Développement d'une plateforme de prédiction in silico des propriétés ADME-Tox

Baptiste Canault

UNIVERSITÉ D'ORLÉANS

ÉCOLE DOCTORALE
SANTE, SCIENCES BIOLOGIQUES ET CHIMIE DU VIVANT

Laboratoire : Institut de Chimie Organique et Analytique

THÈSE présentée par :
Baptiste CANAULT

soutenue le : **1 octobre 2018**

pour obtenir le grade de : **Docteur de l'université d'Orléans**
Discipline/ Spécialité : Chimie et chémoinformatique

Développement d'une plateforme de prédiction *in silico* des propriétés ADME-Tox

THÈSE dirigée par :

Pascal BONNET	Professeur, Université d'Orléans

RAPPORTEURS :

Anne-Claude CAMPROUX	Professeur, Université Paris Diderot
Alexandre VARNEK	Professeur, Université de Strasbourg

PRÉSIDENT DE JURY :

Christel VRAIN	Professeur, Université d'Orléans

JURY :

Anne-Claude CAMPROUX	Professeur, Université Paris Diderot
Alexandre VARNEK	Professeur, Université de Strasbourg
Christel VRAIN	Professeur, Université d'Orléans
Eric ARNOULT	Docteur, GlaxoSmithKline
Pascal BONNET	Professeur, Université d'Orléans
Philippe VAYER	Docteur, Technologie Servier
Stéphane BOURG	Docteur, Université d'Orléans

A GG, JG et DG

Remerciements

Je tiens à remercier à la région Centre-Val de Loire et à l'entreprise Servier, dont les financements m'ont permis de mener à bien ce travail de thèse.

Je remercie les membres de mon jury d'avoir accepté de lire mon manuscrit, d'être venu assister à ma soutenance et d'avoir évalué mon travail : le Professeur Anne-Claude Camproux, le Professeur Alexandre Varnek, le Professeur Christel Vrain et le Docteur Eric Arnoult.

Je remercie le Professeur Pascal Bonnet de m'avoir fait confiance pour mener à bien ce projet, de m'avoir accepté au sein de son équipe, de m'avoir épaulé dans les moments de doutes.

C'est avec une grande gratitude que je remercie le Docteur Philippe Vayer qui a toujours pris le temps de discuter et de me partager ses connaissances. J'ai apprécié tes conseils et l'ouverture d'esprit que tu m'as transmis lors de cette thèse. Ta vision des choses m'a souvent permis de voir différemment les éléments auxquels j'étais confronté.

Je tiens également à remercier la plateforme ADME de la faculté de pharmacie de Lille dirigée par le Professeur Benoit Déprez d'avoir accepté de collaborer avec notre équipe pour la valorisation de notre plateforme *in silico*.

J'aimerais exprimer toute ma reconnaissance aux membres de l'équipe de Biologie Structurale & Chémoinfomratique, et de manière générale à tous les membres de l'ICOA, pour leur aide, leur soutien, leur bonne humeur et pour tous les moments agréables que nous avons passés ensemble.

Pour finir, je tiens à remercier mes amis et ma famille. Les personnes que l'on côtoie sont le reflet de ce que l'on est. Merci d'avoir fait ce que je suis aujourd'hui.

Table des matières

Table des figures

Table des tableaux

Table des équations

Liste des abréviations*

Acc	*Accuracy* (Justesse)
ACF	*Atom-Centred Fragments* (Fragments centrés sur atome)
ACP	Analyse en Composantes Principales
ADN	Acide désoxyribonucléique
ADME	Absorption, Distribution, Métabolisme et Elimination
ADME-Tox	Absorption, Distribution, Métabolisme et Elimination et Toxicité
AEM	Agence Européenne des Médicaments
AGP	Alpha-1-glycoprotéine acide
AMM	Autorisation de Mise sur le Marché
AP	*Atom Pair* (Paire d'atomes)
ARN	Acide ribonucléique
AUC	*Area Under the Curve* (Aire sous la courbe)
BBB	*Blood-Brain Barrier* (Barrière hémato-encéphalique)
BCRP	*Breast Cancer Resistance Protein* (Protéine de résistance au cancer)
CADD	*Computer-Aided Drug Design* *(Conception de médicaments assistée par ordinateur)*
CL_H	Clairance Hépatique
CL_P	Clairance Plasmatique
CL_R	Clairance Rénal
CL_T	Clairance Totale
CYP1A2	Cytochrome P450 1A2
CYP2C9	Cytochrome P450 2C9
CYP2C19	Cytochrome P450 2C19
CYP2D6	Cytochrome P450 2D6
CYP3A4	Cytochrome P450 3A4
CYP450	Cytochrome P450
DA	Domaine d'Applicabilité
DILI	*Drug Induced Liver Injury* (Lésions hépatiques induites par un médicament)
***dk*-NN**	*density k-Nearest Neighbors* (k-plus proches voisins de densité locale)
EBI	*European Bioinformatics Institut* (Institut Européen de Bioinformatique)
EC	*Extended-Connectivity* (Connectivité étendue)
EFPIA	*European Federation of Pharmaceutical Industries and Associations* (Fédération européenne des industries pharmaceutiques)
F	Biodisponibilité
F_{abs}	Fraction absorbée
FC	*Functional Connectivity* (Connectivité fonctionnelle)
FDA	*Food and Drug Administration* (Agence américaine des produits alimentaires et médicamenteux)
FMO	*Flavin-containing monooxygenase* (monooxygénase contenant de la flavine)
Fu_p	*Fraction unbound to plasma* (Fraction libre dans le plasma)
GI	Gastro-Intestinal
GTM	*Generative Topographic Mapping*
HCS	*High-Content Screening* (Criblage à haut contenu)
HDL	*High Density Lipoprotein* (Lipoprotéine haute densité)
HIA	*Human Intestinal Absorption* (Absorption Intestinal Humaine)
HSA	*Human Serum Albumin* (Albumine Humaine)
HTS	*High-Throughput Screening* (Criblage à haut débit)
HTVS	*High-Throughput Virtual Screening* (Criblage virtuel à haut débit)
IMI	Initiative pour les Médicaments Innovants (*Innovative Medicines Initiative*)
InChI	*International Chemical Identifier* (Identificateur Chimique International)
***k*-NN**	*k-Nearest Neighbors* (k-plus proches voisins)

LD_{50}	*median Lethal Dose* (Dose létale médiane)
LDL	*Low Density Lipoprotein* (lipoprotéine basse densité)
LGR	*Logistic Regression* (régression Logistique)
LMO	*Leave-Many-Out*
LOAEL	*Low Observed Adverse Effect Level* (dose minimale avec effet indésirable observé)
LOO	*Leave-One-Out*
MACCS	*Molecular ACCess System*
MAE	*Mean Absolute Error* (erreur moyenne absolue)
MCC	*Matthews Correlation Coefficient* (coefficient de corrélation de Matthews)
MMP	*Match Molecular Pair*
MLR	*Multiple Linear Regression* (*r*égression linéaire multiple)
MRP2	*Multidrug Resistance-associated Protein 2*
NAT	N-acétyl-transférase
NIH	*National Institute of Health* (institut national de la santé)
NME	*New Molecular Entity* (*nouvelle entité m*oléculaire)
NOAEL	*No Observed Adverse Effect Level* (dose sans effet nocif observé)
OCDE	Organisation de Coopération et de Développement Economique
P-gp	Glycoprotéine P
PLS	*Partial Least Square* (régression linéaire selon le critère des moindres carrés partiels)
PPB	*Plasma Protein Binding* (Fraction liée aux protéines plasmatiques)
Pre	Précision
(Q)SAR	*Quantitative or Qualitative Structure-Activity Relationship* (Etudes Quantitatives ou Qualitatives de Relation Structure-Activité)
QSAR	*Quantitative Structure-Activity Relationship* (Etudes Quantitatives de Relation Structure-Activité)
(Q)SPR	*Qualitative or Quantitative Structure-Property Relationship* (Etudes Quantitatives ou Qualitatives de Relation Structure-Propriété)
(Q)STR	*Qualitative or Quantitative Structure-Toxicity Relationship* (Etudes Quantitatives ou Qualitatives de Relation Structure-Toxicité)
REACH	*Registration, Evaluation, Autorisation & Restriction of Chemicals*
RF	*Random Forest* (Forêt aléatoire)
RMSE	*Root Mean Square Error* (Erreur quadratique moyenne)
Ro5	*Rule of five* (Règle de Lipinski)
ROC	*Receiver Operating Characteristic*
RSA	Relation Structure-Activité
RSP	Relation Structure-Propriété
RST	Relation Structure-Toxicité
Sen	Sensibilité
SNC	Système Nerveux Central
SMF	*Substructural Molecular Fragments* (Fragments moléculaires sous-structuraux)
Spe	Spécificité
SVM	*Support Vector Machine* (Machine à vastes marges)
$t_{1/2}$	Temps de demi-vie
TT	Torsion topologique
UGT	UDP-Glucuronosyl-transférase
Vd	Volume de distribution
VHDL	*Very High Density Lipoprotein* (Lipoprotéine très haute densité)

* Ces abréviations prennent en compte les dénominations usuelles. Lorsqu'une abréviation fait référence au terme anglais, la traduction française a été renseignée.

Introduction générale

Il n'y a pas de chef-d'œuvre plus incroyable et complexe dans la Nature que celui du corps humain. Le corps humain possède plusieurs niveaux de complexité dans lesquels, une molécule va pouvoir interagir avec les protéines présentes dans des cellules, qui elles-mêmes définissent des tissus biologiques répondants à des fonctions spécifiques de l'organisme. La recherche de nouveaux produits thérapeutiques a pour objectif d'apporter des solutions optimales aux dysfonctionnements du corps humain. L'enjeu principal est de développer un médicament capable de soigner les maux des patients, tout en limitant les effets indésirables. Un médicament est défini comme une molécule capable de traverser l'ensemble des niveaux de complexité du corps humain pour moduler les processus biologiques. L'action du médicament n'est possible que si ce dernier parvient à atteindre en quantité suffisante la cible thérapeutique visée. Pour cela, le médicament aura plusieurs barrières à traverser. Ces barrières sont décrites par les phénomènes complexes qui interviennent tout au long du voyage du médicament dans l'organisme, notamment lors des étapes d'absorption, de distribution, du métabolisme et de l'élimination (ADME).

Les acteurs qui travaillent à la conception de nouveaux médicaments mesurent des propriétés ADME, afin d'estimer la capacité d'une molécule à traverser l'ensemble de ces barrières, tout en limitant les effets nuisibles pour la santé (Toxicité). Cette détermination nécessite du temps et des moyens financiers important. Par conséquent, la mesure de ces propriétés dès les premières étapes de conception de médicaments est de nos jours difficile. Pour pallier à cette problématique, des méthodes informatiques peuvent être mises en œuvre pour estimer les propriétés ADME-Tox dès les étapes précoces du processus de conception de médicaments. L'intérêt principal est d'apporter une aide décisionnelle aux chimistes médicinaux, afin de prioriser les tests biologiques et d'orienter rapidement les étapes de synthèses nécessaires à la découverte d'un médicament. Ceci s'inscrit dans la dynamique actuelle suivie par la majorité des entreprises pharmaceutiques puisque cela permet d'anticiper les profils ADME-Tox défavorables et de réduire les échecs de certains projets de recherche.

La thèse que nous présentons est née d'une collaboration entre l'équipe de biologie structurale et de chémoinformatique (SB&C) de l'ICOA UMR 7311 et Technologie Servier. Elle a pour but de mettre en œuvre des modèles de prédiction fiables et robustes pour

l'estimation des propriétés ADME-Tox. En s'intéressant à ces propriétés, ce projet constitue une nouvelle thématique de recherche pour l'équipe SB&C. Il a pour enjeu de créer une plateforme *in silico* de criblage ADME-Tox basée sur des données publiques. L'enjeu pour Technologie Servier est de valoriser les données pharmacocinétiques acquises depuis plusieurs années, complétées si besoin de données de référence, afin d'optimiser et/ou d'acquérir des modèles ADME-Tox.

Le travail de cette thèse a pour objectifs : i) de constituer une vaste collection de données expérimentales pour diverses propriétés ADME-Tox, ii) de développer la plateforme MetaPredict, qui est un outil capable de créer automatiquement des modèles de prédiction fiables et robustes, et iii) de rendre l'utilisation facile et efficace des modèles générés par les chimistes. La présentation de ce projet de recherche est répartie en 3 chapitres.

Le chapitre 1 annoncera le contexte global de la conception de nouveaux médicaments. Nous y présenterons également les propriétés ADME-Tox, ainsi que les approches statistiques focalisées sur leur estimation. A la fin de ce chapitre, nous parlerons des enjeux actuels liés à la prédiction des propriétés ADME-Tox et des points d'amélioration envisagés.

Le chapitre 2 sera dédié à la recherche de données expérimentales, réalisée au cours de cette thèse. Nous nous intéresserons également aux méthodes utilisées pour l'extraction des sources de données ADME-Tox identifiées. Au cœur de ce chapitre, nous présenterons les outils informatiques que nous avons développés durant cette thèse pour extraire plus aisément les données d'intérêt.

Enfin, dans le chapitre 3, nous présenterons la méthodologie développée dans le cadre de la plateforme MetaPredict. Chaque étape de la plateforme sera décrite à l'aide d'une application sur les données de solubilité aqueuse. Nous parlerons ensuite de la validation expérimentale de la plateforme pour plusieurs modèles de prédiction de la fraction libre dans le plasma. Nous finirons par la présentation des approches complémentaires que nous avons développées pour rendre les modèles accessibles et plus informatifs pour les chimistes, notamment grâce à la mise en place d'une application en ligne.

Une conclusion générale résumera l'ensemble des travaux et les principaux résultats obtenus. Les perspectives envisagées en termes d'amélioration seront exposées.

Chapitre 1 : Données bibliographiques – Les méthodes *in silico* pour l'optimisation des propriétés ADME-Tox

1. Conception de médicaments

1.1. De l'espace chimique vers une conception rationnelle de nouveaux médicaments

L'espace chimique théorique, comprenant environ 10^{60} petites molécules organiques, est à l'image de notre espace interstellaire, c'est-à-dire tellement vaste que seulement une infime partie a été explorée jusqu'à présent [1]. A titre d'exemple, seules 140 millions de substances ont été répertoriées par l'*American Chemical Society* depuis l'année 1828, date considérée comme décrivant la première synthèse organique effectuée par Frédéric Wöhler [2]. Étant donné que la taille importante de l'espace chimique rend impossible son exploration systématique, une question cruciale est de savoir comment orienter au mieux nos efforts vers des régions susceptibles de contenir des molécules ayant une activité biologique recherchée.

L'Homme a utilisé les molécules que la Nature lui a transmises afin de soulager ses maux et a su s'en inspirer pour créer des médicaments à l'ère de la chimie organique moderne (XIX^{ème}). L'ensemble des molécules issues de la Nature constitue l'espace chimique de produits naturels. Cet espace, estimé à plus de 150 000 molécules [3], contient des produits naturels issus du vivant et donc potentiellement actifs biologiquement. Par exemple, l'acide salicylique, initialement découvert dans les feuilles de saule, est utilisé depuis 400 ans avant J.C. pour soulager douleurs et fièvres [4]. Plus récemment, le taxol, isolé à partir de l'if du Pacifique, est à l'origine de molécules utilisées pour combattre le cancer dans le cadre de la chimiothérapie. Malgré les propriétés anticancéreuses prometteuses de cette molécule, son utilisation à grande échelle nécessitait d'abattre de nombreux arbres en raison de la faible teneur en taxol présente dans l'if [5]. La lente croissance de cet arbre posait alors un problème d'approvisionnement. Pour répondre à cette problématique, l'équipe du Pr. Pierre Potier proposa une synthèse biomimétique du taxol évitant ainsi des déforestations massives.

L'industrie pharmaceutique ou les laboratoires académiques cherchant à développer de façon rationnelle des produits thérapeutiques peuvent également être inspirés de la Nature. Pour cela, les différents acteurs qui travaillent à la conception de nouveaux médicaments utilisent un processus de recherche et de développement similaire.

1.2. Processus de découverte d'un médicament

La conception de nouveaux médicaments (*drug design*) répond au besoin de produits thérapeutiques pour le traitement des maladies existantes. La recherche et le développement de médicaments se déroulent en trois temps : i) la recherche exploratoire va permettre de découvrir et optimiser des molécules actives sur une cible biologique ; ii) les molécules actives les plus prometteuses vont être testées chez l'animal puis l'Homme lors des essais précliniques et cliniques ; iii) si la ou les molécules actives sont considérées comme efficaces vis-à-vis de la pathologie ciblée et sans danger pour l'Homme, elles vont être acceptées par les autorités compétentes, comme par exemple l'agence américaine du médicament (FDA), et obtenir leur autorisation de mise sur le marché (AMM). Le processus de recherche exploratoire utilisé par les institutions privées comme publiques comporte généralement quatre grandes étapes que nous détaillons et discutons ci-après (Figure 1).

