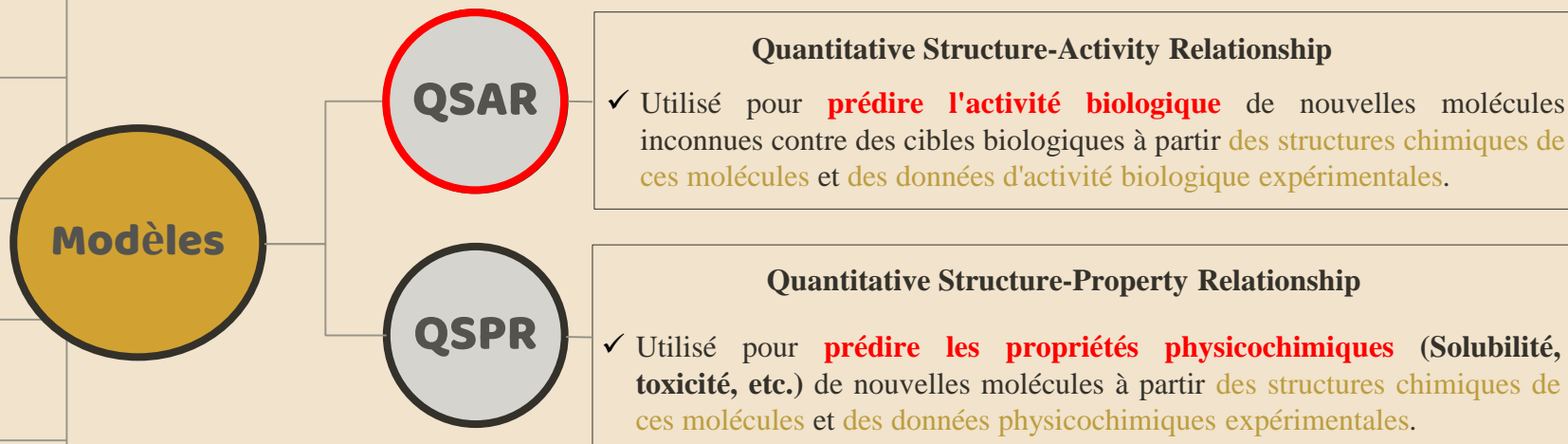


- Importance du **machine learning** en :
 - ✓ **chimio-informatique**
 - ✓ **découverte de médicaments**
- Utilise des applications spécifiques des techniques de **machine learning** pour accélérer le processus de **découverte de médicaments**.





✓ Applications spécifiques

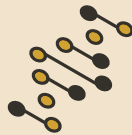


- ✓ Modèles basées sur des modèles statistiques pour identifier des relations entre les **caractéristiques structurales d'une molécule** et ses **propriétés biologiques ou physicochimiques**.



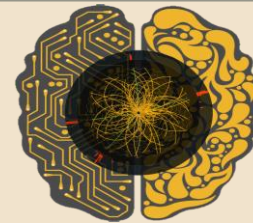
02

Chimio-Informatique





chimio-informatique



- ✓ La découverte de médicament est un **processus complexe**
- ✓ La découverte de médicament implique 4 étapes principales :

1. Identification de la cible biologique
2. Conception des molécules
3. Évaluation des molécules
4. Développement du médicament





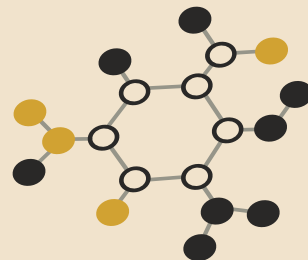
1. Identification de la cible biologique

- ✓ Identifier la cible biologique impliquée dans la maladie qu'ils cherchent à traiter.



← Cible biologique identifiée

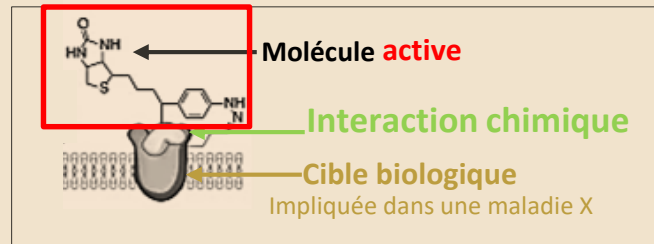
- ✓ La cible biologique peut être :
 - ☐ une protéine
 - ☐ une enzyme
 - ☐ un récepteur ou une autre molécule présente dans le corps humain → qui joue un rôle clé dans la maladie.



