1. Introduction

La recherche de nouveaux candidats médicamenteux constitue un défi majeur dans le domaine de la pharmacologie et de la chimie médicinale. Dans ce contexte, la règle de Lipinski, formulée en 1997, représente un jalon important. Selon cette règle, une molécule est considérée comme "drug-like" si elle respecte certaines propriétés physico-chimiques : un poids moléculaire inférieur à 500 daltons, moins de 10 accepteurs de liaisons hydrogène, moins de 5 donneurs de liaisons hydrogène, une valeur de logP inférieure à 5 et supérieure à -2, et moins de 5 liaisons rotables.
L'objectif de ce travail est donc de réévaluer la définition des molécules "drug-like" en exploitant une base de données comprenant des molécules classées comme "drug-like" et "non drug-like", décrites par un ensemble de 20 descripteurs physico-chimiques.

2. Méthodes et résultats

a) Jeu de données :

Le jeu de données analysé comprend 2,866 observations de molécules (862 dl et 1995 ndl), chacune décrite par 21 descripteurs physico-chimiques et une colonne cible nommée drug qui indique si la molécule est considérée comme drug-like ou non.
Les descripteurs physico-chimiques incluent le radius, a_acc (nombre d'accepteurs de liaison hydrogène), a_acid, a_base, a_don (nombre de donneurs de liaison hydrogène), a_hyd, SlogP (logarithme du coefficient de partage octanol/eau, un indicateur de lipophilicité), a_nB, a_nBr, a_nC, a_nCl, a_nF, a_nI, a_nN, a_nO, a_nP, a_nS, a_aro (nombre d'anneaux aromatiques), a_count (nombre total d'atomes), et a_nH. La variable cible drug indique si une molécule est drug-like (druglike) ou non (nondruglike), selon les critères définis dans l'étude.

b) Nettoyage et éliminations des outliers

Pour assurer la qualité de notre analyse, nous avons procédé à une élimination des observations incomplètes, garantissant ainsi que notre étude repose sur des données fiables et complètes. L'identification de valeurs extrêmes pour chaque descripteur nous a permis de détecter et d'exclure des anomalies spécifiques, comme une valeur exceptionnellement élevée du rayon.

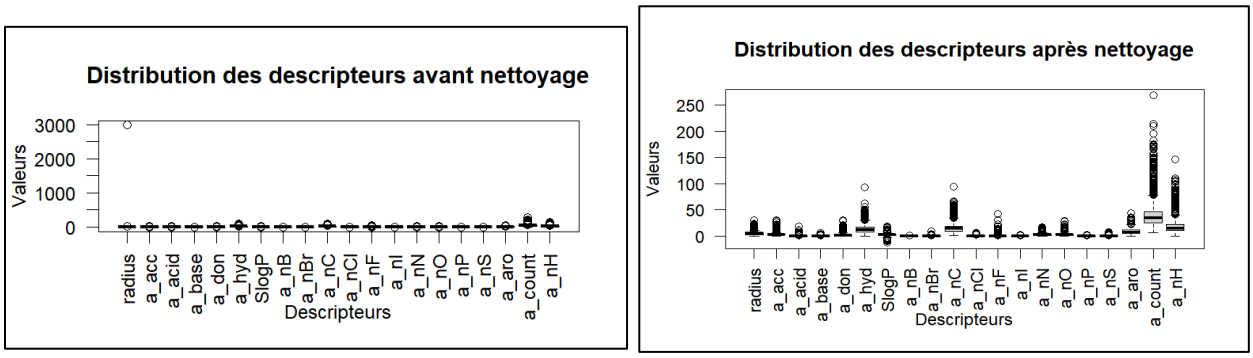

Figure 1: Boxplot montrant la distribution des descripteurs du dataset avant et après nettoyage et élimination de valeurs extrêmes.

On observe qu'avant nettoyage, à part pour le radius où on trouve une valeur extrême > 2700, les autres descripteurs ne semblent pas avoir de valeurs aberrantes. Après nettoyage certaines valeurs de a_count supérieures à 250 et de a_nH supérieures à 100 semblent être des outliers.
Ensuite, nous avons réalisé une analyse en composantes principales (PCA) afin d'explorer la variabilité et les tendances principales au sein des données physico-chimiques des molécules. Les points représentent les

molécules projetées sur les deux premières composantes principales, qui expliquent respectivement 27.55% et 14.27% de la variance totale. Les numéros indiquent les identifiants des molécules présentant des valeurs extrêmes sur ces composantes.