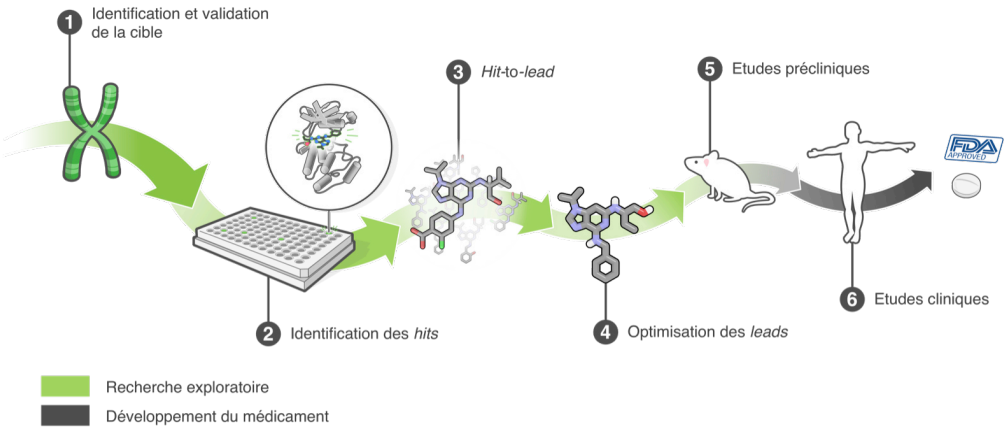


Figure 1 : Schéma simplifié du processus de recherche et de développement d'un nouveau médicament.

1.2.1. Identification et validation d'une cible thérapeutique

Un projet débute par l'étape d'identification de l'origine biologique d'une maladie et des cibles thérapeutiques potentielles pour la combattre. Une cible est une entité biochimique (gène, protéine, ADN, ARN) qui peut être associée à la pathologie visée et qui doit pouvoir être modulée lorsqu'elle interagit avec une molécule [6]. Cette étape d'identification et de validation d'une cible thérapeutique a été grandement facilitée ces dernières décennies grâce à l'avancée des technologies en biologie moléculaire, en analyse de données *omiques* et *in silico* [7–10].

1.2.2. Identification des hits

Une fois qu'une ou plusieurs cibles prometteuses ont été sélectionnées, l'étape suivante consiste à identifier des *hits* (ou touches). Un *hit* peut être défini comme une molécule capable d'interagir avec une cible pour produire l'effet biologique désiré (activation ou inhibition, effet agoniste, antagoniste ou modulateur, etc.). L'identification des *hits* implique le criblage de larges chimiothèques (librairies chimiques de plusieurs millions de molécules) afin d'explorer un sous-espace chimique représentatif. Le développement et la maintenance d'une chimiothèque est coûteux et malgré l'ampleur de la collection à disposition, seule une infime partie de l'espace chimique théorique y est représentée (Ch1 1.1). En réponse à ces défis, de nouvelles stratégies ont été développées durant les deux dernières décennies, afin de créer des chimiothèques de taille modérée, mais disposant tout de même d'une grande diversité chimique. Ainsi, les criblages basés sur les fragments moléculaires [11] ou encore l'utilisation de la chimie combinatoire [12] permettent de créer des chimiothèques qui nécessitent moins de ressources financières et qui disposent d'une plus grande diversité chimique.

De nombreuses stratégies de criblage ont émergé afin d'améliorer l'efficacité et la rapidité des tests effectués lors de cette étape d'identification des *hits*. Les avancées en termes d'automatisation et de robotique, de miniaturisation à l'aide de la microfluidique, le développement de nouvelles sondes fluorescentes ou encore les approches *in silico* de criblage virtuel à haut débit (HTVS) ont permis d'améliorer l'efficacité des tests, de réduire les quantités de produit et le temps nécessaires au criblage de vastes chimiothèques [13]. Les criblages expérimentaux à haut débit (HTS), dont le principe est de tester l'activité d'une chimiothèque sur une cible thérapeutique, sont de plus en plus souvent complétés par l'utilisation de criblage à haut contenu (HCS) pour étudier l'influence des composés sur la physiopathologie cellulaire [14], ou d'un second criblage biochimique orthogonal pour éviter toute dépendance dû au premier criblage HTS.

Pour l'identification des *hits*, un criblage à haut débit comporte généralement trois étapes. La première étape correspond au criblage primaire dont le rôle est d'identifier les composés potentiellement actifs vis-à-vis de la cible. Il est à noter que pour le criblage d'une chimiothèque de 10^6 molécules, moins de 1% d'entre elles seront considérées comme actives [15]. Par ailleurs, il n'est pas rare que des molécules identifiées comme actives soit en réalité inactives biologiquement. Ce sont des faux positifs pouvant être induits par des interférences expérimentales [16]. Par conséquent, la présence de faux positifs peut s'avérer problématique notamment lorsqu'ils sont plus nombreux que les

composés réellement actifs. Une deuxième étape est nécessaire pour trier ces faux positifs. Pour cela, un second criblage est réalisé sur un homologue de la cible visée [17]. Un faux positif est identifié lorsque l'activité observée est la même durant les deux expériences. Au final, environ 1000 *hits* possèdent une activité confirmée à l'issue de ce deuxième criblage. La troisième étape consiste à réaliser une étude plus approfondie à l'aide de courbes dose-réponse dans le but de comparer l'activité de chaque *hit* [18].

1.2.3. Hit-to-lead

Les *hits* sélectionnés sont ensuite regroupés en fonction de leur relation structure-activité (RSA) afin de définir des séries chimiques. La RSA consiste à déterminer les tendances existantes entre des structures chimiques similaires et leurs activités biologiques. Ainsi, les molécules de chaque groupe possèdent une sous-structure commune responsable de l'activité biologique principale (pharmacophore). Les variations structurales au niveau des groupes fonctionnels sont responsables des différences d'activité observées dans la série congénérique. A ce stade, des tests *in vitro* et/ou *in vivo* d'Absorption, de Distribution, de Métabolisme et d'Elimination (ADME) vont être effectués dans le but de définir le profil médicamenteux (*drug-like*) des molécules. L'objectif est d'identifier les composés ayant des propriétés thérapeutiques et des propriétés ADME semblables à celles d'un médicament.

Le concept de profil *drug-like* a été introduit par Lipinski en 2001 et est issu de ses observations faites à partir des médicaments administrés par voie orale. Son analyse descriptive a montré que la plupart des médicaments étaient susceptibles de résider dans des zones de l'espace chimique définies par une gamme limitée de propriétés moléculaires [19]. Cette étude a révélé que 90% des médicaments administrés par voie orale, ayant passés les essais cliniques de phase II, disposaient de moins de 5 donneurs de liaisons hydrogène, moins de 10 accepteurs de liaisons hydrogène, d'une masse moléculaire inférieure à 500 Da et pour finir d'une lipophilie (LogP) inférieure à 5. L'ensemble de ces observations constitue la règle des cinq (*Ro5*). Par extension, d'autres règles basées sur le même principe ont été proposées, comme par exemple les règles de Veber qui reflètent la biodisponibilité orale [20]. De nos jours, ces règles sont souvent mal interprétées et sont largement utilisées pour filtrer les molécules. Par conséquent, elles peuvent être trop restrictives, réduisant l'espace chimique exploré lors de certains projets et difficilement applicables aux nouvelles molécules approuvées par la FDA [21,22].

Les propriétés ADME traduisent l'effet de l'organisme sur la molécule et décrivent plusieurs phénomènes représentant la pharmacocinétique des produits (l'absorption du médicament dans l'organisme, la distribution dans les tissus et la circulation sanguine, le métabolisme par des protéines comme par exemple les cytochromes P450 (CYP450) et enfin l'élimination par les urines, la bile ou les selles du principe actif, soit sous une forme inchangée, soit sous la forme de métabolites [23]). A cause du faible débit des tests expérimentaux et de leur coût important, toutes les molécules nouvellement synthétisées ne peuvent pas être testées sur l'ensemble des criblages ADME disponibles. L'évaluation de ces propriétés, pourtant importantes lors de la conception d'un médicament, se produit de manière irrégulière et intervient trop tardivement dans le processus de recherche et développement. De nos jours, l'anticipation des profils ADME-Tox défavorables est une solution souvent entreprise pour réduire les échecs tardifs [24].

A la fin de cette étape de sélection des *hits*, les molécules les plus prometteuses de chaque groupe vont constituer des têtes de séries (*leads*).

1.2.4. Optimisation des leads

L'étape finale de la recherche d'un médicament a pour but d'obtenir des candidats ayant des propriétés favorables pour une transposition chez l'animal puis chez l'Homme. Pour cela, l'affinité vis-à-vis de la cible thérapeutique, les propriétés pharmacodynamiques (effets de la molécule sur l'organisme) ou encore les propriétés pharmacocinétiques (effets de l'organisme sur la molécule) des *leads* vont être optimisées. Ainsi, le chimiste médicinal va proposer des analogues structuraux à l'aide d'approches rationnelles de relation structure-activité. Durant cette étape d'optimisation, l'affinité et la sélectivité des composés chimiques pour la cible, leurs propriétés pharmaceutiques (stabilité et solubilité), leur profil médicamenteux, ainsi que leur propriétés ADME et leur toxicité (hERG, cancérogénicité, génotoxicité, reprotoxicité, etc.), aussi nommées ADME-Tox, vont être évaluées et optimisées en parallèle. Par conséquent, le chimiste fait régulièrement face à des problèmes complexes d'optimisation multidimensionnelle dont l'importance des différents paramètres change au cours du projet. Dans la plupart des cas, les molécules apportant le meilleur compromis entre l'affinité et les propriétés ADME-Tox seront sélectionnées en tant que candidats médicaments pour poursuivre les essais réglementaires précliniques (chez l'animal) puis cliniques (chez l'homme).

1.2.5. Méthodes in silico pour la recherche thérapeutique

La conception de médicaments assistée par ordinateur (CADD) utilise une recherche beaucoup plus ciblée que les criblages traditionnels à haut débit [25]. La CADD vise non seulement à identifier et à expliquer les raisons moléculaires de l'activité thérapeutique, mais aussi à prédire les modifications chimiques possibles permettant d'améliorer l'activité. Elle permet également de réduire le temps et les coûts nécessaires à l'identification de molécules prometteuses. Cette technologie est habituellement employée pour : i) filtrer de vastes chimiothèques en de plus petits ensembles de molécules prédites actives afin de les tester expérimentalement [26]; ii) guider l'optimisation des *leads*, aussi bien pour augmenter leur activité que pour optimiser leurs propriétés ADME-Tox ; iii) concevoir de nouveaux composés en testant virtuellement une plus grande diversité de chémotypes. Pour parvenir à ces fins, différentes approches ont été élaborées et peuvent être classées selon deux catégories, à savoir : les approches *structure-based* et les approches *ligand-based*.

Les approches *structure-based* s'appuient sur la connaissance de la structure de la cible visée pour calculer les énergies d'interaction entre la protéine cible et les composés testés. Ces approches nécessitent d'avoir identifié la cible biologique impliquée dans la pathologie et de disposer de données structurelles de la cible. Idéalement, la structure tridimensionnelle disposant d'une haute résolution est déterminée à l'aide de la cristallographie aux rayons X ou encore de la RMN. Cependant, il arrive parfois que la structure de la cible visée ne soit pas encore élucidée. Il est alors possible de construire la structure 3D de la macromolécule étudiée à l'aide des approches par homologie. Cette structure tridimensionnelle peut être employée lors d'un amarrage moléculaire (docking). Cette méthode consiste à positionner chaque composé dans le site de liaison de la cible afin d'estimer leur affinité. L'estimation de l'affinité se fait à l'aide d'une fonction de score prenant principalement en compte les interactions hydrophobes, les interactions polaires et les interactions de van der Waals. Ainsi, plus les interactions sont favorables, plus les composés se lient étroitement à la cible. Cette méthode d'amarrage moléculaire permet d'élucider le mode d'interaction entre le site de liaison et les composés étudiés, mais également de prioriser les molécules les plus prometteuses en fonction du score d'amarrage obtenu. Bien que cette méthode apporte des solutions concrètes pour identifier rapidement les composés ayant une affinité vis-à-vis de la cible, elle possède cependant certaines limitations. Généralement, plusieurs caractéristiques de la cible ne sont pas prises en compte lors d'un amarrage moléculaire, comme par exemple l'effet

entropique lors de l'interaction du ligand, le pH local et la teneur en sel du site de liaison. De ce fait, l'état d'ionisation ou la tautomérie des composés dans le site de liaison peuvent être différents entre la réalité et la simulation, ce qui peut fausser les observations faites à partir d'un amarrage moléculaire. De plus, les aspects cinétiques du système (association, fixation et dissociation) ne sont pas explorés dans le cadre de cette méthode. La cinétique du système influence le temps de résidence d'un composé dans le site actif. Même si l'affinité d'un ligand pour la cible est importante, ce dernier sera peu efficace s'il n'arrive pas à accéder au site de liaison ou s'il dispose d'un faible temps de résidence [27]. Afin de répondre aux problématiques posées par la cinétique du système, plusieurs approches basées sur la dynamique moléculaire ont été développées lors de cette dernière décennie [28]. Il est à noter que cette thématique est activement étudiée au sein de l'ICOA [29]. Les succès dans la recherche de nouveaux médicaments à l'aide de ces approches, notamment grâce à l'amélioration de la puissance des calculs, en ont fait des méthodes très utilisée de nos jours [30–32].

Les approches *ligand-based* sont généralement préférées lorsqu'il n'y a pas ou peu de données sur la structure de la cible. Ces approches utilisent les recherches existantes sur les ligands connus comme étant actifs ou inactifs par rapport à la cible étudiée [25,33]. Elles font l'hypothèse que des structures similaires du point de vue moléculaire peuvent exercer une activité biologique équivalente. Ainsi, des structures chimiques d'interactions moléculaires comparables vis-à-vis de la cible vont avoir des activités similaires. Cependant, deux molécules d'activité équivalente peuvent avoir des structures chimiques totalement différentes. Parmi ces approches, les études quantitatives de relation structure-activité (QSAR), qui peuvent être étendues à des approches qualitatives ((Q)SAR), sont utilisées pour créer des modèles statistiques capables de prédire l'activité des molécules pour une cible biologique. Ces approches peuvent également être employées afin de prédire les propriétés physico-chimiques ou physiologiques liées à la structure des composés. On parlera alors d'études quantitatives ou qualitatives de relation structure-propriété ((Q)SPR). Il existe également une dénomination similaire pour les méthodes de prédiction de la toxicité ((Q)STR). Pour plus de clarté, nous avons choisi de n'utiliser que le terme (Q)SAR dans la suite de ce manuscrit.

Le processus de conception d'un médicament est une entreprise colossale qui comporte des risques importants d'échecs, qui se traduit par une probabilité de succès inférieure à 9,6 % [34]. Ainsi, la recherche et le développement de médicaments nécessitent beaucoup de temps et des investissements financiers considérables. En règle générale,

15 années peuvent s'écouler de la création du projet jusqu'à la commercialisation d'un nouveau médicament, pour un budget moyen d'environ 1 milliard de dollars (environ 800 millions d'euros) [35].

1.3. Enjeux actuels de la recherche thérapeutique

Les études précliniques et cliniques nécessaires à l'obtention de l'AMM sont de plus en plus longues et coûteuses, afin de prouver l'efficacité et l'innocuité absolues du médicament. Les tests pharmacologiques et ADME-Tox sont introduis de plus en plus tôt dans les étapes de conception dans le but d'accélérer les phases de sélection de candidats prometteurs et de limiter les risques d'échec tardif. Cependant, le faible débit et le coût important des tests ADME-Tox représentent un frein pour le criblage de vaste chimiothèques. De ce fait, l'introduction précoce des approches *in silico* pour la prédiction des propriétés ADME-Tox présentent un intérêt grandissant. En effet, elles donnent la possibilité de prédire les propriétés de plusieurs milliers de composés avant les étapes de synthèse organique, afin d'identifier les structures qui ont le plus de chance d'aboutir. Il est important de noter que ces approches n'ont pas pour but de remplacer les tests expérimentaux. Elles permettent d'améliorer la prise de décision pour prioriser les prochaines étapes de conception.

La mise en place de cette stratégie a poussé des entreprises pharmaceutiques, comme AstraZeneca, à reconsidérer leur processus de recherche de nouveaux médicaments [36]. L'analyse de leurs données internes, sur la période de 2005 à 2010, a révélé que la toxicité était la cause majeure de l'échec des candidats médicaments lors des essais précliniques et des essais cliniques de phase I. La même étude a été proposée sur la période de 2012 à 2016 et montre que les mesures prises par AstraZeneca ont permis de réduire l'impact de la toxicité de 32 % et 24,5 % pour les essais précliniques et les essais cliniques de phase I, respectivement [37]. Ainsi, leur taux de succès a connu une nette amélioration en passant de 4 % en 2005 à 19 % en 2016. Cette amélioration notable a été possible, en partie, grâce à l'incorporation de plateformes de prédiction ADME-Tox robustes dès l'étape *hit-to-lead*. Cette étude prouve que des changements sont nécessaires afin de s'adapter aux nouveaux enjeux de la recherche thérapeutique et augmenter les avancées scientifiques [38].

La détermination « en temps réel » des propriétés ADME-Tox représente à l'heure actuelle un des enjeux majeurs pour accélérer et optimiser la recherche de nouveaux médicaments. Ce besoin a été reconnu dans le cadre de l'Initiative pour les Médicaments

Innovants (IMI), une initiative conjointe entre l'Union européenne et l'association des entreprises pharmaceutiques (EFPIA), qui finance actuellement une cinquantaine de projets visant à améliorer le processus de découverte de médicaments. Ainsi de nombreux projets, comme par exemple BIGCHEM [39], ToxCast EPA [40] ou encore eTOX [41], ont vu le jour afin de répondre à cette demande à partir des années 2010. Le projet eTOX représente les efforts combinés de 25 participants industriels (Pfizer, Novartis, Bayer, Servier, Lhasa, etc.) et de laboratoires publics dans le but d'accélérer le processus d'introduction de médicaments plus sûrs et plus efficaces, en développant des stratégies et des outils *in silico* pour prédire l'innocuité et les effets secondaires des candidats médicaments. Ce projet est né d'une prise de conscience au sujet des données détenues par l'industrie pharmaceutique. En effet, l'industrie pharmaceutique est en possession d'un grand nombre de données non publiées qui ont été acquises au cours du processus de développement de médicaments. Ces données n'ont pas été rendues publiques, car elles sont soumises à la propriété intellectuelle. Les principaux objectifs du projet eTOX sont : i) de créer une vaste base de données qui contiendrait à la fois des données publiques et privées ; ii) de mettre en œuvre des méthodes qui permettraient d'accéder à ces informations exclusives, tout en protégeant la propriété intellectuelle ; et iii) d'utiliser cette vaste base de données pour développer ou améliorer les modèles de prévision de toxicité [42].