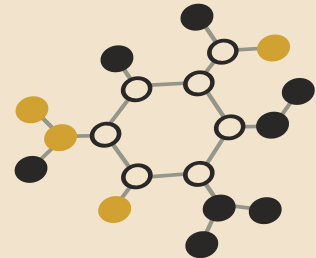
2. Conception des molécules



- ✓ Une fois la cible biologique identifiée
- ✓ chercher à identifier des molécules capables de se lier à cette cible biologique de manière spécifique et sélective (**Interaction chimique avec la cible**)



- ✓ Le but : détecter l'activité de ces molécules et traiter la maladie.
- ✓ Ces molécules s'appellent **des molécules actives** (candidates) ou **inactives** selon leur interaction ou non avec la cible.





3. Évaluation des molécules



- ✓ Les molécules **actives** (candidates) doivent être testées pour évaluer leur efficacité.
- ✓ Ces tests incluent des essais sur des :
 - ☐ Cellules en laboratoire.
 - ☐ Tests sur des animaux.
 - ☐ Essais cliniques sur des êtres humains.

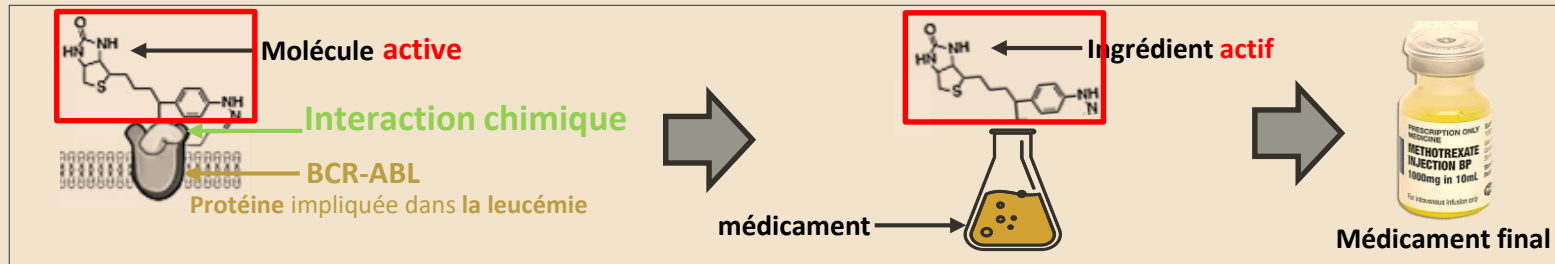




4. Développement du médicament



- ✓ Si la molécule **active** est efficace alors :
 - commencer à développer un médicament en utilisant cette molécule **active** comme **ingrédient actif**.



Si une molécule est capable de se lier spécifiquement à la protéine **BCR-ABL** (cible biologique impliquée dans la leucémie) et d'inhiber son activité, elle peut être utilisée pour traiter la leucémie.





Méthodes de recherche de nouveaux médicaments

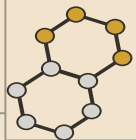


- ✓ Méthodes traditionnelles.
- ✓ Méthodes impliquant souvent la synthèse de **grandes bibliothèques de molécules**, qui sont ensuite **testées** pour leur **activité** contre une cible biologique spécifique.
- ✓ Méthodes sont donc :



✓ Longues

✓ Coûteuses





Méthodes de recherche de nouveaux médicaments



- ✓ Méthodes récentes apparues .
- ✓ Méthodes capables d'accélérer le processus de découverte de médicaments
- ✓ Méthodes basées sur l'apprentissage automatique (Machine learning)

Champ d'étude d'intelligence artificielle qui se fonde sur des algorithmes pour donner aux machines la capacité d'apprendre seul à partir des données sans être explicitement programme pour le faire.