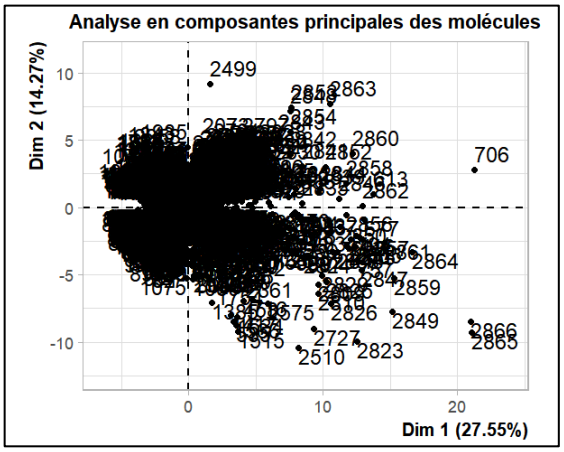

Figure 2 : Analyse en composantes principales (PCA) des données physico-chimiques des molécules.

Avant le nettoyage des données, à l'exception d'une valeur extrême de radius supérieure à 2700, les valeurs des autres descripteurs ne semblaient pas contenir d'outliers significatifs. Cependant, après nettoyage, des valeurs particulièrement élevées pour a_count et a_nH ont été identifiées comme potentiellement extrêmes. Les individus correspondants ont été isolés : l'individu 703 pour a_count, et les individus 610, 703, 2849, 2852, 2854 pour a_nH.
Lors de la visualisation par PCA, ces points extrêmes ne se distinguent pas de façon prononcée par rapport à l'ensemble des données, ce qui suggère que leur impact sur les deux principales composantes est minime.

c) Vérification des paramètres pour Y

Puis, une analyse statistique a été réalisée pour évaluer la signification des différentes variables par rapport à leur capacité à distinguer entre des molécules "drug-like" et "non drug-like". Pour chaque variable, un Student's t-test a été utilisé pour comparer les moyennes des deux groupes et déterminer si les différences observées sont statistiquement significatives. La p-value est une mesure statistique qui aide à décider si l'hypothèse nulle (aucune différence de moyenne entre les groupes) peut être rejetée. Le seuil pour déterminer la signification est fixé à 0.05. Les variables avec des pvalues inférieures à ce seuil sont considérées comme significatives, ce qui suggère qu'il existe une différence statistiquement significative pour ces variables entre les molécules DL et nDL . En conclusion, les variables "a_nB" et "a_nBr" ont été identifiées comme non significatives et ont donc été exclues du jeu de données nettoyé, en vue d'une analyse plus approfondie.

radius	a_acc	a_acid	a_base	a_don
1.447792e-70	3.527258e-29	1.513008e-02	1.804374e-06	6.166498e-36
a_hyd	SlogP	a_nB	a_nBr	a_nC
7.870503e-08	1.222480e-02	1.825509e-01	4.461058e-01	9.612260e-12
a_nCl	a_nF	a_nI	a_nN	a_nO
3.968229e-10	8.418186e-11	1.786444e-02	1.570227e-11	3.967946e-09
a_nP	a_nS	a_aro	a_count	a_nH
4.839028e-03	1.371692e-02	1.002983e-02	2.563977e-18	4.039543e-19

Figure 3 : Pvalues des t-test sur les différentes variables pour comparer les moyennes des 2 groupes (druglike et nondruglike).

Après avoir retiré les variables non significatives, une matrice de corrélation de Pearson a été calculée sur le jeu de données nettoyé.

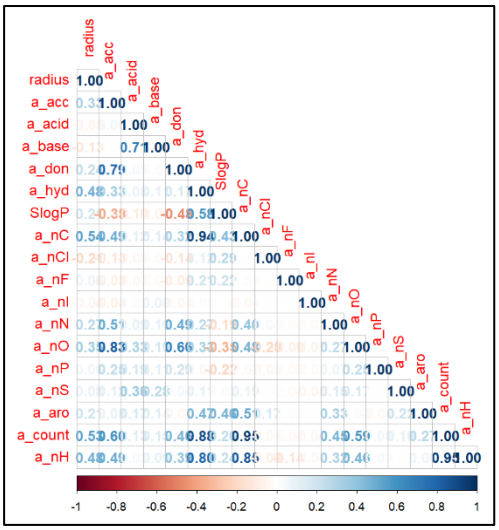

Figure 3 : Matrice de corrélation de Pearson des descripteurs physico-chimiques des molécules. Les valeurs numériques dans chaque case reflètent l'intensité de la corrélation entre les paires de descripteurs.