Le sujet de thèse que nous présentons est née d'une collaboration entre les Technologie Servier et l'Institut de Chimie Organique et Analytique (ICOA) et s'inscrit dans cette problématique. Elle a pour but de proposer de nouvelles approches *in silico* visant à améliorer la robustesse des modèles de prédiction ADME-Tox. L'objectif pour Technologie Servier est de valoriser les données pharmacocinétiques acquises depuis plusieurs années afin d'optimiser et/ou d'acquérir des modèles ADME-Tox. L'objectif pour l'ICOA est de se doter d'une plateforme *in silico* de criblage ADME-Tox basée sur des données publiques. L'objectif commun de ces deux entités, privée et publique, est d'anticiper les profils ADME-Tox à l'aide de méthodes *in silico*, afin de focaliser le travail des chimistes sur les structures les plus prometteuses durant leurs travaux de recherche.

2. Principes d'ADME-Tox

Entre le moment où il est administré et le moment où il atteint sa cible thérapeutique, le principe actif rencontre plusieurs barrières ADME dans l'organisme. Chaque passage d'une barrière ADME est susceptible d'engendrer une diminution de la concentration en principe actif. Or, la concentration en xénobiotique (principe actif) est déterminante pour obtenir un temps d'exposition assez long avec la cible, afin d'engendrer un effet thérapeutique efficace. Ainsi, une optimisation des *leads* uniquement orientée sur l'activité biologique peut donner des composés qui sont très efficaces comme ligands pour la cible visée, mais dont les propriétés ADME-Tox peuvent être inadéquates, les empêchant de devenir des médicaments à succès [43]. De ce fait, l'optimisation des *leads* doit prendre en compte l'amélioration de l'affinité pour la cible thérapeutique, mais également l'optimisation des propriétés ADME-Tox indispensables pour le passage des barrières physiologiques (Ch1 1.2.4) et l'innocuité du médicament. L'optimisation de ces propriétés ADME-Tox peut être réalisée à l'aide des approches de relation structure-propriété (RSP) basées sur le même principe que les approches de RSA [44]. Les propriétés ADME-Tox sont des propriétés physiologiques gouvernées par des principes complexes dépendants de la structure chimique du principe actif, mais également de l'environnement biologique. L'ensemble de ces propriétés physiologiques permet de définir le profil ADME-Tox d'une molécule. Ce profil traduit la capacité d'un principe actif à traverser chaque barrière ADME de son site d'administration, en passant par les interactions avec les transporteurs, les enzymes du métabolisme (phase I et phase II) ainsi qu'avec les protéines plasmatiques, jusqu'à sa rencontre avec la cible thérapeutique. Le voyage du médicament dans l'organisme débute par son administration. Parmi les différentes voies d'administration, la voie orale est la plus utilisée en raison de sa commodité d'utilisation pour les patients (80% des médicaments commercialisés). A ce titre, la présentation des barrières ADME prendra en considération le cas d'une administration par voie orale. Par ailleurs, les phénomènes qui interviennent dans l'organisme sont dynamiques et interactifs. Nous avons essayé de présenter ci-après les phénomènes physiologiques permettant d'expliquer les principes d'ADME-Tox tels qu'ils surviennent dans l'organisme.

2.1. Absorption

Une fois le médicament ingéré, son voyage dans l'organisme débute par la phase d'absorption. L'absorption correspond à la pénétration du médicament dans l'organisme. Après administration orale, la phase d'absorption (A) regroupe les phénomènes impliqués dans le transfert d'un xénobiotique du tractus gastro-intestinal (GI), et plus

particulièrement de l'intestin grêle, vers la circulation sanguine. Après la désagrégation de la forme galénique et la dissolution du principe actif dans les sucs digestifs, le xénobiotique en solution va devoir traverser les cellules épithéliales, et plus particulièrement, une ou plusieurs membranes cellulaires via des mécanismes de transport paracellulaire (entre les cellules) ou transcellulaire (à travers les cellules) (Figure 2.a). Les mécanismes paracellulaires sont généralement observés pour les molécules de petites tailles pouvant être hydrophiles ou polaires. Les membranes plasmiques, constituées d'une bicouche phospholipidique, représentent une barrière presque infranchissable pour les molécules non lipophiles. Les mécanismes transcellulaires permettent de laisser passer une plus grande diversité de molécules. Ces mécanismes peuvent être passifs (diffusion [45]) pour les molécules lipophiles, peuvent faire intervenir des transporteurs [46] (transport facilité ou actif) et dans certains cas des vésicules (transcytose [47]) pour les molécules hydrophiles ou de poids moléculaire plus élevé. Ainsi, la solubilité, la lipophilie ou encore le pKa sont les principales propriétés physico-chimiques qui influencent l'absorption d'une molécule dans l'organisme.

La grande majorité des médicaments possèdent des groupes fonctionnels ionisables qui peuvent prendre la forme d'un acide, d'une base ou d'un zwitterion selon le principe de Brønsted–Lowry [48,49]. Le pKa d'une molécule traduit le degré de dissociation, c'est à dire le pH pour lequel la moitié de la substance est sous forme protonée ou non protonée. Ce paramètre va impacter l'état d'ionisation d'une molécule à un pH donné et par conséquent joue un rôle prépondérant sur la solubilité et la lipophilie d'un composé. En effet, une molécule ionisée est plus soluble dans les milieux aqueux qu'une molécule neutre, car elle est plus polaire. De plus, une molécule ionisée aura moins de chance de traverser les membranes cellulaires qu'une molécule neutre. Une molécule neutre est plus lipophile, facilitant son passage à travers les membranes cellulaires par le biais de la diffusion passive. Lors d'une administration par voie orale, la solubilité et la lipophilie du principe actif vont être fonction du pKa du xénobiotique et des variations de pH observées dans le tractus digestif. Ainsi, un principe actif disposant de groupements basiques sera sous la forme cationique (protonée) lors de son passage au niveau de l'estomac et du jéjunum (pH ~ 1-3). Ceci améliorera la solubilité du principe actif dans la lumière intestinale, mais défavorisera son absorption. Les molécules acides sont sous la forme neutre lors de leur entrée dans l'intestin, ce qui facilite leur absorption par le biais de la diffusion passive. Le pH augmente tout au long de l'intestin et les molécules acides vont progressivement acquérir leur forme anionique (non protonée) facilitant la solubilité du

principe actif et réduisant alors son absorption. Le logarithme de la solubilité aqueuse (LogS), le logarithme du coefficient de partition entre les phases H_2O/Octanol (LogP), ou encore le LogP à un pH spécifique (LogD) sont des grandeurs généralement utilisées dans le but d'estimer la capacité du composé à être absorbé par l'organisme par le biais d'une diffusion passive. Généralement, le xénobiotique doit avoir un LogP modéré (compris entre 0 et 3), pour avoir une absorption intestinale optimale. D'autre part, le LogS est comprise entre 0 et -4 pour la majorité des médicaments administrés par voie orale. La lipophilie au pH 7,4 (LogD7,4) doit être comprise entre 1 et 3 pour une absorption intestinale favorable du principe actif. Cependant, ces propriétés physico-chimiques ne prennent pas en compte les phénomènes actifs. Ainsi, une sélection orientée uniquement à l'aide de ces critères ne garantie pas l'absorption des composés dans l'organisme.

Même si une molécule réunit toutes les caractéristiques favorables à son absorption, la totalité de la dose (quantité de principe actif administrée) ne sera pas retrouvée dans la circulation sanguine. En effet, le principe actif peut être confronté à plusieurs obstacles de taille. Le premier de ces obstacles est l'hydrolyse enzymatique [50,51]. Le tractus digestif, dont la fonction première est de digérer et d'absorber les nutriments contenus dans les aliments, possède une multitude d'enzymes « digestives » sécrétées par l'estomac, le pancréas ou la salive. Ces enzymes (peptidases, estérases, ribonucléases, phosphatases, etc.) tapissent les muqueuses de l'intestin grêle en grande quantité. Par conséquent, un principe actif est susceptible d'être hydrolysé par ces enzymes avant même de pouvoir essayer de traverser la membrane plasmique des cellules épithéliales. Les composés contenants des groupements ester, amide ou encore carbamate sont particulièrement sensibles à cette hydrolyse enzymatique. Cependant, des stratégies ont été élaborées dans le but de tirer parti de ce phénomène de dégradation, comme dans le cas des promédicaments [52]. Un promédicament (*prodrug*) utilise ces réactions enzymatiques afin de produire *in situ* la forme biologiquement active du médicament. Ainsi, dans le cas d'un principe actif peu soluble il est possible de lui ajouter une sous-structure (phosphate, ester, etc.) permettant d'augmenter sa solubilité dans l'intestin. Ce promédicament, rendu soluble, va ainsi pouvoir atteindre la surface des cellules épithéliales. Les enzymes vont alors l'hydrolyser et libérer la forme active du composé à proximité du site d'absorption ciblé.

Les molécules qui ont réussi à traverser la membrane des cellules épithéliales peuvent être confrontées au mécanisme d'efflux et d'influx. L'efflux est un mécanisme par lequel les cellules vont pouvoir rejeter certaines molécules dans le milieu extracellulaire par le

biais de transporteurs. La famille des transporteurs ABC, dont la glycoprotéine P (P-gp) fait partie, contribue à l'efflux des xénobiotiques du milieu intracellulaire vers le milieu extracellulaire [53]. Un principe actif va être soumis à ce mécanisme d'efflux s'il est substrat du transporteur incriminé. D'autres transporteurs, tels que les protéines résistantes aux médicaments (MRP2) ou les transporteurs BCRP participent également de façon plus modérée à ce phénomène d'efflux dans l'intestin [54]. Ce phénomène contribue donc à l'élimination de certains xénobiotiques dans les selles.

Si le principe actif présent dans le cytoplasme des cellules épithéliales est peu enclin au phénomène d'efflux, il peut néanmoins faire face à l'obstacle du métabolisme dès la phase d'absorption. En effet, le cytochrome P450 3A4 (CYP3A4) est une enzyme du métabolisme présente de façon extra-hépatiques également dans les cellules intestinales [55]. Ce cytochrome métabolise divers composés chimiques [56] et peut travailler conjointement avec les transporteurs P-gp [44,57]. Ainsi, les glycoprotéines P et transporteurs réduisent la concentration intracellulaire du principe actif et de ses métabolites dans les cellules intestinales, ce qui permet au CYP3A4 de catalyser l'oxydation du xénobiotique en réduisant le risque de saturation. Ce phénomène du métabolisme, présent dès la phase d'absorption, est souvent nommé « premier passage intestinal ». Il est étroitement lié au premier passage hépatique. En effet, une fois que le principe actif a traversé les cellules épithéliales et a gagné les capillaires sanguins de la muqueuse intestinale, la fraction absorbée va ensuite être dirigée vers le foie par le biais de la veine porte. Le foie, considéré comme l'organe de détoxification, va être le siège de plusieurs réactions métaboliques par les cytochromes P450 (CYP450). Ces acteurs de réactions biochimiques seront présentés plus en détails dans la partie consacrée au métabolisme (Ch1 2.2). Le premier passage hépatique va réduire la concentration biodisponible en principe actif dans la circulation sanguine.

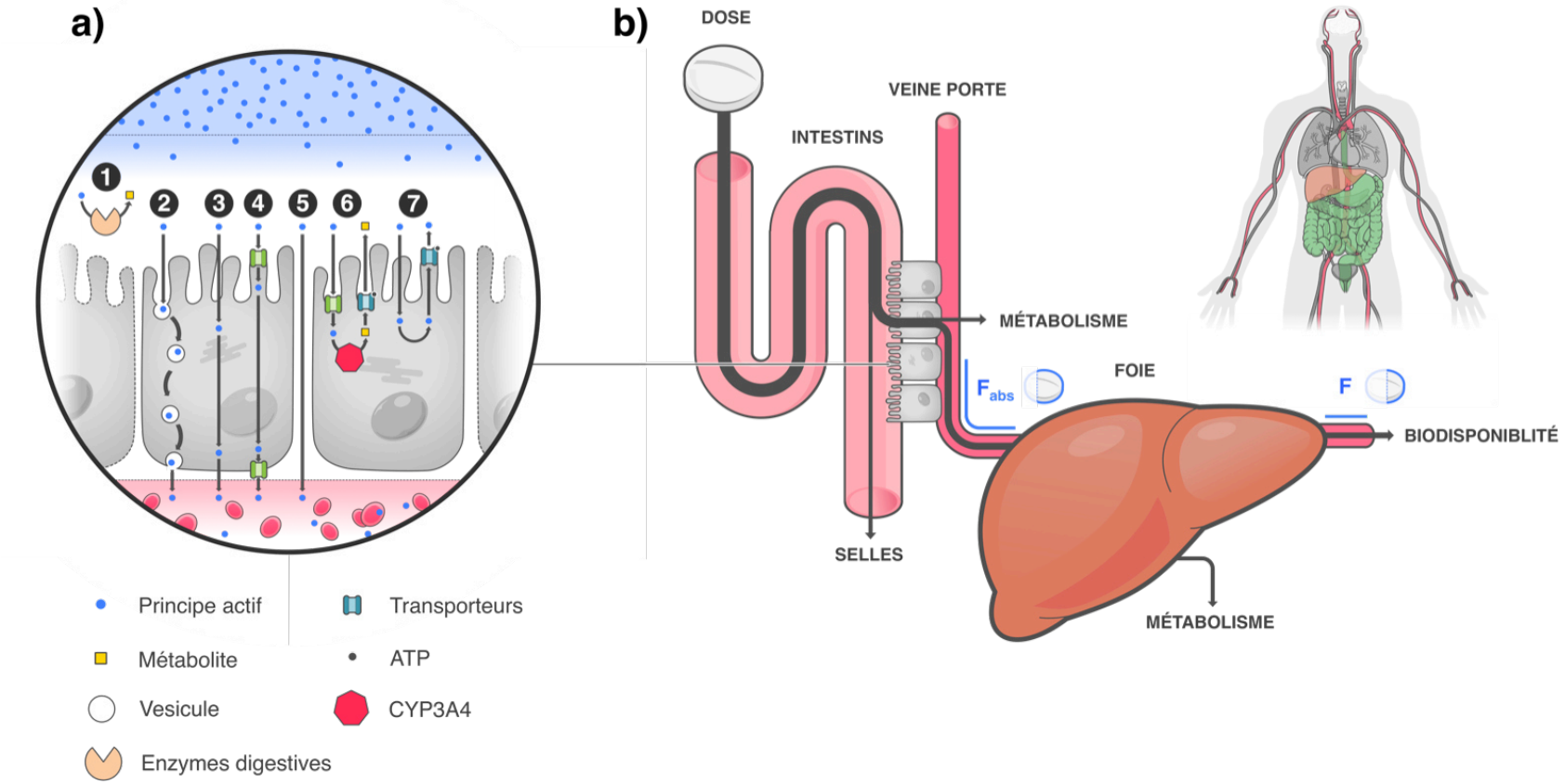


Figure 2 : Vue d'ensemble des phénomènes qui interviennent lors de la phase d'absorption.

a) Illustration des mécanismes enzymatiques et des mécanismes de transport mis en œuvre lors du passage des cellules épithéliales : 1) Hydrolyse enzymatique ; 2) Transcytose ; 3) Diffusion passive ; 4) Transport actif via un transporteur (mécanisme d'influx) ; 5) Mécanisme paracellulaire ; 6) Métabolisme via le CYP3A4 ; 7) Mécanisme d'efflux via les transporteurs P-gp. b) Schéma global de la phase d'absorption suite à une administration par voie orale.

Plusieurs propriétés permettent d'estimer l'ensemble des phénomènes impliqués lors de la phase d'absorption. La première propriété est la fraction absorbée (F_{abs}) qui reflète le pourcentage de la dose qui a réussi à traverser les cellules épithéliales pour gagner la circulation sanguine (Figure 2.b) [58]. Cette propriété à elle seule ne permet pas cependant de déterminer la disponibilité du principe actif dans la circulation systémique. Ceci est représenté par la biodisponibilité (F), qui reflète la fraction biodisponible dans la circulation sanguine après les phases d'absorption et le premier passage hépatique (Figure 2.b) [59]. Cette fraction biodisponible va pouvoir ensuite être distribuée dans l'organisme.

2.2. Distribution

La phase de distribution (D) regroupe les phénomènes impliqués dans le transport via la circulation sanguine et le partage du xénobiotique entre le sang et les différents tissus de l'organisme. Le transport du principe actif va être assuré par le sang, qui est constitué d'environ 44% de globules rouges, 1% de globules blancs et de 55% de plasma. Le plasma contient approximativement 8% de protéines plasmatiques, qui jouent un rôle central dans le maintien du pH sanguin, la régulation de la pression osmotique, ainsi que la fixation et le transport des composés endogènes nécessaires au fonctionnement de l'organisme. Les protéines plasmatiques impliquées dans ce phénomène de fixation sont l'albumine (HSA), les α-1-glycoprotéines acides (AGP), les lipoprotéines et les globulines [60]. L'albumine représente 60% des protéines plasmatiques totales et fixe préférentiellement les composés acides ou neutres. Les α-1-glycoprotéines acides représentent environ 3% des protéines plasmatiques totales et fixent préférentiellement les composés basiques ou neutres. Les autres protéines plasmatiques, telles que les lipoprotéines (transporteurs endogènes du cholestérol VHDL, HDL et LDL) ou les globulines, interviennent également lors de la fixation des composés exogènes.