- ✓ Méthodes utilisent des algorithmes d'apprentissage automatique
 - ✓ Supervisés
 - ✓ Non Supervisés





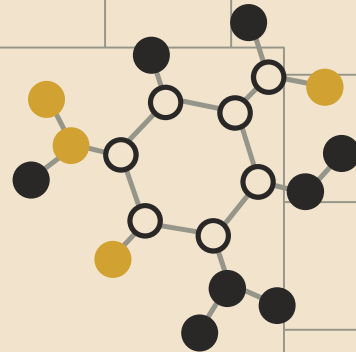
Pour utiliser ces techniques de machine learning supervisé et non supervisé il est nécessaire d'avoir :

✓ **données de bonne qualité.**



03

Préparation et Extraction de données





Préparation de données

01

- ✓ Collecter les données moléculaires auprès de ChEMBL

une base de données publique en ligne contenant des informations sur des molécules bioactives, des cibles biologiques et des essais de médicaments.



02

- ✓ Nettoyer les données en supprimant / corrigeant
 - les données dupliquées
 - les valeurs manquantes
 - les données erronées etc.

03

- ✓ Transformer les données moléculaires brutes en un format de fichier standardisé pour stocker des informations sur les structures moléculaires, (coordonnées atomiques et de liaisons chimiques).

04



- ✓ Convertir la structure moléculaire en un graphe où chaque nœud représente un atome et chaque arête représente une liaison chimique.

Cette représentation sous forme de graphe permet de capturer la structure de la molécule pour ensuite être utilisée pour extraire des données à partir de cette molécule




✓ Jeu de données



Caractéristiques moléculaires

Étiquette

ID	Nom de la molécule	SMILES		Cible biologique	Nb. de liaisons	Nb. d'atomes	Activité biologique
CHEMBL112	Caffeine	<chem>Cn1cnc2c1c(=O)n(C)c(=O)n2C</chem>		Adenosine receptor A2A	34	18	oui
CHEMBL415	Theophylline	<chem>Cn1cnc2c1c(=O)n(C)c(=O)n2C</chem>		Adenosine receptor A2A	34	18	non
CHEMBL28	Paracetamol	<chem>CC(=O)NC1=CC=C(C=C1)O</chem>		Enzyme COX-2	24	11	oui
CHEMBL25	Aspirin	<chem>CC(=O)OC1=CC=CC=C1C(=O)O</chem>		Enzyme COX-1	21	13	oui
CHEMBL521	Ibuprofen	<chem>CC(C)CC1=CC=C(C=C1)C(C)C(=O)O</chem>		Enzyme COX-1	30	18	oui



Extraction de données

- ❑ extraire des informations supplémentaires utiles et pertinentes sur la structure moléculaire
But : aider à améliorer la précision et les performances des modèles de machine learning

01

- ✓ Extraire de descripteurs chimiques à partir du graphe de la molécule

- Exemple : le poids, la surface, la polarité, les groupes fonctionnels

Un **groupe fonctionnel** est un ensemble d'atomes dans une molécule qui confère des propriétés chimiques spécifiques à cette molécule.

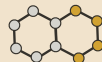
les **groupes fonctionnels** peuvent être représentés sous forme de sous-graphes dans la molécule.

02

- ✓ Calculer des empreintes chimiques à partir du graphe de la molécule

- Les empreintes sont des **représentations binaires** de la structure moléculaire.

- Exemple : 0011011001



- Méthode de calcul : **Morgan**

- ✓ utilise des algorithmes de marche aléatoire pour explorer le graphe moléculaire et générer des fragments (une partie de la molécule qui est séparée du reste de la molécule en coupant une ou plusieurs liaisons chimiques)

- ✓ les fragments sont ensuite codés en binaire pour former l'empreinte.

03

- ✓ Analyser la similarité entre les molécules à partir des empreintes.

- Exemple :

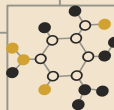
Empreinte 1: 0011**101**1001

Empreinte 2 : 110**101**0010

- Méthode de calcul : **Tanimoto**

$$\text{Similarité} = \frac{4}{20 - 4}$$

- ✓ le nombre de bits communs divisé par le nombre total de bits dans les deux empreintes - nombre de bits communs