La matrice de corrélation illustre les relations linéaires entre les descripteurs, avec un accent particulier sur les liens entre le nombre de carbones (a_nC), le nombre d'hydrogènes (a_nH), le compte total d'atomes (a_count) et l'hydrophilie (a_hyd). Les corrélations supérieures à 0.9 entre ces variables suggèrent une forte liaison, cependant, dans un contexte chimique, ces liens sont attendus et représentent des aspects fondamentaux de la structure moléculaire.

d) Construction des échantillons et du modèle

Pour la construction du modèle prédictif, il est crucial de diviser le jeu de données en deux ensembles distincts : un pour l'apprentissage et l'autre pour la validation. La répartition traditionnelle suit souvent une règle de deux tiers pour l'apprentissage et un tiers pour la validation. Afin d'assurer l'équilibre et la représentativité des échantillons, la méthode sample est utilisée pour réaliser un échantillonnage aléatoire des données. Cette approche est appliquée séparément aux données des molécules qualifiées de "druglike" et "nondruglike".

Pour l'échantillon d'apprentissage, deux tiers des données de chaque catégorie (DL et NDL) sont sélectionnés au hasard. Le tiers restant constitue l'échantillon de validation. Ces échantillons sont ensuite fusionnés séparément : l'ensemble d'apprentissage avec les deux tiers et l'ensemble de validation avec le dernier tiers.

Homogénéité des variables quantitatives : pour vérifier l'homogénéité entre les ensembles d'apprentissage et de validation, un test t de Student pour échantillons indépendants est appliqué sur chaque variable quantitative. Le test t est utilisé ici pour déterminer s'il existe une différence statistiquement significative entre les moyennes des deux ensembles pour chaque variable. Les pvalues obtenues pour chaque variable sont ensuite comparées au seuil de 0.05. Dans ce cas, les pvalues indiquent qu'aucune différence significative n'a été détectée entre les échantillons d'apprentissage et de validation pour les descripteurs, suggérant une bonne homogénéité et permettant de procéder avec le modèle construit sur ces ensembles.

```
   radius     a_acc    a_acid    a_base     a_don     a_hyd     SlogP
0.4182403 0.9018979 0.6385030 0.3716075 0.4333698 0.6655107 0.2979738
     a_nC     a_nCl      a_nF      a_nI      a_nN      a_nO      a_nP
0.7590455 0.7553651 0.5746784 0.2012410 0.8248570 0.7418411 0.8265824
     a_nS     a_aro   a_count      a_nH
0.5523001 0.4045530 0.9155406 0.6877594
```

Figure 4 : Pvalues des t-test sur les différentes variables pour comparer les moyennes des 2 échantillons.

En superposant les histogrammes des deux ensembles, on peut comparer directement les distributions et s'assurer que les modèles entraînés sur l'ensemble d'apprentissage seront pertinents et applicables aux données de validation.

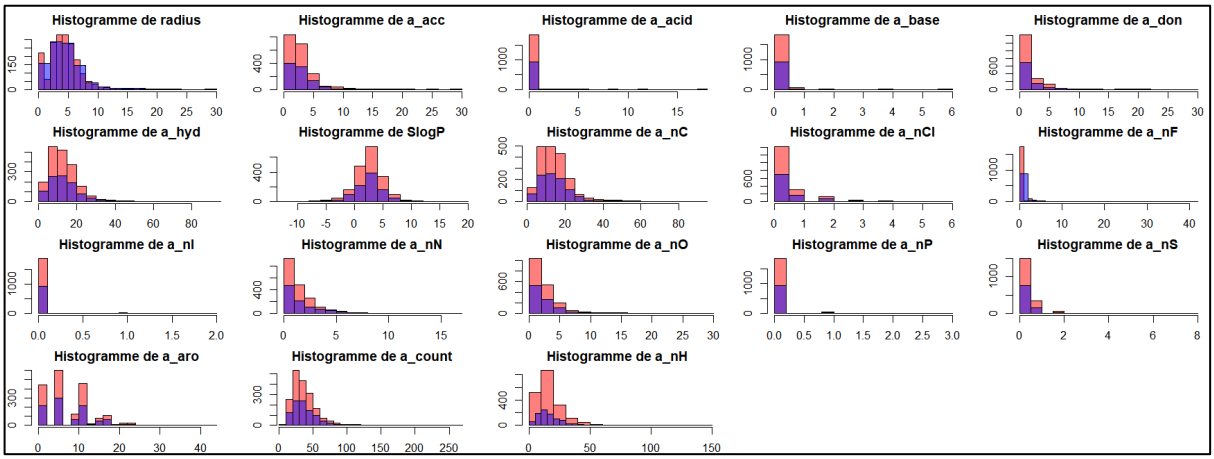

Figure 5 : Série d'histogrammes qui compare la distribution des variables quantitatives entre les ensembles d'apprentissage et de validation. Chaque histogramme affiche la fréquence des valeurs pour une variable spécifique, avec l'ensemble d'apprentissage en rouge et l'ensemble de validation en bleu.