Le principe actif présent dans la circulation sanguine se fixe de façon réversible avec les protéines plasmatiques (Figure 3.a). L'affinité du xénobiotique pour ces protéines va déterminer le rapport entre la fraction libre en solution et la fraction liée aux protéines. Seule la fraction libre est biodisponible pour induire l'effet thérapeutique désiré, être métabolisée puis éliminée. Les molécules libres peuvent être soumises à plusieurs dégradations induites par les enzymes plasmatiques, comme par exemple les aldolases, les lipases, ou les phosphatases, qui vont pour la plupart hydrolyser le principe actif [61]. Telle une réserve, la fraction liée est libérée progressivement afin de maintenir un équilibre avec la fraction libre lors de la phase de distribution. Le phénomène de fixation

sera exprimé expérimentalement par le pourcentage de la fraction liée aux protéines plasmatiques (PPB) ou par le pourcentage de la fraction libre dans le plasma (fu_p).

Le principe actif véhiculé par le flux sanguin se retrouve à présent distribué dans l'organisme et à proximité des tissus. Seule la fraction libre du xénobiotique peut quitter la circulation systémique en traversant les membranes des capillaires sanguins, afin de diffuser dans les tissus. Cette distribution dans les tissus va suivre différentes phases de pénétration (Figure 3.b). La distribution va se faire dans un premier temps du volume intravasculaire vers le volume interstitiel (extracellulaire). Si le principe actif libre est capable de traverser les membranes cellulaires, la distribution va se prolonger du volume interstitiel vers le volume cellulaire (intracellulaire), pouvant aboutir à une distribution presque homogène dans les tissus. Cette diffusion tissulaire doit être spécifique au mode d'action du médicament. Prenons l'exemple de l'edoxaban, un inhibiteur du facteur Xa. Le facteur Xa joue un rôle central dans le mécanisme de coagulation en activant la prothrombine en présence de phospholipides (d'origine tissulaire ou plaquettaire), entrainant ainsi la génération de thrombine et convertit le fibrinogène en fibrine. Ceci implique la formation de caillot et une agglomération des plaquettes. De ce fait l'inhibition du facteur Xa par le biais de l'edoxaban permet un contrôle de la thrombogénèse. Le mode d'action du l'edoxaban nous indique qu'une diffusion dans les volumes interstitiels et cellulaires n'est pas nécessaire, dans ce cas, pour obtenir l'effet thérapeutique désiré.

Le volume de distribution (Vd) est généralement utilisé afin de pouvoir déterminer le mode de diffusion tissulaire du principe actif. Le Vd correspond au volume de liquide nécessaire pour contenir la quantité de médicament présent dans l'organisme à la même concentration que dans le plasma. Connaissant la dose de médicament présente dans l'organisme (Dose) et la concentration plasmatique du principe actif (C_p), il est alors possible de déterminer simplement le volume de distribution selon l'Equation 1.

$$Vd = \frac{Dose}{C_p}$$

Equation 1 : Volume de distribution.

Il s'agit en fait d'un volume apparent, car l'Equation 1 présuppose que la distribution du médicament dans l'organisme est homogène. Par conséquent, un médicament qui s'accumule dans les tissus aura une concentration plasmatique relativement faible par rapport à la dose administrée. Son Vd calculé sera élevé. Les médicaments à très faible Vd sont principalement confinés au fluide intravasculaire (C_p élevée). Cela peut se

produire pour deux raisons : i) soit la molécule est trop volumineuse pour quitter la circulation sanguine, ou ii) soit la molécule se lie fortement aux protéines plasmatiques.

En réalité, cette distribution n'est pas homogène et dépend de plusieurs facteurs. En effet, la condition physique, l'âge ou le poids d'un individu ; et à l'échelle cellulaire, l'accumulation du xénobiotique libre dans certains tissus, ainsi que la composition cellulaire des endothéliums, sont autant de paramètres influençant la distribution du principe actif dans l'organisme. Concernant ce dernier facteur, le système nerveux central, aussi nommé névraxe, est doté de la barrière hémato-encéphalique. Cette barrière presque infranchissable est formée par les cellules endothéliales qui possèdent des jonctions serrées rendant impossible le transport paracellulaire. Une molécule doit obligatoirement traverser les membranes luminales (au contact du sang) et basales (au contact du tissu) de ces cellules pour être distribuée dans le névraxe. De plus, comme dans le cas de l'absorption, les glycoprotéines P et transporteurs sont présentent dans les cellules endothéliales afin d'expulser toutes les molécules étrangères vers la circulation sanguine [62]. Ainsi, ces mécanismes de protection rendent difficile la distribution du principe actif libre dans le système nerveux central (SNC). Il a été estimé qu'uniquement 2% des médicaments ayant une application sur le SNC étaient capables de traverser cette barrière entre le sang et le cerveau [63]. Afin de déterminer la capacité du principe actif libre à traverser la barrière hémato-encéphalique, aussi nommée *Blood-Brain Barrier* (BBB), la perméabilité est généralement utilisée. Au même moment dans l'organisme, le métabolisme et l'élimination du principe actif libre réduisent progressivement la concentration biodisponible.

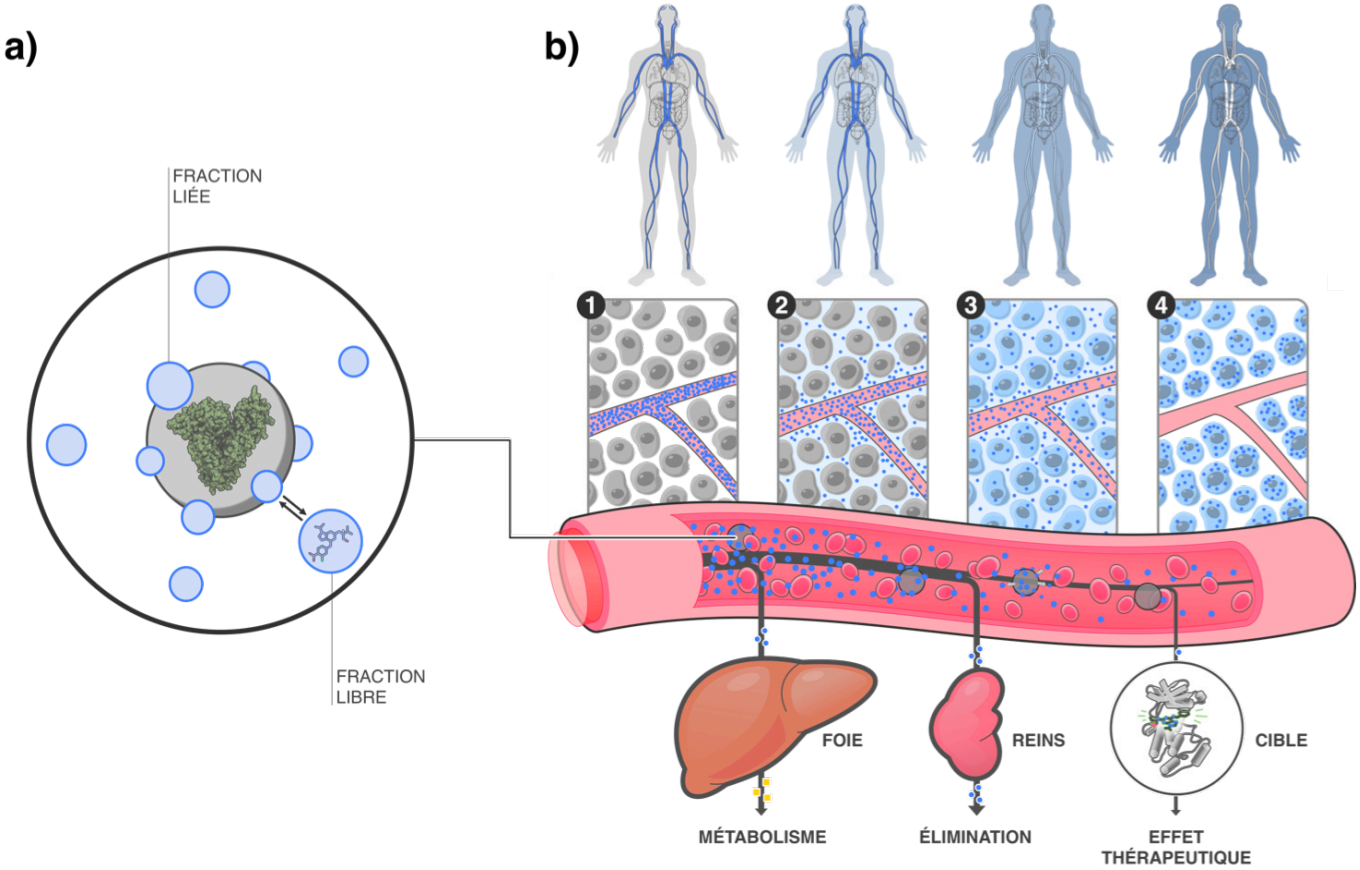


Figure 3 : Vue d'ensemble des phénomènes qui interviennent lors de la phase de distribution.

a) Représentation schématique de la fixation du principe actif (cercle bleu) aux protéines plasmatiques (cercle gris). La protéine plasmatique représentée est l'albumine (PDB : 4L9K). b) Schéma globale de la phase de distribution suite à l'absorption du principe actif. Les différentes phases de diffusion tissulaire sont représentées dans la partie supérieure: 1) Volume intravasculaire ; 2) Volume intravasculaire et interstitiel ; 3) Volume extra- et intracellulaire ; 4) Enrichissement cellulaire. Les différentes possibilités de distribution du principe actif sont représentées dans la partie inférieure.

2.3. Métabolisme

La phase du métabolisme (M) peut être décrite comme un processus de biotransformation du principe actif par lequel il va être plus polaire, et par conséquent plus facilement dissout dans les milieux aqueux, tels que la bile ou l'urine, afin d'être éliminé par l'organisme [64]. Comme nous l'avons vu précédemment, une molécule peut être confrontée à des dégradations enzymatiques dès les premières étapes d'absorption et de distribution. Ainsi, un principe actif peut être soumis à une ou plusieurs transformations biochimiques tout au long de son voyage dans l'organisme, réduisant drastiquement la concentration biodisponible avant de pouvoir impulser l'effet thérapeutique désiré.

Les réactions métaboliques responsables des décompositions successives du principe actif sont généralement divisées en deux phases. Les réactions de phase I regroupent les biotransformations qui oxydent la structure moléculaire du xénobiotique (oxydation, désalkylation, hydroxylation et désamination) [65]. Les réactions de phase II ont pour but de conjuguer un ou plusieurs groupes fonctionnels polaires pour rendre le principe actif plus soluble ou reconnu par d'autres protéines de l'organisme [66]. Majoritairement, ces réactions de phase II suivent les réactions de phase I, mais elles peuvent aussi s'appliquer directement sur le principe actif.

Plusieurs enzymes participent à ces réactions biochimiques, comme par exemple les CYP450, les monooxygénases contenant de la flavine (FMO), les UDP-Glucuronosyl-transférases (UGT), ou encore les N-acétyl-transférases (NAT) [67]. Certaines de ces enzymes sont plus impliquées que d'autres dans le métabolisme des médicaments. J. A. Williams *et al.* ont déterminé les causes principales du métabolisme pour les 200 médicaments les plus prescrits aux USA en 2002. Leur analyse est basée sur les rapports de la FDA [68]. Ainsi, cette étude a montré que les enzymes de la famille des CYP450 était la cause majoritaire du métabolisme pour environ 70% des médicaments (Figure 4).

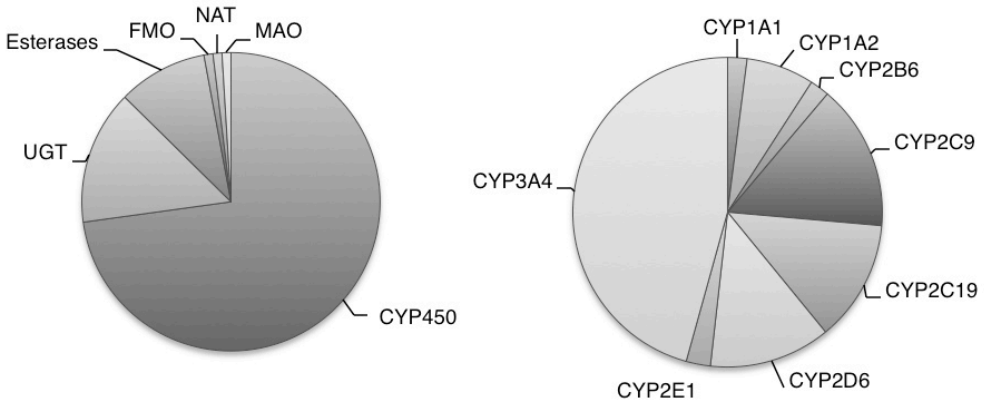


Figure 4 : Enzymes impliquées dans le métabolisme des médicaments les plus prescrits aux USA en 2002.

A gauche est représentée la contribution de chaque enzyme sur le métabolisme des médicaments étudiés. A droite est représentée la proportion de médicaments considérés comme substrats des isoenzymes de la famille des CYP450.

Les cytochromes P450 sont une famille d'hémoprotéines impliquées dans les réactions de phase I, lors des processus enzymatiques qui se déroulent dans les intestins et le foie. Le génome humain contient 57 gènes exprimant divers cytochromes [69]. Beaucoup d'entre eux ne sont pas impliqués dans la décomposition de molécules exogènes. Les isoenzymes CYP3A4, CYP2C9, CYP2C19, CYP2D6 et CYP1A2 sont majoritairement responsables du métabolisme des médicaments chez l'homme (Figure 4). Ces cytochromes présentent une large diversité de substrats, si bien que des molécules très différentes peuvent être métabolisées par la même isoenzyme [56,70,71]. Il est intéressant de noter qu'un principe actif peut être substrat et il sera alors métabolisé par le cytochrome, il peut également avoir un effet inhibiteur et bloquer le cytochrome, et enfin il peut aussi avoir un effet inducteur vis-à-vis de la cellule et favoriser l'expression d'un ou plusieurs cytochromes [72–74]. Tout ceci peut conduire à des interactions médicamenteuses cliniquement significatives [75–77].

La détermination de l'affinité ou les études cinétiques d'une molécule pour un cytochrome P450 spécifique sont généralement utilisées pour déterminer si une molécule est inhibitrice, inductrice ou substrat. Cependant, ceci ne permet pas d'estimer la quantité du principe actif qui est métabolisée par l'organisme. En effet, le métabolisme d'un xénobiotique est un phénomène complexe faisant intervenir une multitude d'enzymes impliquées dans plusieurs réactions biochimiques. La stabilité métabolique du principe actif vis-à-vis de l'ensemble de ces réactions enzymatiques est donc à prendre en considération [78]. La stabilité métabolique peut être étudiée à l'aide de deux propriétés

physiologiques ADME, à savoir le temps de demi-vie ($t_{1/2}$) et la clairance hépatique (CL_H). Ces deux propriétés sont également étroitement liées à la phase d'élimination.

2.4. Elimination

La phase d'élimination (E) regroupe les phénomènes impliqués l'excrétion du principe actif et de ses métabolites. Ces phénomènes sont d'une importance cruciale, car toute insuffisance d'un organe responsable de l'élimination se traduit par un ralentissement de l'excrétion et des risques d'accumulation du principe actif dans les tissus pouvant engendrer des effets indésirables. Cette phase d'élimination du xénobiotique est principalement assurée par le foie et les reins.

Le principe actif présent dans le sang est acheminé vers le foie par le biais de la veine porte et de l'artère hépatique. La circulation sanguine dans le foie est constituée d'un réseau capillaire hautement ramifié (sinus hépatiques). L'endothélium des sinus hépatiques présente une structure particulière favorisant un contact étroit entre les cellules du foie (hépatocytes) et le sang. En effet, des pores de grande taille définissent un espace intermédiaire (espace de Disse) permettant l'échange direct des substances entre le sang et les hépatocytes. Les enzymes impliquées dans le métabolisme (Ch1 2.3) sont localisées au niveau des membranes cellulaires et des mitochondries des cellules hépatiques. Le principe actif et ses métabolites vont ainsi pouvoir traverser les membranes des hépatocytes selon les mécanismes de transport transcellulaire, dans le but d'être sécrétés dans les canicules biliaires complètement séparées de la circulation sanguine. De la sorte, la bile est un suc digestif sécrété par les hépatocytes, mais également une voie d'élimination pour le foie. Le xénobiotique et ses métabolites sécrétés dans la bile vont regagnés l'intestin par le biais du canal cholédoque avant d'être potentiellement réabsorbés par l'organisme ou être excrétés dans les selles. La fraction non modifiée du principe actif est véhiculée dans le sang par le biais des veines sus-hépatiques puis quitte le foie via la veine cave inférieure, afin d'être redistribuée dans l'organisme.