✓ Nouveau jeu de données


Caractéristiques moléculaires

Descripteurs

Empreintes

similarités

Étiquette

ID	Nom de la molécule	SMILES		Cible biologique	Nb. de liaisons	Nb. d'atomes	Poids moléculaire	Surface moléculaire	Polarité	Groupes fonctionnels	Empreinte chimique	Similarité	Activité biologique
CHEMBL112	Caffeine	<chem>Cn1cnc2c1c(=O)n(C)c(=O)n2C</chem>		Adenosine receptor A2A	34	18	194.19	253.09	polaire	amine, amide, hétérocycles	empreinte1 110011	0.0	oui
CHEMBL415	Theophylline	<chem>Cn1cnc2c1c(=O)n(C)c(=O)n2C</chem>		Adenosine receptor A2A	34	18	180.16	244.36	polaire	amine, amide, hétérocycles	empreinte2 100001	0.43	non
CHEMBL28	Paracetamol	<chem>CC(=O)NC1=CC=C(C=C1)O</chem>		Enzyme COX-2	24	11	151.16	200.56	polaire	amine, amide, phénol	empreinte3 101111	0.19	oui
CHEMBL25	Aspirin	<chem>CC(=O)OC1=CC=CC=C1C(=O)O</chem>		Enzyme COX-1	21	13	180.16	238.66	polaire	acide carboxylique, phénol	empreinte4 101010	0.27	oui
CHEMBL521	Ibuprofen	<chem>CC(C)CC1=CC=C(C(=C1)C(C)C(=O)O</chem>		Enzyme COX-1	30	18	206.28	266.38	polaire	acide carboxylique, hétérocycles	empreinte5 110001	0.21	oui

✓ Les caractéristiques sont utilisées pour entraîner plusieurs modèles de machine learning **supervisés** et **non supervisés**

04

Modèles de machine learning supervisé





Modèles d'apprentissage supervisé



- ✓ Les modèles sont entraînés à partir de données moléculaire d'entraînement étiquetées i.e. variable cible connue (**activité biologique**).
- ✓ Les modèles sont ensuite utilisés pour **prédire** la variable cible (**activité biologique**) pour de nouvelles données (molécules) en fonction de leur caractéristiques moléculaires.

Caractéristiques moléculaires											Étiquette variable cible	
ID	Nom de la molécule	SMILES	Cible biologique	Nb. de liaisons	Nb. d'atomes	Poids moléculaire	Surface moléculaire	Polarité	Groupes fonctionnels	Empreinte chimique	Similarité	Activité biologique
CHEMBL112	Caffeine	Cn1cnc2c1c(=O)n(C)c(=O)n2C	Adenosine receptor A2A	34	18	194.19	253.09	polaire	amine, amide, hétérocycles	empreinte1	0.0	oui
CHEMBL415	Theophylline	Cn1cnc2c1c(=O)n(C)c(=O)n2C	Adenosine receptor A2A	34	18	180.16	244.36	polaire	amine, amide, hétérocycles	empreinte2	0.43	non
CHEMBL28	Paracetamol	CC(=O)NC1=CC=C(C=C1)O	Enzyme COX-2	24	11	151.16	200.56	polaire	amine, amide, phénol	empreinte3	0.19	oui
CHEMBL25	Aspirin	CC(=O)OC1=CC=CC=C1C(=O)O	Enzyme COX-1	21	13	180.16	238.66	polaire	acide carboxylique, phénol	empreinte4	0.27	oui
CHEMBL521	Ibuprofen	CC(C)CC1=CC=C(C(=C1)C(C)C(=O)O	Enzyme COX-1	30	18	206.28	266.38	polaire	acide carboxylique, hétérocycles	empreinte5	0.21	?

Données moléculaires d'entraînement

Nouvelle donnée moléculaire



K plus proches voisins (KNN)