Les histogrammes sont superposés pour une comparaison directe, montrant l'homogénéité entre les deux ensembles, ce qui est crucial pour l'entraînement du modèle.

Homogénéité de la variable qualitative : l'homogénéité des proportions de la variable cible entre les ensembles d'apprentissage et de validation est cruciale pour la validation d'un modèle prédictif. Pour cela, un test du Chi-deux de Pearson a été réalisé pour vérifier l'indépendance entre les ensembles. L'hypothèse nulle (H0) pour ce test est qu'il n'y a pas de différence significative dans la répartition de la variable cible (ici, "drug") entre les deux ensembles.
Les résultats du test Chi-deux de Pearson montrent une valeur très faible de X-squared et une p-value très élevée (0.9931), bien au-dessus du seuil habituel de 0.05. Cela indique qu'il n'y a pas de preuve statistique pour rejeter l'hypothèse nulle, ce qui signifie que les ensembles d'apprentissage et de validation sont homogènes en ce qui concerne la proportion de molécules qualifiées de "druglike" et "nondruglike".

```
        Pearson's Chi-squared test

data:  contingency_table
X-squared = 7.5608e-05, df = 1, p-value = 0.9931
```

Figure 5 : Résultat du Chi² d'homogénéité

e) Apprentissage

La régression logistique est appropriée dans ce contexte car elle permet de gérer une variable de réponse binaire et de produire des probabilités. Dans le modèle de régression logistique, chaque prédicteur (variable indépendante) est associé à un coefficient qui mesure son effet logarithmique sur la probabilité que la molécule soit considérée comme "druglike". L'hypothèse nulle (H0) postule que ce coefficient est égal à zéro, ce qui suggérerait que le prédicteur n'a aucun impact sur la probabilité de l'issue. À l'opposé, l'hypothèse alternative (H1) soutient que le coefficient est différent de zéro et donc que le prédicteur a un effet significatif.

Les pvalues dans le résumé du modèle sont des indicateurs de la significativité statistique des prédicteurs sur la variable dépendante, ici la classification des molécules en tant que "druglike" ou non. L'intercept a une pvalue extrêmement faible (7.07e-06) ce qui indique que le modèle est significatif.
L'analyse du modèle met en lumière l'importance de caractéristiques spécifiques qui pourraient compléter ou affiner la règle des cinq de Lipinski pour définir une molécule comme "druglike". En particulier, le radius moléculaire et l'acceptation de liaisons hydrogène (a_acc) se distinguent comme des facteurs clés, suggérant que la petite taille et une capacité élevée à établir des interactions hydrogène pourraient être cruciales pour la bioactivité.
En outre, la lipophilie confirme l'importance de l'équilibre entre solubilité et perméabilité. Cela suggère qu'une attention plus nuancée au spectre lipophilique pourrait améliorer les critères de sélection des candidats médicamenteux, au-delà des seuils généraux proposés par Lipinski.
L'effet significatif des variables a_nC (nombre d'atomes de carbone), a_nI (nombre d'atomes d'iode), et a_nP (nombre d'atomes de phosphore) sur la "druglikeliness" indique que la composition atomique spécifique joue également un rôle dans la détermination du potentiel médicamenteux d'une molécule.

L'évaluation du modèle révèle un taux d'erreur de 15.65% et un taux de bonnes prédictions de 84.35%, soulignant une efficacité notable dans la classification des molécules comme "druglike" ou "non-druglike". La spécificité élevée (95.49%) démontre que le modèle excelle à identifier correctement les molécules "non-druglike", minimisant ainsi les faux positifs. Cette performance suggère que le modèle est particulièrement apte à éliminer les molécules ne correspondant pas aux critères "druglike".
La sensibilité (58.54%), bien qu'inférieure à la spécificité, indique la capacité du modèle à détecter les véritables molécules "druglike".
Toutefois, cette sensibilité plus modeste peut en partie s'expliquer par la distribution déséquilibrée des données, avec 1995 molécules "non-druglike" contre seulement 864 "druglike". Ce déséquilibre dans le jeu de données peut influencer la capacité du modèle à identifier correctement les molécules "druglike", puisque la majorité des exemples d'entraînement sont classés comme "non-druglike", ce qui pourrait biaiser le modèle en faveur de cette catégorie.