A l'instar du foie, le sang acheminé par l'artère rénale va emprunter un réseau capillaire hautement vascularisé dans les reins. Ces capillaires vont desservir plusieurs millions de néphrons, qui représentent les unités primaires de l'élimination rénale. Ainsi, la phase d'élimination débute par la filtration glomérulaire des molécules dissoutes dans le sang et leur transfert dans l'urine primitive au niveau de la capsule de Bowman. Les membranes de cette capsule possèdent de larges jonctions permettant une perméation paracellulaire

élevée de l'eau, des xénobiotiques et d'autres composants sanguins à l'exception des protéines et des cellules. En règle générale, les molécules ayant une masse moléculaire inférieure à 5 000 Da subissent une filtration importante. A la sortie du glomérule, les capillaires sanguins vont former une artériole dite efférente. Cette artériole efférente va donner naissance à un réseau de capillaires péritubulaires intimement liés aux tubules du néphron. Dans le tubule proximal, certaines molécules médicamenteuses peuvent être activement sécrétées du flux sanguin dans la lumière tubulaire par le biais de transporteurs. Par exemple, le passage des composés pénicillines et glucuronides vers le tubule urinaire est facilité par des transporteurs d'anions, la morphine et la procaïne sont transportées par des transporteurs de cations, ou encore la digoxine est transportée par les P-gp. Les molécules ayant traversé cette barrière physiologique vont se retrouver dans l'urine primitive pour suivre l'écoulement tubulaire du néphron. Au cours de cet écoulement, l'eau contenue dans l'urine primitive va être réabsorbée, ce qui aura pour conséquence une diminution du volume urinaire et une augmentation de la concentration du principe actif dans l'urine définitive. Un gradient de concentration va alors être observé entre l'urine et le liquide interstitiel ou le sang. Ce gradient de concentration entraînera la réabsorption partielle des molécules lipophiles dans le sang selon un phénomène de diffusion passive. La réabsorption des molécules polarisables dépend de leur degré de dissociation au pH urinaire. Ainsi, seule l'entité neutre de la molécule peut quitter la lumière du tubule en suivant le gradient de concentration. La réabsorption des molécules dans la circulation sanguine se produit également par des mécanismes de transport passifs et actifs. Le comportement du principe actif dépend des mêmes mécanismes actifs que ceux rencontrés durant la phase d'absorption. D'autre part, les métabolites dans le sang sont plus facilement éliminés par les reins que le principe actif, car l'augmentation de leur polarité augmente leur solubilité dans l'urine primitive lors de la filtration glomérulaire. A la fin de ce processus d'élimination rénale, l'urine s'écoule ensuite à travers l'uretère jusqu'à la vessie d'où elle est excrétée.

Plusieurs paramètres physiologiques permettent d'estimer l'élimination d'un xénobiotique. L'un des plus importants est la clairance totale, qui correspond à la capacité de l'organisme à éliminer la molécule après avoir regagné la circulation systémique. Elle représente le volume de plasma épuré par heure via les différents organes impliqués dans la phase d'élimination. La clairance totale est déterminée en cumulant les clairances rénale et hépatique. Un autre paramètre physiologique est le temps de demi-vie d'élimination correspondant au temps nécessaire pour que la concentration du

xénobiotique dans le plasma diminue de moitié. Ce temps de demi-vie est dépendant du volume de distribution et de la clairance. D'autres paramètres peuvent être déterminés comme par exemple la fraction du principe actif non métabolisé retrouvée dans les selles ou l'urine.

2.5. Toxicité

La mise au point de nouveaux médicaments exige que des études toxicologiques soient effectuées afin d'estimer l'innocuité de ces nouvelles entités moléculaires avant la demande d'AMM auprès des autorités compétentes. La toxicologie est une discipline scientifique qui étudie les effets indésirables d'une molécule chimique sur les organismes vivants, pouvant provoquer des dommages sévères, voire mortels dans les cas les plus graves. Ces effets indésirables sont les résultats visibles d'une action pharmacologique excessive du médicament, induite par le surdosage du principe actif ou par une exposition prolongée de l'organisme vis-à-vis de la molécule incriminée. La dose d'un principe actif peut être responsable de sa toxicité. Cette toxicité liée à la dose est connue depuis le XVI^{ième} siècle lorsque Paracelse énonça ce principe fondamental de la toxicologie moderne. De nos jours, la toxicité aigüe permet de déterminer les doses toxiques chez l'animal et dans les organes. Les tests *in vivo* doivent obligatoirement être réalisés chez au moins deux espèces de mammifères et avec deux voies d'administration différentes [79]. Le médicament va être administré à dose croissante pendant plusieurs jours. Lors de ces expérimentations, la dose administrée va être unique pour chaque animal afin de tester un large panel de concentrations. D'autres paramètres peuvent être pris en compte comme par exemple le sexe des animaux, afin d'étudier la toxicité à l'échelle de sous-populations. Plusieurs doses vont être déterminées : i) la dose létale (LD_{50}) est définie comme la dose à laquelle 50% des animaux décèdent ii) la dose maximale tolérée est la dose administrée à partir de laquelle des effets toxiques sont observés, mais n'affectent pas les fonctions vitales des animaux iii) la dose maximale sans effet toxique, comme son nom l'indique, est la dose maximale pour laquelle aucun effet toxique n'a été observé chez l'animal. D'autres mesures de la toxicité sont utilisées pour les études chez l'animal, mais également chez l'Homme lors des essais précliniques et des essais cliniques, comme par exemple la NOAEL qui est la dose pour laquelle aucun effet indésirable n'est observé, ou encore la LOAEL qui est la dose minimale pour laquelle des effets indésirables sont observés. Seuls les médicaments toxiques à la dose usuelle (posologique) seront considérés comme dangereux. Les causes de la toxicité des

molécules peuvent être induis par cinq phénomènes biologiques que nous présentons ci-dessous [80].

Le premier phénomène est la toxicité hors cible (*off-target*). Le médicament, normalement conçu pour interagir avec une cible spécifique, peut interagir avec d'autres macromolécules lors de son voyage dans l'organisme, induisant des effets indésirables non prévus. Par exemple, l'inhibition des CYP450 peut entraîner une diminution de l'élimination des xénobiotiques et de leurs métabolites. Si un médicament est métabolisé principalement par une voie unique du métabolisme, l'inhibition des CYP450 peut entraîner une accumulation de ce composé dans l'organisme due à la saturation des processus enzymatiques. Ainsi, l'inhibition des CYP450 peut induire une toxicité accrue [81–84]. Un autre exemple concerne la cardiotoxicité induite par les médicaments bloqueurs des canaux potassiques hERG provoquant des arythmies cardiaques. Ce phénomène de toxicité peut être évité grâce à l'élaboration de candidats médicaments plus sélectifs [85].

Le deuxième phénomène est la toxicité induite par la modulation de la cible primaire (*on-target*). Le principe actif se lie à la cible thérapeutique, mais cette cible peut être présente dans plusieurs tissus non souhaités entraînant une toxicité. C'est par exemple le cas des composés statines, dont les propriétés hypercholestérolémiques sont dues à l'inhibition de la 3-hydroxy-3-méthylglutaryl CoA réductase (HMGCoA) dans le foie. Les effets indésirables des statines sont engendrés par l'inhibition de l'HMGCoA réductase dans les muscles, perturbant les modifications post-traductionnelles des protéines pouvant causer une myopathie [86,87].

Le troisième phénomène est l'hypersensibilité immunologique. Le concept de cette toxicité, développé en grande partie sur la base des travaux pionniers de Landsteiner en 1935 [88], considère que le médicament ou ses métabolites peuvent réagir avec les protéines de l'organisme (sous forme d'haptène) pour induire la production d'anticorps et des réponses immunitaires [89]. En d'autres termes, le principe actif n'est pas complètement stable et a le potentiel de se lier de façon covalente aux protéines afin d'induire la production d'anticorps responsables des réactions allergiques. C'est par exemple le cas des pénicillines pour lesquelles les réactions allergiques sont connues depuis de nombreuses années.

Le quatrième phénomène de toxicité est associé au métabolisme des médicaments. Les xénobiotiques peuvent être métabolisés en espèces chimiques actives (bioactivation). Les métabolites bioactivés peuvent se lier de façon covalente aux protéines natives,

modifiant de la sorte leurs fonctions biologiques ou entraînant des réponses immunitaires comme dans le cas des haptènes [90]. Par exemple, l'acétaminophène est métabolisé en une espèce bioactive hépatotoxique. Ainsi, la prédiction du métabolisme des xénobiotiques, et plus particulièrement l'identification des métabolites potentiellement bioactifs, est l'un des défis actuels de la recherche thérapeutique [91].

Pour finir, les réactions idiosyncratiques représentent le dernier phénomène responsable de la toxicité des médicaments. Ce sont des réactions indésirables imprévisibles qui ne se produisent pas chez tous les patients en raison de leurs différences génétiques pouvant engendrer un métabolisme différent du xénobiotique ou des réponse immunitaires spécifiques à chaque individu [92]. Elles introduisent un degré important d'incertitude dans le processus de développement de nouveaux médicaments, car elles ne peuvent pas être anticipées à l'aide des tests chez l'animal et ne sont détectées qu'après de nombreux tests chez l'Homme.

L'exploration de la toxicité d'un xénobiotique peut être conduite à l'aide d'approches *in vitro*, *in vivo*, mais également *in silico*. L'un des thèmes centraux de la toxicologie computationnelle est l'élaboration de modèles prédictifs à l'aide des données déjà publiées obtenues lors de tests expérimentaux. Ainsi, une grande diversité de méthodes *in silico* sont utilisées pour estimer la toxicité de nouvelles molécules, comme par exemple les modèles QSAR ou les approches d'amarrage moléculaire. Les approches QSAR consistent à rechercher une relation mathématique entre les descripteurs moléculaires, utilisés pour décrire la structure chimique des molécules, et les valeurs expérimentales de toxicité. Ces approches sont fondées sur l'hypothèse que des molécules dont les caractéristiques structurelles sont similaires peuvent être associés aux mêmes effets biologiques, et par conséquent qu'elles possèdent des profils de toxicité équivalents. Elles représentent une alternative à l'utilisation des tests *in vivo* chez l'animal et *in vitro*, afin de limiter le temps et les coûts des tests expérimentaux lors de la découverte de nouvelles molécules bioactives. Ainsi, le système législatif européen REACH est le meilleur exemple de l'utilisation des approches QSAR à cet effet. Ce programme européen promeut l'utilisation de méthodes alternatives aux tests *in vivo*, comme par exemple les approches *in silico*. L'objectif est alors d'intégrer le principe des 3R (*Reduction, Refinement, Replacement)*, qui vise à la réduction, au perfectionnement ou au remplacement de l'utilisation des tests sur les animaux [93].

Dans le cadre des approches *in silico*, la toxicité des candidats médicaments peut être étudiée à différentes échelles. En effet, des méthodes traitent les phénomènes de toxicité globale, comme par exemple les modèles de cancérogénicité, tandis que d'autres méthodes traitent les facteurs de la toxicité conduisant à des manifestations locales, comme par exemple la toxicité propre à un organe. De ce fait, plusieurs facteurs de toxicité sont étudiés lors de la recherche et du développement de nouveaux médicaments et peuvent être la cause de l'arrêt d'un projet (Figure 5) [94].

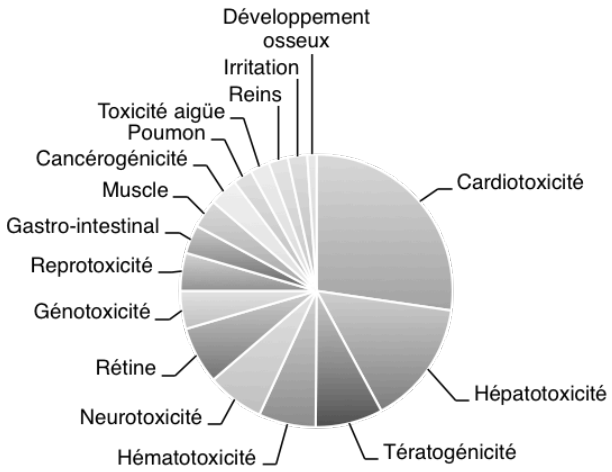


Figure 5 : Facteurs de toxicité responsables de l'échec des candidats médicaments lors des études toxicologiques.

F. Peter Guengerich a publié en 2011 les causes majoritaires des échecs lors des études toxicologiques basées sur les expériences de Merck et Bristol-Myers Squibb sur la période de 1993 à 2006.

Il existe donc des modèles *in silico* pour un grand nombre des phénomènes toxicologiques [95,96]. Cependant, la création de modèles fiables et robustes dépend essentiellement de la qualité et de la disponibilité de données toxicologiques. Ainsi, la génotoxicité et la cancérogénécité, sont des toxicités très étudiées car des criblages à haut débit ont été récemment rendus publics par le biais de plusieurs bases de données gouvernementales [97–99]. La génotoxicité traduit la capacité du médicament à modifier les gènes. Cette toxicité spécifique est étudiée à l'aide d'une multitude de tests *in vitro* et *in vivo* recommandés par les agences réglementaires. Des tests *in vitro* (Ames) sont réalisés afin d'estimer la mutagénicité, c'est-à-dire la capacité d'un xénobiotique à engendrer la mutation des gènes [100]. De plus, un test *in vivo* doit être obligatoirement réalisé pour déterminer la capacité d'un xénobiotique à endommager l'ADN. Cette toxicité est étroitement liée aux effets cancérogénotoxiques, c'est-à-dire qu'un médicament peut

provoquer des cancers. La cancérogénécité d'un médicament est estimée à l'aide de test *in vivo* sur une période de 2 ans minimum. L'objectif durant ces études est de déterminer si l'exposition à des doses répétées du médicament provoque au long terme le cancer.

D'autre part, l'hépatotoxicité et la cardiotoxicité sont de plus en plus étudiées de nos jours, en raison de leur impact important dans le processus de R&D de nouveaux médicaments (Figure 5) et du nombre croissant d'études expérimentales publiées [101–103]. L'hépatotoxicité traduit la capacité d'un médicament à provoquer des lésions hépatiques qui se manifeste par des inflammations, voire dans les cas les plus graves des nécroses du foie [104]. Elle est explorée à l'aide d'études *in vivo* (DILI) qui consiste à identifier les médicaments réponsables de lésions hépatiques lors des essais cliniques [105]. Les causes de ces phénomènes sont difficilement explicables et sont généralement considérées commme des phénomènes idiosyncratiques.

En revanche, il existe peu de modèles traitant de la reprotoxicité, de la neurotoxicité et de l'hématotoxicité du fait d'un manque de données avéré dans la littérature [106,107]. L'hématotoxicité représente les effets toxiques d'un médicament sur les composants du sang, comme par exemple les globules rouges (anémie), les lymphocytes (lymphopénie) ou encore les plaquettes (thrombopénie), se manifestant par une altération sévère des fonctions hématopoïétiques du sang. De plus, la tératogénécité, se manifestant par des anomalies ou des malformations du fœtus chez les femmes enceintes, est une autre toxicité majoritairement responsable du taux d'échec des candidats médicaments.

En résumé, les études toxicologiques utilisées pour prouver l'innocuité des candidats médicaments sont traditionnellement réalisées lors des essais précliniques et des essais cliniques. De ce fait, le retrait des molécules à ce stade avancé du processus de R&D entraîne des pertes économiques considérables. En réponse à cette problématique, la tendance actuelle est d'identifier les dangers potentiels dès l'étape de sélection des *leads*, afin de déceler les propriétés indésirables à un stade plus précoce. Par conséquent, les approches (Q)SAR ont un rôle prépondérant à jouer dans l'élaboration de ces nouvelles stratégies, en permettant de filtrer des composés qui présentent un profil ADME-Tox défavorable.

3. Les approches (Q)SAR orientées vers la prédiction ADME-Tox

Les groupes de recherche universitaires comme privés, ainsi que les entreprises pharmaceutiques s'intéressent tous à la réduction du nombre d'échecs lors de la mise au point de nouveaux médicaments. Anticiper les profils ADME-Tox défavorables à l'aide des approches *in silico* dès l'étape d'optimisation des *leads* est une solution pour réduire les coûts et le temps nécessaire à la découverte d'un médicament. Afin de répondre à cette problématique actuelle, lors de cette thèse nous nous sommes intéressés à l'élaboration de modèles (Q)SAR basés sur un apprentissage automatique (*machine learning*) [108]. Pour cette raison, nous avons jugé important de présenter plus en détails ces approches *in silico* ainsi que les méthodologies et stratégies indispensables pour la création de modèles de prédiction. Nous présenterons également les limitations actuelles des prédictions ADME-Tox et l'implication de ces approches dans le processus de découverte de nouveaux médicaments.

3.1. Généralités sur les approches (Q)SAR

3.1.1. Historique

La modélisation (Q)SAR est née il y a plus de 150 ans, lorsque des scientifiques ont essayé de quantifier les relations entre la structure chimique et l'activité de petites molécules organiques. Crum-Brown et Fraser ont émis l'hypothèse en 1868 que l'action physiologique d'une substance était fonction de sa constitution chimique. Ainsi, ils ont proposés le postulat selon lequel une modification de la structure chimique engendrait une modification de l'activité biologique [109]. A la fin du XIX^{e} siècle, des scientifiques comme Richet, Meyer ou encore Overton démontrent que l'activité d'un composé est fortement corrélée à ses propriétés constitutionnelles [110,111]. D'autres approches de ce type ont été élaborées tout au long du XX^{e} siècle, comme par exemple les travaux de Hammett ou de Taft [112,113]. Mais c'est en 1964 que Hansch *et al.* parviennent à élaborer une équation mathématique permettant de prédire le coefficient de partition H_2O/Octanol à partir des constantes électroniques de Hammett développées quelques années plus tôt : c'est le premier modèle QSAR de l'histoire [114]. Dans les années qui ont suivies, la nécessité de résoudre de nouveaux problèmes et la contribution de nombreux chercheurs, ont généré des milliers de variations de la méthodologie proposée par Hansch, ainsi que des approches complètement nouvelles, comme par exemple l'introduction des études qualitatives des relations structure-activité. A l'heure actuelle,

fondé sur l'utilisation systématique de modèles mathématiques, les approches (Q)SAR sont les outils de base pour la conception contemporaine de nouveaux médicaments, se situant à l'intersection de la chimie, des statistiques et de la biologie.