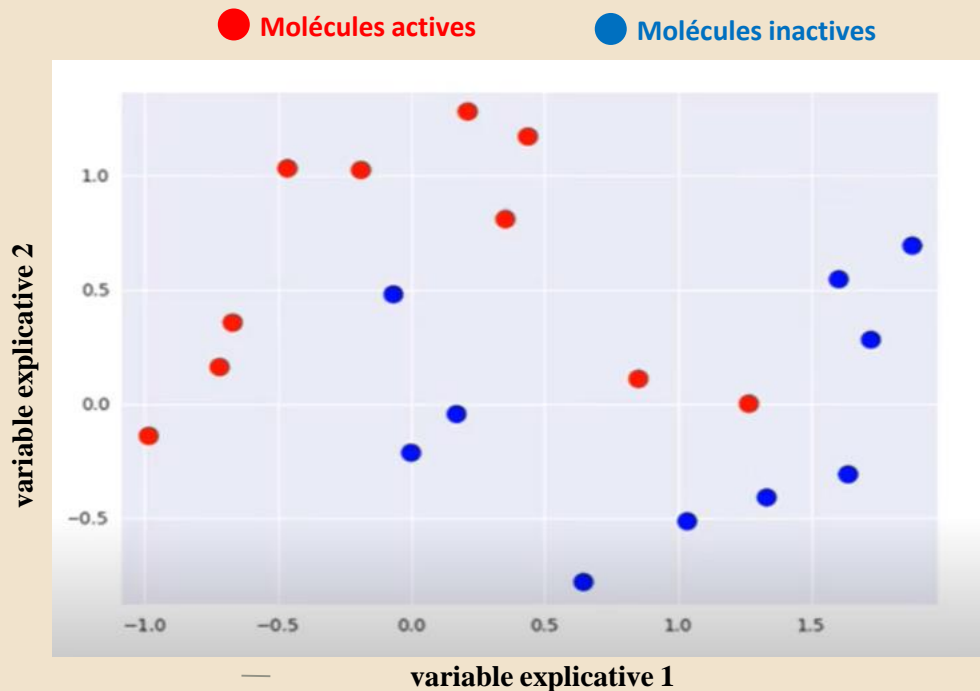


Prédire **l'activité biologique** d'une molécule en :

- ✓ **1.** Recherchant les k molécules les plus similaires (proches) de la base de données d'entraînement en termes de caractéristiques des molécules.
- ✓ **2.** Prenant la classe (**active** | **inactive**) constituant la majorité de ces molécules.



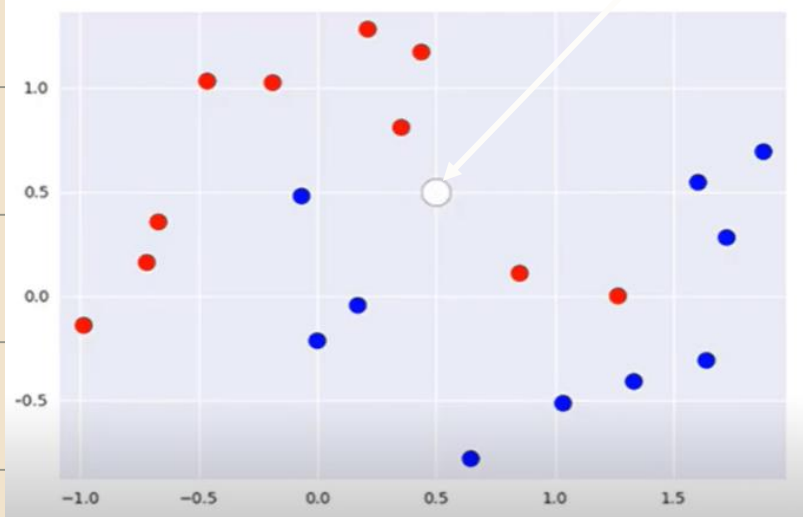
✓ Jeu de données d'entraînement avec deux classes :



✓ La variable cible est la couleur à classer soit **bleue** ou bien **rouge** (l'**activité** ou **non** de la molécule).

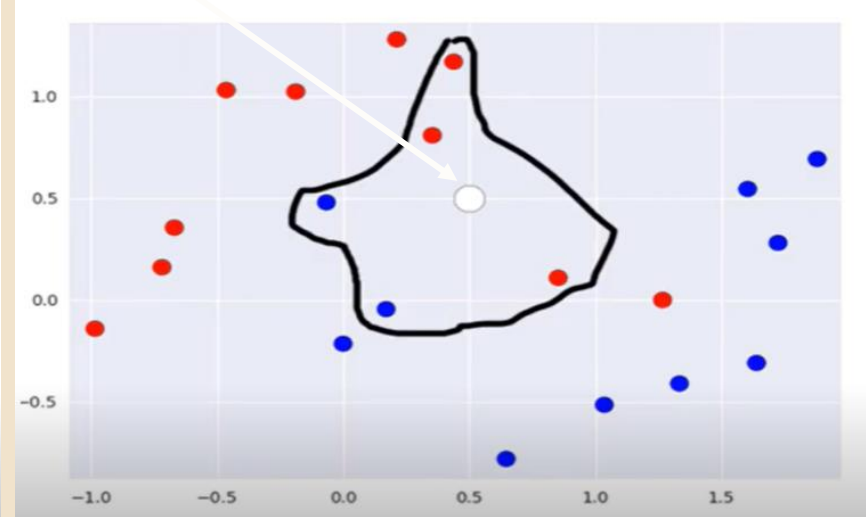
K-NN

nouvelle donnée



- ✓ regarder les K voisins les plus proches de ce point.
- ✓ regarder quelle classe constitue la majorité de ces points afin d'en déduire la classe du nouveau point.