Dans le processus d'optimisation de notre modèle, nous avons employé la méthode stepwise pour affiner la sélection des variables prédictrices. Cette approche permet d'ajuster le modèle en ajoutant ou supprimant des variables basées sur leur impact sur le critère AIC (Akaike Information Criterion qui permet de mesurer la qualité d'un modèle statistique), visant ainsi à améliorer la performance du modèle tout en conservant une complexité modérée.
À travers cette démarche, certaines variables ont été éliminées du modèle initial, à savoir a_acid, a_hyd, a_nCl, a_nF, a_nO, a_nS, et a_aro. Le modèle final affiné, avec un AIC réduit à 1579.8, suggère une amélioration par rapport au modèle initial.

En termes de performance de prédiction, ce modèle optimisé présente un taux d'erreur de 15.65%, un taux de bonnes prédictions de 84.35%, une sensibilité de 57.84%, et une spécificité de 95.79%. Ces résultats indiquent que le modèle est particulièrement efficace pour identifier les molécules "non drug-like", avec une spécificité élevée.

	Avant optimisation	Après optimisation
Taux d'erreur	15.65%	14.81%
Taux bien prédit	84.35%	85.18%
Sensibilité	58.54%	58.28%
Spécificité	95.49%	96.84%

Figure 6 : Résultats des paramètres optimisés

L'élimination de certaines variables dans le processus d'optimisation révèle des pistes intéressantes pour la discussion autour de la définition des molécules "drug-like".

3. Conclusion

À travers notre démarche, certaines variables ont été éliminées du modèle initial, à savoir a_acid, a_hyd, a_nCl, a_nF, a_nO, a_nS, et a_aro. Ces variables ont été jugées moins pertinentes pour prédire le caractère "drug-like" d'une molécule, ce qui suggère qu'elles ont un impact limité sur la probabilité qu'une molécule soit considérée comme potentiellement thérapeutique.

Le modèle final affiné, avec un AIC réduit à 1579.8, suggère une amélioration par rapport au modèle initial. Ce modèle inclut des variables telles que le rayon (radius), l'accepteur d'hydrogène (a_acc), la base (a_base), le donneur d'hydrogène (a_don), le logP (SlogP), le nombre d'atomes de carbone (a_nC), d'iode (a_nI), d'azote (a_nN), de phosphore (a_nP), le nombre total d'atomes (a_count), et le nombre d'atomes d'hydrogène (a_nH), comme étant significatives pour la prédiction.

L'élimination de certaines variables dans le processus d'optimisation révèle des pistes intéressantes pour la discussion autour de la définition des molécules "drug-like". Bien que certaines des variables retirées puissent sembler pertinentes d'un point de vue chimique, leur contribution limitée au modèle suggère que d'autres caractéristiques physico-chimiques pourraient être plus cruciales pour définir une molécule comme étant "drug-like".

Ainsi une définition révisée pourrait être :

- Un rayon efficace (radius) qui influence négativement la "drug-likeness", suggérant une limite supérieure au-delà de laquelle les molécules deviennent moins susceptibles d'être "drug-like".
- Un nombre acceptable d'accepteurs de liaisons hydrogène (a_acc) et de donneurs (a_don), avec une influence positive sur la "drug-likeness", soulignant l'importance de l'équilibre hydrophile/hydrophobe.
- Une considération de la présence de certains groupes fonctionnels spécifiques (comme indiqué par les variables a_nI, a_nP), qui montre une influence positive sur la "drug-likeness", suggérant l'importance de certaines caractéristiques chimiques spécifiques.
- Un nombre de liaisons rotables (a_count) qui, lorsqu'il est trop élevé, peut diminuer la probabilité qu'une molécule soit "drug-like", indiquant une limite sur la flexibilité moléculaire pour une bonne "drug-likeness".

Cette définition enrichie ne remplace pas les règles de Lipinski mais les complète en incorporant des facteurs supplémentaires liés à la structure et à la composition chimique des molécules, offrant ainsi une approche plus holistique pour évaluer le potentiel "drug-like" d'une molécule.