Par conséquent, un modèle (Q)SAR peut être décrit comme un modèle statistique qui approxime une fonction (f) à partir descripteurs moléculaires (X) et d'une activité biologique (Y) selon l'équation Y = f(X). L'objectif d'un modèle est alors de capter la relation existante entre les descripteurs moléculaires et l'activité, afin de créer des règles génériques permettant d'expliquer l'activité étudiée. Le but est ensuite d'appliquer ces règles afin de prédire l'activité de molécules inconnues à partir de leurs descripteurs moléculaires. L'élaboration d'un modèle nécessite trois composantes : i) un ensemble de données constitué des mesures expérimentales de l'activité biologique pour un groupe de molécules (Y) ; ii) des valeurs de descripteurs moléculaires pour décrire la structure des molécules (X) ; iii) des méthodes statistiques, pour identifier la relation entre les deux ensembles de données (f). Plusieurs méthodologies peuvent être employées afin de créer un modèle de prédiction, mais toutes doivent respecter les règles de bonnes pratiques approuvées par la communauté scientifique, comme par exemple les principes de l'Organisation de Coopération et de Développement Economique (OCDE).

3.1.2. Principes de l'OCDE et bonnes pratiques

Lors du congrès QSAR de Setubal (Portugal) en mars 2002, les lignes directrices pour déterminer la validité des modèles (Q)SAR, en particulier à des fins réglementaires [115] ont été définies. Suite à ce congrès, les membres de l'OCDE ont convenus de 5 principes fondamentaux à suivre pour établir la validité scientifique d'un modèle (Q)SAR. Il est intéressant de noter que des observations similaires ont été proposées par Unger et Hansch en 1973 [116]. Ces principes sont un aperçu des points impératifs auxquels doit répondre le modèle pour être considéré comme cohérent, fiable et reproductible [117]. Les cinq principes adoptés par l'OCDE sont les suivants :

i) Une activité définie – pour s'assurer que les données modélisées soient homogènes (même activité, même unité et dans la mesure du possible même protocole expérimental).

ii) Un algorithme non ambigu – Les méthodes statistiques utilisées pour construire un modèle (Q)SAR doivent être explicitement détaillées dans la mesure du possible, afin d'assurer la reproductibilité des prédictions.

iii) Un domaine d'applicabilité défini (Ch1 3.2.4.7) – Les modèles sont construits sur des sous-ensembles spécifiques de l'espace chimique. Des prédictions peu fiables peuvent être obtenues pour des molécules qui n'appartiennent pas au sous-espace chimique couvert par le modèle.

iv) Des mesures appropriées de la qualité, de la robustesse et de la prédictivité du modèle – Les performances des modèles doivent être évaluées à l'aide de métriques détaillées suite à une validation interne puis une validation externe (Ch1 3.2.4.5). La validation interne consiste à estimer la qualité statistique du modèle sur le jeu d'apprentissage (jeu de données utilisées pour créer le modèle). La validation externe consiste à estimer la capacité du modèle à prédire de nouvelles molécules (pouvoir prédictif) à l'aide d'un jeu de test (Ch1 3.2.4.3).

v) Une interprétabilité du modèle (si possible) – Le modèle doit être interprétable chimiquement, c'est-à-dire qu'il doit permettre d'expliquer l'importance de chaque descripteur moléculaire sur la propriété modélisée, afin de définir des règles génériques. Le respect de ce principe n'est pas toujours évident à mettre en œuvre, car certaines méthodes d'apprentissages (algorithmes) ainsi que des descripteurs peu explicatifs, comme les empreintes moléculaires, ne permettent pas une interprétation facile. Toutefois, si ce dernier principe n'est pas respecté, un modèle statistique disposant de bonnes performances peut tout de même être utilisé s'il respecte les principes précédemment énoncés.

Comme nous venons de le voir, cette liste fait appel à des termes spécifiques dans la construction de modèles QSAR qui seront détaillés par la suite. L'OCDE a également présenté des indications pour l'interprétation et la mise en œuvre de ces principes [118]. Cependant, ces règles fondamentales ne sont pas toujours suivies. Dearden, Cronin et Kaiser ont fourni en 2009 une liste de 21 erreurs rencontrées lors de l'élaboration, l'interprétation et l'utilisation d'un modèle (Q)SAR [119]. Les erreurs rencontrées comprennent i) l'utilisation de données inadéquates, incorrectes ou non homogènes, ii) l'utilisation de descripteurs moléculaires colinéaires, incompréhensibles et/ou incorrects, iii) l'utilisation d'un nombre excessif de descripteurs, iv) le manque de normalisation des descripteurs, v) la présence de doublons dans le jeu de données utilisé pour élaborer le modèle, vi) la mauvaise sélection du jeu d'apprentissage et du jeu de test, vii) l'omission injustifiée de points de données, viii) le sur-apprentissage des données, ix) l'absence ou l'utilisation d'un domaine d'applicabilité inadéquat, x) ou encore l'absence, la mauvaise

utilisation de méthodes statistiques ou l'incapacité de valider correctement le modèle. En 2009, Scior *et al.* ont suggéré des pistes pour reconnaître de tels pièges et de les éviter [120]. D'autres initiatives de ce type ont été proposées par Varnek et Baskin en 2012 et Cherkasov *et al.* en 2014 [121,122]. Ces travaux sont importants pour la communauté scientifique et constituent de véritables bases de travail et de réflexion pour l'élaboration de nouveaux modèles (Q)SAR.

En résumé, le défi actuel n'est plus de développer un modèle capable de prédire l'activité pour le jeu d'apprentissage d'une manière statistiquement valable, mais de développer un modèle qui a la capacité de prédire avec précision l'activité de composés chimiques encore non testés [123]. De ce fait, nous allons voir les aspects pratiques à prendre en compte lors de la création d'un modèle de prédiction en respect avec les bonnes pratiques énoncées précédemment.

3.2. Aspects pratiques : élaboration de modèles (Q)SAR

Comme énoncé par les principes de l'OCDE, la création d'un modèle de prédiction nécessite quatre éléments comme illustrés par la Figure 6, à savoir : i) des données expérimentales (Y) ; ii) des descripteurs moléculaires (X) ; iii) la sélection d'une méthode d'apprentissage ; iv) une validation adaptée du modèle statistique.

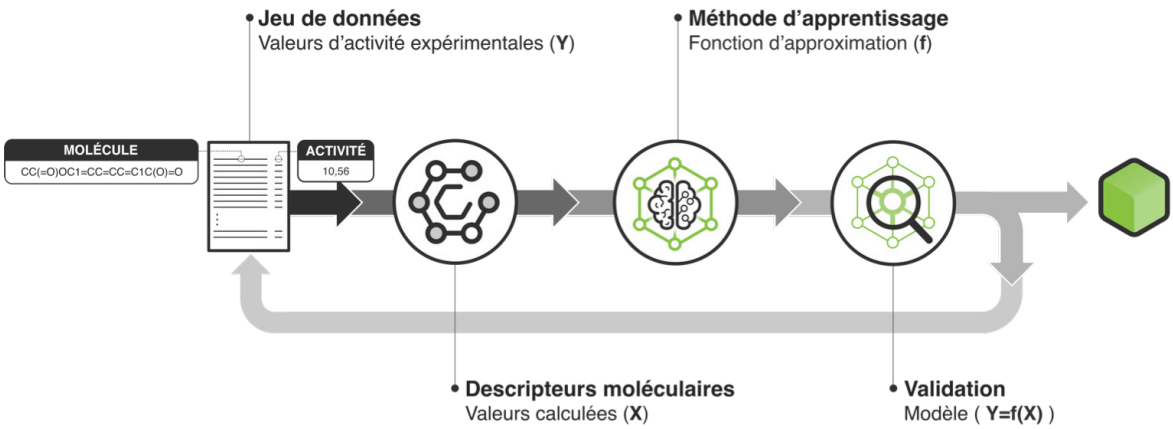


Figure 6 : Représentation des 4 composantes indispensables à la création d'un modèle (Q)SAR.

3.2.1. Jeux de données expérimentaux

La mise en place d'un modèle de prédiction nécessite avant tout une étape de fouille et de collecte de données. L'objectif est de constituer un jeu de données, c'est-à-dire une liste de composés chimiques pour lesquels des valeurs d'activité ont été

mesurées expérimentalement. Pour cela, différentes sources de données peuvent être exploitées à savoir des sources de données internes ou publiques.

Les données internes à un laboratoire présentent plusieurs avantages, comme par exemple l'accès à une méthode expérimentale permettant de mesurer l'activité biologique et l'information sur le protocole utilisé. Dans le cas où aucune mesure de l'activité n'est disponible pour l'étude d'un espace chimique précis, il est alors possible de générer un jeu de données, *a priori* homogène (même méthode), permettant d'élaborer un modèle performant pour une série chimique précise. Cependant, l'obtention de ces données homogènes requiert du temps et des coûts supplémentaires du fait de la réalisation expérimentale des mesures. Lorsque le laboratoire ne dispose pas des ressources nécessaires pour la détermination expérimentale de l'activité souhaitée, il lui est possible de constituer un jeu de données à partir d'une grande variété de sources publiques. Certaines de ces sources proposent des données formatées et spécifiques à une activité (jeux de données publiés), tandis que d'autres proposent une grande diversité de mesures expérimentales extraites de la littérature pour une multitude d'activités biologiques (bases de données). Quelle que soit la nature de la source utilisée, un jeu de données doit posséder trois caractéristiques essentielles, à savoir des données fiables, des données homogènes et une taille suffisante.

3.2.1.1. Fiabilité et homogénéité des données

Avant toute étape de modélisation, une préparation du jeu de données est nécessaire afin de s'assurer de la qualité i) des représentations moléculaires et ii) des données expérimentales.

Plusieurs études ont été menées pour estimer et améliorer la qualité des informations transmises par les sources de données publiques [124–126]. Bien que le nombre d'erreurs soit faible, les résultats de ces études montrent que toutes les sources transmettent des informations structurales erronées, qui peuvent être introduites par inadvertance lors de la conversion de structures moléculaires à l'aide d'outils automatiques ou lors de la transcription par un humain. Ainsi, l'activité expérimentale d'une molécule n'est pas associée à la bonne structure moléculaire. Dans ce cas, une molécule peut disposer de plusieurs structures qui vont être considérées comme indépendantes mais disposant d'une valeur d'activité similaire. Ceci peut avoir pour conséquence une perte de sensibilité du modèle pour les séries congénériques. En plus des structures erronées, des problèmes peuvent également survenir en raison de la duplication des structures, de la

présence de mélanges, ou encore de la présence d'isomères. Dans ce cas, les descripteurs moléculaires calculés à partir de deux représentations non homogènes ne disposeront pas des mêmes valeurs numériques. Il est important de noter qu'une structure peut avoir différentes formes tautomériques qui ne constituent en aucun cas une erreur de représentation moléculaire. Ainsi, la vérification des structures et leur standardisation sont indispensables pour la création de modèles de prédiction robustes.

Les valeurs d'activité expérimentales doivent être dans la mesure du possible homogènes, c'est-à-dire que toutes les mesures doivent exprimer la même propriété biologique, avoir la même unité, et pour finir lorsque plusieurs protocoles ont été utilisés, qu'ils soient comparables. Cependant, il n'est pas rare de rencontrer des erreurs de retranscription ou de conversion d'unité dans les bases de données publiques. Ainsi, le modèle peut avoir des difficultés pour identifier les tendances existantes entre les molécules et leur activité. Il existe néanmoins des solutions pour identifier les valeurs aberrantes ou erronées comme la comparaison de plusieurs points de mesure pour une même molécule, ou encore l'identification des individus mal prédits par le modèle (outliers). Il est à noter que l'utilisation de ces solutions est limitée lorsque des sauts d'activité (activity cliffs) sont observés pour des séries congénériques. D'autre part, la majorité des jeux de données rencontrés dans la littérature n'apportent pas l'information sur la variabilité inhérente des données biologiques, et ne nous transmettent pas les conditions opératoires utilisées pour réaliser la mesure. Une incertitude supplémentaire s'applique lorsque plusieurs sources de données sont combinées, dans le but de créer un modèle (Q)SAR couvrant un plus vaste espace chimique. Le risque de ce genre de pratique est de mélanger des mesures non homogènes rendant le jeu de données inutilisable pour une application en (Q)SAR.

Les données ADME-Tox sont généralement déterminées tardivement dans le processus de *drug discovery* (Ch1 1.2) et ne sont pas rendues publiques dans la grande majorité des cas (Ch1 1.3). Par conséquent, un nombre limité de données expérimentales est disponible dans la littérature. Il n'est pas rare de rencontrer des jeux de données ADME-Tox qui combinent des mesures provenant de protocoles expérimentaux quelque peu différents. Cependant, la question est de savoir quelle est la taille optimale du jeu de données pour la création d'un modèle de prédiction.

3.2.1.2. Taille du jeu de données

La taille des jeux de données doit être suffisante pour représenter de manière significative l'espace chimique désiré. Certains modèles peuvent être créés à partir de jeux de données non congénériques dans le but de couvrir un vaste espace chimique (modèles globaux), tandis que d'autres peuvent être conçus pour examiner une série chimique spécifique (modèles locaux). Dans les deux cas, les données doivent représenter aussi largement que possible l'espace chimique souhaité afin que le modèle ait un domaine d'applicabilité optimal (Ch1 3.2.4.7). Aucune règle générique à ce sujet n'a été proposée dans la littérature, mais certaines études ont été entreprises afin de déterminer l'effet de la taille du jeu de données sur la construction et les performances d'un modèle. Ainsi, Roy et al. ont montrés que la réduction de la taille du jeu de données avait un effet négatif sur le pouvoir prédictif d'un modèle 127. Ils recommandent que la taille optimale soit fondée sur la capacité du jeu de données à couvrir l'espace chimique ciblé, mais également sur les descripteurs utilisés ou encore sur la méthode d'apprentissage employée pour générer le modèle. D'autres recommandations proposées par Tropsha sont fondées sur les limitations techniques et la qualité du modèle 128. Les jeux de données trop grands peuvent rendre difficile la construction du modèle, tandis que les jeux de données trop petits peuvent souffrir des phénomènes de corrélation aléatoire ou de sur-apprentissage (Ch1 3.2.4.1). Ainsi, si nous réduisons le nombre d'individus (molécules) tout en conservant le nombre de descripteurs, les chances de sur-apprentissage du modèle augmentent, ce qui entraîne une diminution du pouvoir prédictif du modèle. De ce fait, une sélection des descripteurs (Ch1 3.2.2.3) est indispensable, et des méthodes de validation doivent être adoptées pour vérifier que le modèle ne soit pas soumis à ces phénomènes de corrélation aléatoire et de sur-apprentissage (Ch1 3.2.4.1).

En résumé, la fiabilité et la taille sont deux éléments à prendre en considération afin de constituer un jeu de données valide pour la création de modèles de prédiction. Il n'est donc pas surprenant de voir que les principes de l'OCDE prêtent une attention particulière à la préparation des données. Une fois le jeu de données vérifié et validé, le calcul des descripteurs moléculaires peut être effectué.

3.2.2. Descripteurs moléculaires

La structure moléculaire d'un composé contient implicitement toutes ses informations chimiques. Théoriquement, il est possible de définir des données numériques (descripteurs) capables d'extraire une partie des informations chimiques [129].

Depuis plusieurs décennies, des recherches se sont concentrées sur la façon de capturer et de convertir l'information codée dans la structure moléculaire en un ou plusieurs descripteurs. L'intérêt de la communauté scientifique pour les descripteurs moléculaires est attesté par le grand nombre de descripteurs proposés et calculables à l'aide d'outils logiciels dédiés (Table 1). Le nombre de descripteurs augmente continuellement avec la complexité croissante des systèmes chimiques étudiés. De ce fait, nous ne présenterons que les descripteurs moléculaires liés aux petites molécules organiques.

Logiciels	Licence	Description
RDKit [130]	Libre	RDKit permet de calculer 200 descripteurs et plusieurs types d'empreintes moléculaires.
CDK [131]	Libre	CDK propose environ 286 descripteurs (constitutionnels, topologiques, géométriques et électroniques), ainsi que qu'une dizaine d'empreintes moléculaires.
MORDRED [132]	Libre	MORDRED permet de calculer 1825 descripteurs (constitutionnels, topologiques, géométriques et électroniques).
ISIDA[133]	Libre	ISIDA propose plusieurs combinaisons permettant de calculer des descripteurs basés sur le nombre de fragments structuraux.
VolSurf+ [134]	Commerciale	VolSurf+ propose 128 descripteurs majoritairement physico-chimiques adaptés pour la prédiction ADME-Tox.
MOE [135]	Commerciale	.3.2.2.1.a.1 MOE possède une grande variété de descripteurs : des clés structurales, des indices topologiques, E-state ainsi que des propriétés physico-chimiques.
Dragon [136]	Commerciale	Dragon 7 permet de calculer 5270 descripteurs moléculaires. Cet outil est très utilisé par la communauté scientifique.
ADRIANA.Code [137]	Commerciale	ADRIANA.Code propose des descripteurs de formes, physico-chimiques, topologiques et électroniques.
ChemAxon [138]	Commerciale	ChemAxon propose des outils permettant de calculer des descripteurs physico-chimiques et structuraux.

Table 1 : Logiciels proposant des descripteurs moléculaires utilisés pour la prédiction des propriétés ADME-Tox.