5-NN



- ✓ exemple : **K=5** (les cinq plus proches voisins)
- ✓ Prédiction : la nouvelle donnée appartient à la classe **rouge** (3 points rouges et 2 points bleus dans son entourage).



Algorithme du modèle KNN

-1-

Sélectionner le nombre
K de voisins

-2-

Calculer la distance

$$\sum_{i=1}^n |x_i - y_i|$$

Euclidienne

$$\sqrt{\sum_{i=1}^n (x_i - y_i)^2}$$

Manhattan

-3-

Prendre les K voisins les plus proches selon la
distance calculée.

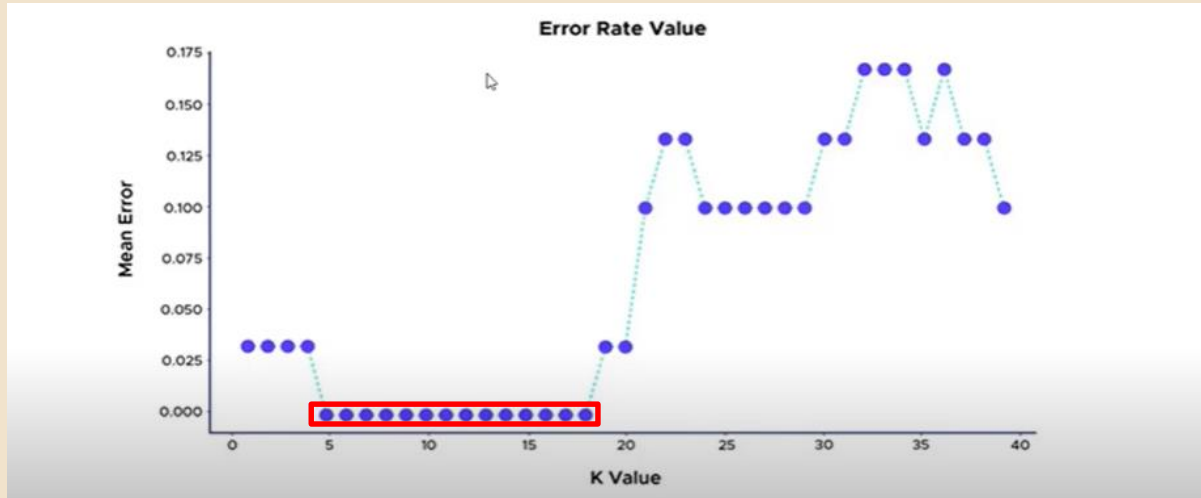
-4-

Parmi ces K voisins, compter le nombre de
points appartenant à chaque catégorie.

-5-

Attribuer le nouveau point à la catégorie la plus
présente parmi ces K voisins.

- ✓ **La question : comment choisir le bon nombre de voisins k pour lequel la classification soit meilleure ?**
- ✓ **La réponse : en fonction de l'erreur de généralisation du modèle**
il faut choisir un k optimal qui minimise au max la fonction d'erreur



variation de la fonction d'erreur en fonction de valeurs de K.

- ✓ les meilleures prédictions obtenues sont entre un **k = 5 et 18** où l'erreur a été minimisée au max (presque nulle).

05

Modèles de machine learning non supervisé





Modèles d'apprentissage non supervisé



- ✓ Les modèles sont entraînés à partir de données moléculaire d'entraînement non étiquetées i.e. variable cible pas connue.
- ✓ L'algorithme essaye **seul** de trouver des similarités et distinctions au sein de ces données moléculaires en **regroupant ensemble celles qui partagent des caractéristiques communes** afin de classifier les molécules de notre jeu de données en groupes (**clusters**).