Il existe six familles de descripteurs moléculaires [139] : i) Les descripteurs constitutionnels sont simples et couramment utilisés pour refléter l'information chimique sans prendre en compte la connectivité des atomes (nombre de liaisons, nombre d'atomes, poids moléculaire, etc.). ii) Les descripteurs structuraux ou fragmentaux représentent l'absence, la présence, ou encore le nombre d'occurrences d'un fragment spécifique dans la structure d'une molécule, comme par exemple les empreintes moléculaires. iii) Les

descripteurs topologiques sont fondés sur la théorie des graphes et traduisent la connectivité des atomes observée dans la représentation symbolique des molécules (indices de connectivité de Wiener, de Zagreb, etc.). iv) Les descripteurs électroniques sont utilisés pour décrire les aspects électroniques des atomes, des liaisons, voire même des fragments de la structure moléculaire (moments dipolaires, charges partielles ou formelles, etc.). v) Les descripteurs thermodynamiques établissent un lien entre la structure et le comportement chimique d'une molécule et représentent en général des paramètres physico-chimiques comme par exemple le $LogP_{o/w}$. vi) Les descripteurs géométriques sont déterminés à partir des coordonnées 3D des atomes et ils permettent de capturer des informations tridimensionnelles concernant la taille, la forme, le volume, la surface et la distribution des atomes. Les descripteurs 3D nécessite la minimisation de la structure moléculaire à l'aide de la mécanique moléculaire ou de la mécanique quantique, ce qui peut augmenter drastiquement le temps nécessaire pour les calculer.

Les caractéristiques qui rendent un descripteur idéal pour la construction d'un modèle (Q)SAR sont les suivantes [140] : i) un descripteur doit être corrélé avec l'activité étudiée et montré une corrélation négligeable avec les autres descripteurs. ii) un descripteur doit être applicable à une vaste diversité de composés. iii) un descripteur doit être calculé rapidement et ne doit pas nécessiter d'expérimentation. iv) un descripteur doit générer des valeurs dissemblables pour des molécules structurellement différentes et cela même si les différences structurelles sont minimes. v) et pour finir, un descripteur doit être simple à calculer, répétable et interprétable dans la mesure du possible.

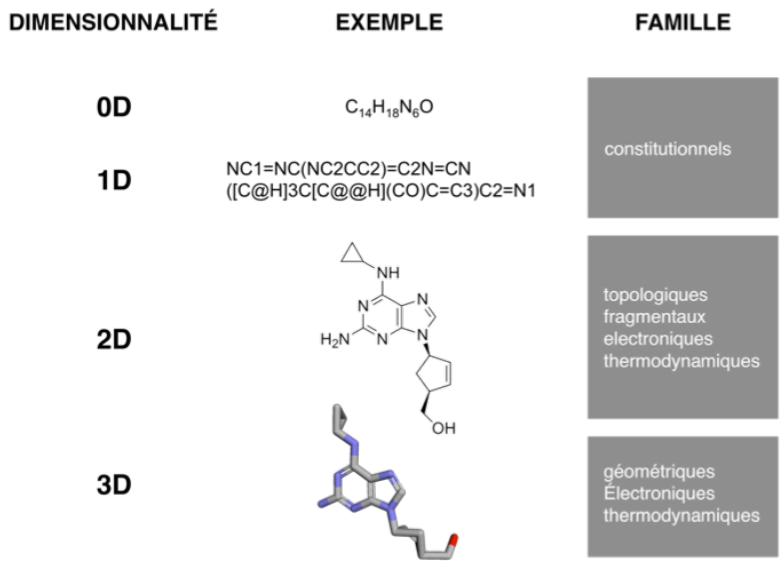


Figure 7 : Classification des descripteurs moléculaires en fonction de leur dimensionnalité.

Les descripteurs peuvent être classés en fonction de leur dimensionnalité comme illustré par la Figure 7 [129,141]. Lors cette thèse, nous nous sommes intéressés aux descripteurs de type 0D, 1D et 2D uniquement, car nous ne connaissons pas la conformation bioactive d'une molécule, c'est-à-dire la conformation biologiquement responsable du ou des phénomène(s) ADME-Tox étudié(s). Par conséquent, nous avons fait abstraction de phénomènes importants liés à la structure 3D des composés, comme par exemple les liaisons hydrogènes intramoléculaires indispensables pour le passage de la barrière hémato-encéphalique [142].

Les descripteurs fragmentaux nécessitent plus d'explications et seront présentés plus en détails ci-après. D'autre part, les valeurs brutes de descripteurs, une fois calculées, ne peuvent pas être utilisées en l'état pour la création d'un modèle. Des étapes de préparation du jeu de données sont nécessaires, dont la sélection des descripteurs fait partie.

3.2.2.2. Descripteurs fragmentaux

Les descripteurs fragmentaux indiquent la présence, l'absence, ou l'occurrence de fragments structuraux dans la représentation 2D des molécules étudiées [143]. Ces descripteurs peuvent être encodés sous différents formats, à savoir : i) une séquence binaire (séquence de *bits*) où une valeur de 0 indique l'absence et une valeur de 1 indique la présence du fragment considéré ; ii) une séquence de nombres entiers qui définit l'occurrence des fragments. Ces séquences sont des empreintes moléculaires. Conçues initialement pour effectuer des recherches par similarité dans de vastes librairies chimiques [144], ces empreintes ont été détournées en tant que descripteurs moléculaires pour créer des modèles de prédiction [145], car elles permettent de lier directement l'information structurale à l'activité biologique. Nous ferons la distinction entre les clés structurales et les empreintes moléculaires hachées.

a) Clés structurales

Ces empreintes moléculaires sont basées sur l'utilisation d'un dictionnaire de fragments, c'est-à-dire que les fragments recherchés sont prédéfinis dans une liste de fragments potentiels. L'utilisation de ce dictionnaire pour définir les *bits* de l'empreinte moléculaire permet une interprétation non ambiguë en associant directement une clé (*bit*) à une sous-structure (fragment). Ainsi, les informations non codées dans l'empreinte moléculaire ne peuvent pas être représentées sans l'ajout d'une clé supplémentaire dans la liste des fragments potentiels. La société Molecular Design Limited (MDL) a défini dans

les années 2000 une série de 166 clés structurales, aussi nommées clés MACCS [146,147], qui est optimisée pour la recherche par similarité moléculaire. Bien que ces empreintes soient simples à mettre en œuvre, leur utilisation peut être limitée par le temps nécessaire à leur détermination, car elles requièrent une recherche par sous-structure. Ces descripteurs sont très populaires pour la prédiction de la toxicité à l'aide des systèmes basés sur des règles (*expert and rule-based systems*). Les descripteurs générés vont prendre en compte les fragments toxiques communément rencontrés lors de la conception de nouveaux médicaments et vont permettre de créer des modèles robustes, comme par exemple ceux proposés dans le cadre du projet NCI-60 [148,149]. D'autres empreintes moléculaires de ce type peuvent être rencontrées comme les empreintes PubChem [150], BCI [151], ou TGT et TGD [152].

b) Empreintes moléculaires hachées (ou hashed fingerprints)

Ces empreintes moléculaires sont générées à partir de la représentation moléculaire 2D sur laquelle une fonction de hachage va être appliquée afin d'en extraire des fragments structuraux permettant de décrire la molécule. Plusieurs modes de fragmentation peuvent être utilisés pour hacher le graphique moléculaire. Pour cette raison, nous les avons regroupé en deux grandes catégories, à savoir les fragmentations basées sur un chemin de connectivité linéaire et les fragmentations circulaires (Figure 8).

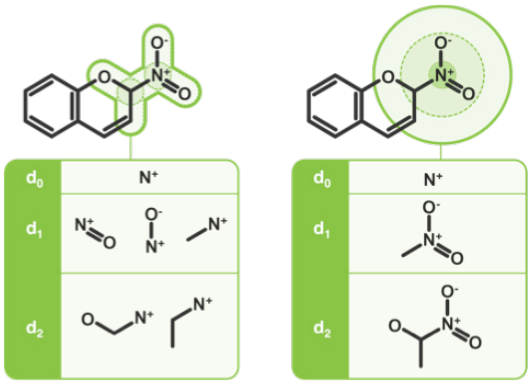


Figure 8 : Chemins de fragmentation linéaire et circulaire pour le nitrogène à une distance comprise entre 0 et 2.

Les empreintes qui se basent sur une fragmentation linéaire reposent sur l'énumération de toutes les combinaisons existantes entre plusieurs atomes en considérant le chemin le plus court (connectivité linéaire) permettant de les relier. Le chemin de connectivité linéaire est ensuite haché et va représenter un *bit* dans l'empreinte moléculaire. Les empreintes basées sur une fragmentation circulaire, ou *Atom-centred fragments* (ACF), consistent à définir une sphère d'inclusion autour d'un atome central afin d'extraire les

chemins (connectivité multidirectionnelle) qui relient cet atome central aux atomes lourds voisins. Ces processus de fragmentation sont répétés à partir de tous les atomes de la représentation moléculaire et pour toutes les distances comprises entre 0 et une distance maximale prédéfinie (longueur de l'empreinte désirée). Cependant, un fragment ne représente qu'une fraction de la structure chimique. Ceci signifie qu'un seul *bit* ne peut pas être relié directement à une caractéristique structurale donnée. Par conséquent, il n'est pas rare qu'un *bit* soit encodé par plusieurs paramètres différents pouvant engendrer des collisions [153]. Ceci peut avoir des conséquences désastreuses lors de la mise en place d'un modèle de prédiction, car l'information d'un *bit* est basée sur deux ou plusieurs motifs structuraux indépendants, ce qui entraîne des descripteurs ambigus pouvant avoir une importance trompeuse dans le modèle de prédiction.

Historiquement, les paires d'atomes (AP) [154] sont les premières empreintes de fragmentation linéaire à avoir été proposées. Les paires d'atomes ont été utilisées pour créer des modèles de prédiction et ont permis d'envisager de nouveaux descripteurs de ce type dans le cadre des approches (Q)SAR [145]. Cependant, elles ne prennent en compte ni le type de liaison, ni l'environnement local de chaque atome. De ce fait, les torsions topologiques (TT) [155] ont été créées par Nilakantan *et al.* afin de répondre partiellement aux limitations rencontrées dans le cadre des paires d'atomes. De nos jours, les fragments moléculaires sous-structuraux (SMF) [156], proposé par Varnek *et al.* dans le cadre de la plateforme ISIDA [157], sont des empreintes topologiques plus perfectionnées qui comptabilisent l'occurrence de chaque fragment dans les représentations moléculaires des composés étudiés. ISIDA propose trois modes de fragmentation [158], dont deux modes sont basés sur l'approche de fragmentation linéaire (*Sequences*, *Triplets*) et un mode basé sur l'approche de fragmentation circulaire (ACF). Ces descripteurs donnent une information plus explicite que les AP et les TT au sujet des propriétés atomiques et des types de liaisons pour définir chaque fragment. Les descripteurs SMF sont d'une importance majeure pour la communauté scientifique et ont été utilisés avec succès pour l'élaboration de modèles (Q)SAR.

Le concept des empreintes circulaires a été proposé par Morgan dès 1965 et a donné naissance aux empreintes de Morgan [159]. Durant la dernière décennie, plusieurs empreintes ont été créés comme les ECFP, ECFC, FCFP et FCFC [130,138] qui prennent en compte des informations chimiques complémentaires lors de la fragmentation. Les empreintes de connectivité étendue (acronyme EC) utilisent les propriétés atomiques, les propriétés des liaisons et la chiralité lors de la fragmentation. Les empreintes de

connectivité fonctionnelle (acronyme FC) considèrent les propriétés pharmacophoriques durant la fragmentation. D'autre part, des chaînes de *bits* (FP) et des chaînes d'occurrence (FC) peuvent être différentiées selon le type d'encodage souhaité dans l'empreinte finale. De plus, le diamètre utilisé pour définir la sphère d'inclusion, ou encore le nombre de *bits* désirés dans l'empreinte finale, sont des paramètres variables qui donnent accès à une multitude d'empreintes moléculaires. Plusieurs études ont été menées afin de déterminer les empreintes qui permettent de décrire et de différentier de façon optimale des structures chimiques [153,160,161]. Ces études montrent que des empreintes moléculaires de 1024 *bits,* obtenues à l'aide de sphères d'inclusion de rayon 4 ou 6, permettent d'obtenir des modèles de bonnes performances dans la majorité des cas. Il est important de noter que ces empreintes sont de nos jours les plus populaires pour l'élaboration de modèles (Q)SAR.

3.2.2.3. Sélection des descripteurs

Comme nous venons de le voir, nous pouvons décrire une structure chimique à l'aide de plusieurs milliers de descripteurs moléculaires (Table 1). Néanmoins, tous ne sont pas pertinents pour l'élaboration d'un modèle. En effet, plusieurs d'entre eux peuvent apporter des informations chimiques redondantes (colinéarité) ou peuvent ne pas être pertinents pour l'activité biologique étudiée, et ainsi affecter la découverte de la relation descripteurs-activité. Pour cette raison, la sélection des descripteurs les plus pertinents est considérée comme l'une des tâches les plus difficiles et cruciales pour la modélisation (Q)SAR [162,163]. L'objectif est alors de choisir les descripteurs qui apportent une aussi bonne, voire une meilleure précision tout en exigeant moins de données. Ainsi, cette réduction est souhaitable car elle permet de réduire la complexité du modèle afin de mieux généraliser la relation structure-activité en la rendant plus facile à comprendre et à interpréter. Il est à noter que cette notion de simplicité est l'une des exigences abordées par les principes de l'OCDE (Ch1 3.1.2). Trois stratégies principales peuvent être utilisées pour la sélection de sous-ensemble de descripteurs pertinents [164] :

i) **Stratégie *Filter*** : La pertinence de chaque descripteur (score) est calculée en considérant les propriétés intrinsèques de ces derniers. La sélection des descripteurs se fait en fonction d'un seuil à partir duquel les descripteurs donnant un score inférieur vont être supprimés. Ces filtres sont très utilisés car ils permettent de sélectionner les descripteurs les plus avantageux parmi un jeu de données de haute dimension (plusieurs milliers de descripteurs). Plusieurs exemples permettent d'illustrer la stratégie *Filter*, comme la suppression des

descripteurs de variance nulle voire quasi-nulle qui n'apportent aucune information au modèle, ou encore la suppression des descripteurs corrélés qui apportent une information chimique redondante dans le modèle. L'avantage premier de cette stratégie est qu'elle est simple et efficace cependant, son inconvénient majeur est qu'elle se base sur des analyses univariées ou bivariées qui ne prennent pas en considération les relations étroites entre les descripteurs lors de l'apprentissage.

ii) **Stratégie *Wrapper*** : Des sous-ensembles de descripteurs sont utilisés conjointement à une méthode d'apprentissage afin de déterminer l'influence de chacun d'entre eux sur la prédiction. Les performances du modèle temporaire permettent de sélectionner un sous ensemble de descripteurs apportant l'erreur minimale. L'un des exemples les plus connus de la stratégie *Wrapper* est l'approche graduelle (*stepwise*) ascendante ou descendante à l'aide d'une régression linéaire multiple, qui consiste à ajouter ou supprimer un descripteur de l'ensemble de données jusqu'à ce que ce ne soit plus rentable en termes de performance du modèle. L'avantage de cette méthode est qu'elle permet de considérer les dépendances existantes entre les descripteurs. Cependant, elle requiert un temps de calcul plus long que la stratégie *Filter* et elle n'est pas spécifique à une méthode d'apprentissage.

iii) **Stratégie *Embedded*** : Cette approche est identique à la stratégie *Wrapper*, à la seule différence qu'elle est spécifique à une méthode d'apprentissage. Ainsi, l'algorithme utilisé pour la sélection des descripteurs est le même que celui employé pour créer le modèle de prédiction.

En résumé, cette étape de sélection peut être effectuée selon différentes stratégies et a pour but ultime de réduire l'impact des descripteurs apportant une information inutile (variance nulle), une information redondante (colinéarité) ou une information non pertinente à la prédiction de l'activité étudiée. Une fois les descripteurs sélectionnés, une méthode d'apprentissage va devoir être choisie pour créer le modèle de prédiction.

3.2.3. Apprentissage et algorithmes

Il existe une grande diversité d'algorithmes pouvant être appliqués dans le cadre des approches (Q)SAR. Le défi consiste à choisir la méthode d'apprentissage qui convienne le mieux à l'exploration de la propriété en court d'investigation. Pour cela, des

méthodes supervisées et non supervisées peuvent être utilisées et seront décrites par la suite (Figure 9).

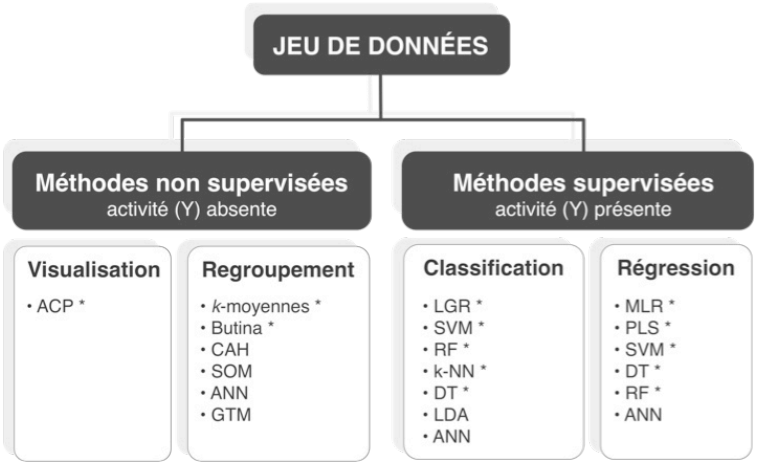


Figure 9 : Liste des méthodes d'apprentissage (non exhaustive) .

L'astérisque (*) représente les méthodes explorées dans le cadre de cette thèse. Abréviations : (ACP : Analyse en Composantes Principales ; *k*-NN : *k-Nearest Neighbors* ; CAH : Classification Ascendante Hiérarchique ; SOM : Self Organizing Map ; ANN : Artificial Neural Network ; LGR : Logistic regression ; SVM : Support Vector Machine ; RF : Random Forest ; LDA : Linear Discriminant Analysis ; MLR : Multiple Linear Regression ; DT : Decision Tree) ; GTM : Generative Topographic Mapping.