Caractéristiques moléculaires											Étiquette variable cible	
ID	Nom de la molécule	SMILES	Cible biologique	Nb. de liaisons	Nb. d'atomes	Poids moléculaire	Surface moléculaire	Polarité	Groupes fonctionnels	Empreinte chimique	Similarité	Activité biologique
CHEMBL112	Caffeine	Cn1cnc2c1c(=O)n(C)c(=O)n2C	Adenosine receptor A2A	34	18	194.19	253.09	polaire	amine, amide, hétérocycles	empreinte1	0.0	oui
CHEMBL415	Theophylline	Cn1cnc2c1c(=O)n(C)c(=O)n2C	Adenosine receptor A2A	34	18	180.16	244.36	polaire	amine, amide, hétérocycles	empreinte2	0.43	non
CHEMBL28	Paracetamol	CC(=O)NC1=CC=C(C=C1)O	Enzyme COX-2	24	11	151.16	200.56	polaire	amine, amide, phénol	empreinte3	0.19	oui
CHEMBL25	Aspirin	CC(=O)OC1=CC=CC=C1C(=O)O	Enzyme COX-1	21	13	180.16	238.66	polaire	acide carboxylique, phénol	empreinte4	0.27	oui
CHEMBL521	Ibuprofen	CC(C)CC1=CC=C(C(=C1)C(C)C(=O)O)	Enzyme COX-1	30	18	206.28	266.38	polaire	acide carboxylique, hétérocycles	empreinte5	0.21	oui

Données moléculaires d'entraînement



K-Means



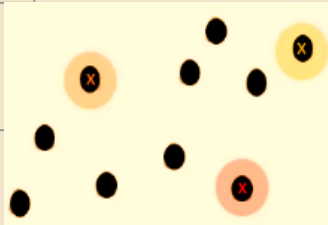
- ✓ Algorithme de Clustering.
- ✓ Étant donné **des points** et un **entier K** alors l'algorithme vise à diviser les points en k groupes appelés clusters homogènes et compacts.



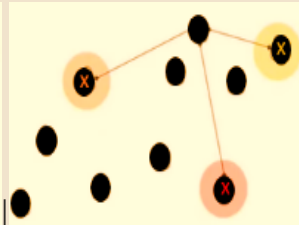
Exemple : (entrées)



- ✓ **10 données moléculaires** (points noirs)
- ✓ **K=3** (le nombre de clusters qu'on souhaite former)



- ✓ tirer aléatoirement 3 centroides (x).
- ✓ ces 3 centroides correspondent aux centres initiaux de nos 3 classes.



- ✓ calculer la distance entre les points et les centroides.
- ✓ utiliser par exemple **la distance euclidienne.**



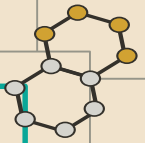
- ✓ affecter chaque point au centroïde le plus proche.



- ✓ calculer les centres de gravité des clusters qui deviennent les nouveaux centroides



- ✓ recommence les étapes suivantes pour affecter chaque point au centroïde le plus proche jusqu'à ce que les nouveaux centroides ne bougent plus des précédents.

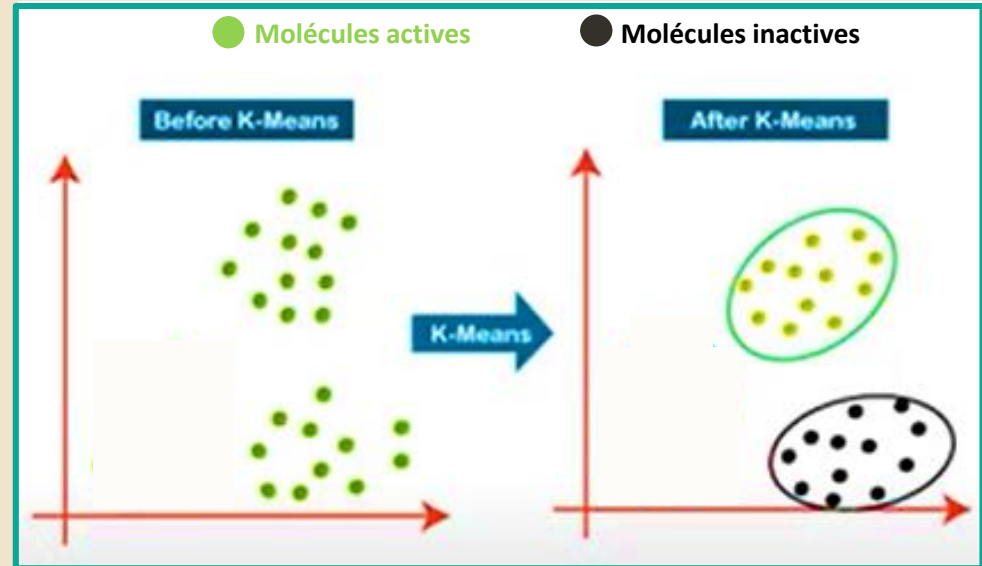


K-means c'est un algorithme itératif qui fonctionne en deux étapes :

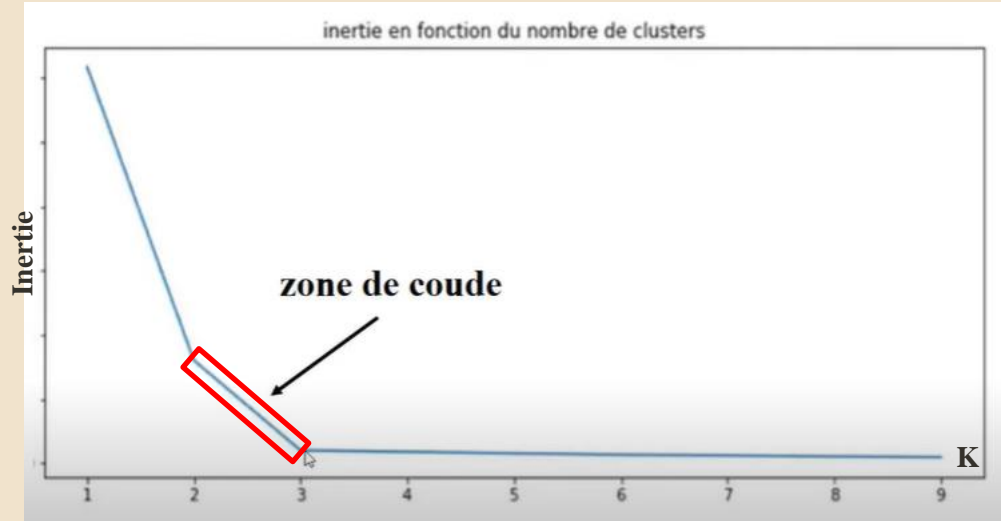
- ✓ affectation des points au centre le plus proche.
- ✓ déplacement du centre à la moyenne du cluster par le calcul de centres de gravité.

Pour classer les molécules en fonction de leur activité biologique (dans notre cas) :

- ✓ le nombre idéal de clusters est égale à 2 (**active** / **inactive**) car on connaît déjà la réponse étant donné qu'on possède déjà le jeu de données étiqueté.



- ✓ **La question : comment choisir le nombre idéal de clusters k pour lequel la classification soit meilleure ?**
- ✓ **La réponse : La méthode de coude « Elbow »**



INERTIE : peut être considérée comme une fonction d'erreur ou une fonction de coût, car elle mesure l'écart entre **chaque point de données** et son **centroïde** associé.

ZONE DE COUDE : cette zone indique le nombre de clusters K optimal qui réduit au max l'inertie de notre modèle

- ✓ À partir du 3ème cluster on remarque qu'il y a un changement très minime dans l'inertie
→ le nombre idéal de cluster est = 3.

06

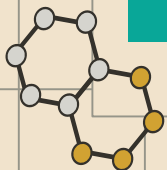
Conclusion





Conclusion

- ✓ **Combinaison** des techniques de machine learning **supervisé** et **non supervisée** pour améliorer la précision des prédictions de l'activité des molécules contre une cible biologique donnée.
- ✓ Malgré leurs **avantages**, l'utilisation du machine learning en chimio-informatique et en découverte de médicaments est **limité** :
 - ❑ les modèles de machine learning sont souvent complexes et difficiles à interpréter, ce qui peut rendre difficile la compréhension de la relation entre l'activité biologique et les prédictions du modèle.
 - ❑ Utilisation des approches complémentaires pour valider les résultats à partir de modèles de machine learning telles que **les expériences en laboratoire**.





MERCI DE VOTRE ATTENTION