3.2.3.1. Méthodes non supervisées

Les méthodes non supervisées sont descriptives et n'utilisent pas directement l'information sur l'activité biologique (Y). Certaines d'entre elles, comme par exemple l'Analyse en Composantes Principale (ACP), permettent de visualiser et de décrire l'organisation des molécules dans l'espace multidimensionnel des descripteurs. D'autres permettent de regrouper les molécules, comme par exemple lors des approches de regroupement (*clustering*) de type *k*-moyennes (*k-means*) ou encore de type Butina. Ces méthodes peuvent être utilisées afin d'identifier les éventuelles valeurs aberrantes (*outliers*) présentes dans le jeu de données [141].

a) Analyse en Composantes Principales (ACP)

L'ACP est une technique de réduction de dimensionnalité qui est largement utilisée pour l'analyse de données. Elle décompose un jeu de données multivariées (plusieurs descripteurs) à l'aide d'un ensemble de composantes orthogonales successives qui expliquent la variance maximale observée dans le jeu de données. Ces composantes orthogonales vont être appelées composantes principales et correspondent à des combinaisons linéaires des descripteurs moléculaires. Chaque composante

principale va exprimer une part de la variance expliquée présente dans le jeu de données. Ainsi, l'ACP définit un espace de plus faible dimensionnalité que le jeu de données initial ce qui permet au modélisateur d'analyser de façon plus aisée les données sur lesquelles il travaille. L'ACP permet alors de visualiser les molécules et les descripteurs sur l'ensemble des plans bidimensionnels définis par les combinaisons de deux composantes (Figure 10), ce qui est plus facilement interprétable.

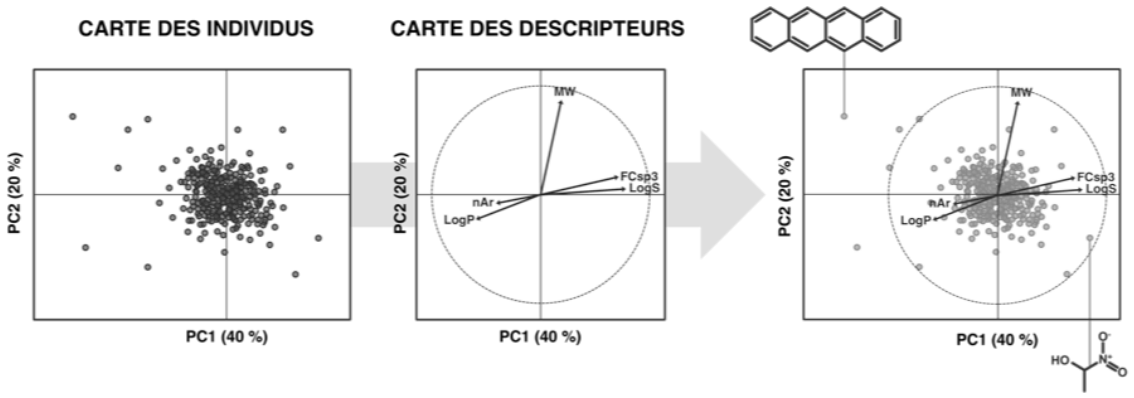


Figure 10 : Représentation schématique d'une ACP sur les deux premières composantes principales.

b) k-moyennes

La méthode des *k*-moyennes est un algorithme de regroupement qui identifie *k* groupes à l'intérieur d'un espace multidimensionnel définit par les descripteurs moléculaires [165–167]. L'objectif de cette méthode est de déterminer l'appartenance de chaque individu à l'un des *k* groupes fixés par l'utilisateur. L'algorithme utilisé pour réaliser ce regroupement est itératif descendant, c'est-à-dire qu'il va effectuer plusieurs fois la même opération unitaire afin de ne retenir qu'une solution optimale. L'opération unitaire de l'algorithme comporte plusieurs étapes (Figure 11) : i) les centroïdes des *k* groupes vont être positionnés dans l'espace des descripteurs moléculaires ; ii) les distances entre chaque individu et les centroïdes des *k* groupes sont ensuite calculées ; iii) un individu est affecté au groupe dont il est le plus proche en fonction de sa distance avec les centroïdes des *k* groupes. Lors de la première itération, les centroïdes des groupes vont être positionnés aléatoirement, tandis que pour les itérations ultérieures les centroïdes moyens de chaque groupe vont être utilisés (étape i). Ce processus se répète jusqu'à la convergence vers un minimum local, où la dissimilarité moyenne à l'intérieur de chaque groupe n'évolue plus [168].

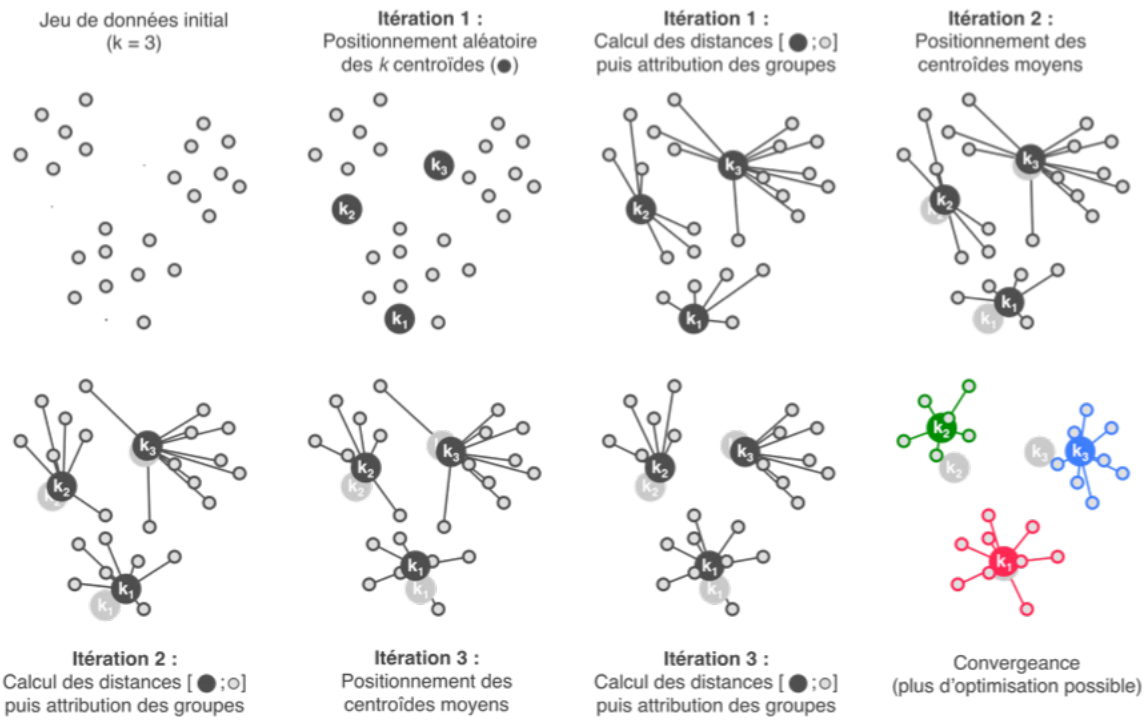


Figure 11 : Principe des k-moyennes.

Bien que cette méthode soit simple à mettre en œuvre et efficace, elle possède cependant des inconvénients. La solution que nous apporte cette méthode n'est pas unique, car elle est dépendante du positionnement aléatoire des centroïdes lors de l'étape initiale, des itérations répétées durant l'opération de regroupement, ou encore de la distance utilisée (Euclidienne, City Block, Mahalanobis, etc.) pour assigner les individus à un groupe. Ainsi, les résultats obtenus ne sont pas reproductibles d'une expérience à une autre. Le deuxième inconvénient est que cette méthode ne permet pas de déterminer le nombre optimal de groupes pour séparer les individus selon leurs caractéristiques propres. En effet, le nombre *k* de groupes à explorer est fixé par l'utilisateur. L'utilisation de cette méthode nécessite donc une connaissance approfondie du jeu de données exploré.

c) Butina

Du nom de son créateur, Darko Butina, cette méthode utilise une approche similaire au célèbre regroupement de Jarvis-Patrick. L'objectif de ce dernier est de définir des groupes de molécules en fonction de la densité locale observée pour chaque individu (ou molécule) représenté dans l'espace multidimensionnel décrit par les descripteurs moléculaires. Une matrice de distances est ensuite calculée pour définir la proximité de tous les individus projetés dans l'espace cartésien. Pour chaque individu un nombre k_{min} de plus proches voisins va être sélectionné. Deux individus vont être regroupés s'ils sont considérés comme voisins et s'ils possèdent un nombre p_{min} de voisins communs. Dans

le cadre du regroupement Butina, l'indice de Tanimoto est utilisé en tant que distance et il est calculé à partir des empreintes moléculaires. De ce fait, la distance est directement reliée à la similarité structurale des composés du jeu de données. D'autre part, Le nombre k_{min} n'est pas fixé, mais il est déterminé à l'aide d'un seuil de similarité selon lequel toutes les molécules ayant un indice de Tanimoto supérieur ou égal à ce seuil vont être considérées comme voisines (Figure 12). Ainsi des listes de plus proches voisins vont être obtenues pour chaque molécule et vont être ordonnées en fonction de leur taille. Le *clustering* se fait ensuite de manière itérative en déterminant le composé ayant le plus grand nombre de voisins. La liste du composé sélectionné va constituer un groupe, et toutes les molécules de ce groupe vont être supprimées des listes encore non explorées. Le regroupement se termine lorsque toutes les listes initialement définies ont été explorées.

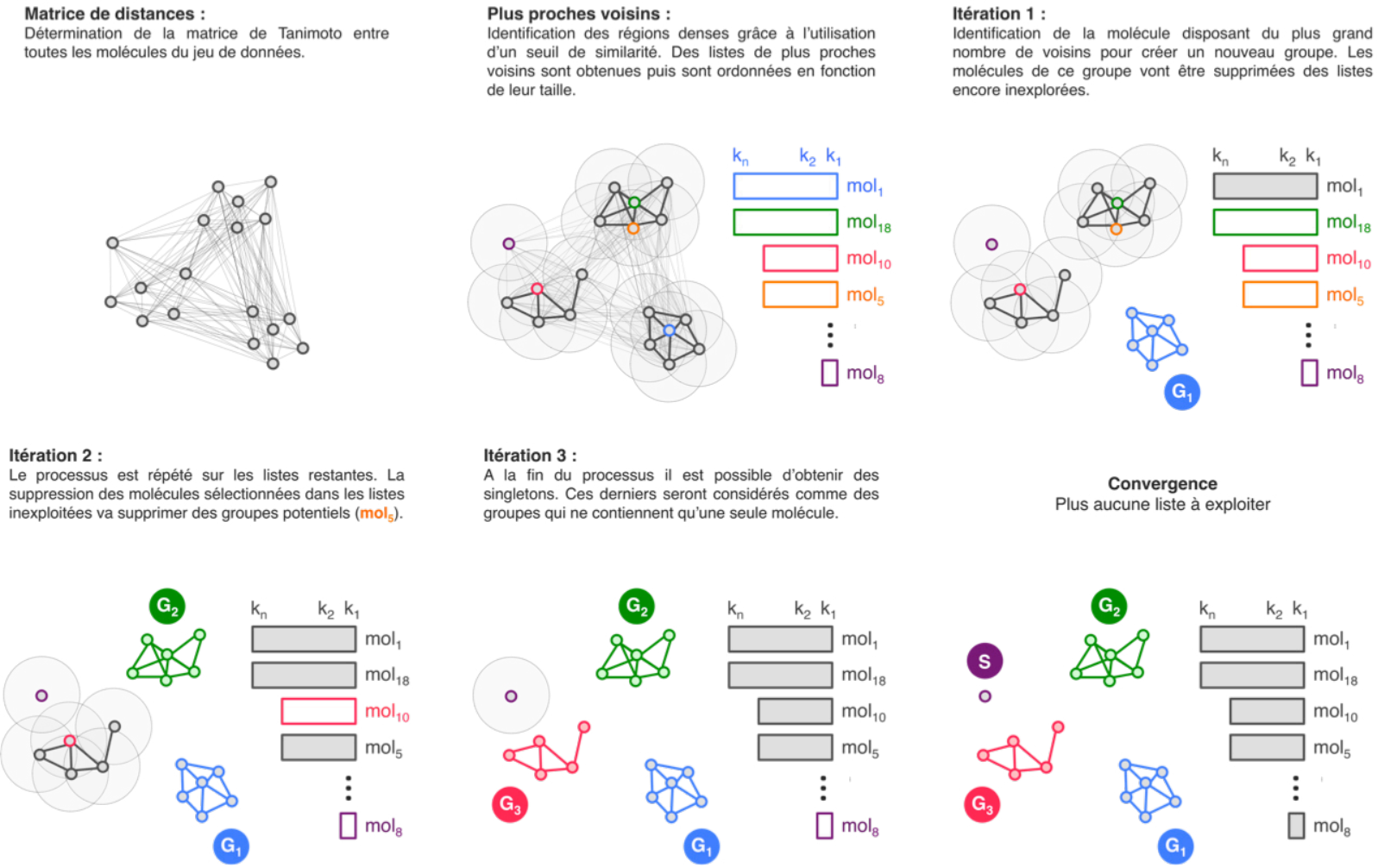


Figure 12 : Schéma de principe du *clustering* Butina.

L'avantage de ce regroupement par rapport à celui de Jarvis-Patrick est que seul le seuil de similarité est utilisé pour créer les groupes. En effet, l'approche de Jarvis-Patrick nécessite de définir au préalable les paramètres k_{min} et p_{min}. Ces paramètres ne sont pas concrets et nécessitent selon nous une étape de paramétrage afin de définir les valeurs optimales à utiliser pour le cas étudié. Dans le cas du regroupement Butina, le seuil de similarité est concret et permet de paramétrer rapidement l'algorithme pour obtenir des groupes de similarité désirée. D'autre part, ces approches de regroupement peuvent créer des groupes ne contenant qu'une seule molécule. Ces molécules isolées vont être appelées singletons. L'obtention de singletons peut être un réel désavantage lors d'un regroupement, car ces molécules seules ne sont attribuées à aucun groupe. Cependant dans le cadre du regroupement Butina, les groupes sont identifiés de façon à optimiser l'homogénéité des molécules qui les constituent, ce qui peut représenter dans certains cas un avantage.

3.2.3.2. Méthodes supervisées

Dans le cadre des méthodes supervisées, l'information sur l'activité biologique (Y) est utilisée lors d'une étape d'apprentissage. Ces méthodes sont prédictives et permettent d'établir la relation structure-activité des molécules entre les descripteurs moléculaires (X) et leur activité (Y). Les fonctions d'approximation (f) sont divisées en deux catégories en fonction de la nature de l'activité Y étudiée : régression et classification.

i) Une régression consiste à définir une fonction (f) à partir des descripteurs moléculaires (X) afin de prédire une activité biologique continue (Y). La valeur prédite est donc quantitative.

ii) Une classification consiste à définir une fonction (f) à partir des descripteurs moléculaires (X) et d'une activité biologique discrète (Y). La valeur prédite, appelée classe, est donc qualitative. Dans ce type de modèle, l'activité biologique modélisée est représentée la plupart du temps par deux classes (actif et inactif), mais des classes supplémentaires peuvent être ajoutées. Cependant, il est courant que les modèles de classification apportent également une information continue sur la valeur prédite [169]. Cette valeur correspond à la probabilité d'un composé d'appartenir à chacune des classes étudiées. On parlera alors de classification probabiliste. Cette probabilité varie entre 0 et 1, avec une valeur de 1 lorsque le modèle à une bonne probabilité de prédire correctement la classe

attribuée à la molécule testée. Une probabilité de 0,5 représente un cas aléatoire pour lequel le modèle ne parvient pas à classer correctement une molécule.

Il existe de nombreux algorithmes dont la complexité et le degré d'interprétation varient. La complexité du jeu de données et la nature de la propriété à modéliser (continue ou discrète) oriente le choix de l'algorithme à utiliser. Parmi les méthodes d'apprentissage les plus employées pour la prédiction des propriétés ADME-Tox [170], seuls les algorithmes utilisés dans le cadre de cette thèse seront présentés plus en détail ci-après.

a) Régression Linéaire Multiple

La régression linéaire multiple, aussi nommée *Multiple Linear Regression* en anglais (MLR), est une extension de la régression linéaire simple utilisée pour étudier la relation entre plusieurs descripteurs moléculaires indépendants et une activité biologique continue selon une équation linéaire, comme représentée par l'Equation 2 [168].

$$Y = \varepsilon + \alpha_j x_j + \cdots + \alpha_p x_p$$

Equation 2 : Equation générique d'une régression linéaire multiple.

Avec Y pour l'activité biologique modélisée, j l'indice du descripteur compris entre 1 et *p*, x pour les valeurs numériques du descripteur moléculaire, α pour la contribution du descripteur x dans le modèle linéaire, ε pour la déviation du modèle.

Cette technique est simple à mettre en œuvre et convient à l'analyse de petits ensembles de données, mais elle peut devenir problématique lorsque plusieurs descripteurs moléculaires présentent une corrélation élevée [120]. En effet, en présence de descripteurs corrélés, la MLR dispose de coefficient de régression instable et d'une diminution de l'interprétabilité du modèle. Par exemple, certains coefficients α_p peuvent être surestimés, voire dans certains cas avoir des signes erronés [171]. Par conséquent, l'utilisation de cette méthode d'apprentissage peut être limitée lorsque le mécanisme étudié est complexe.

b) Régression linéaire selon le critère des moindres carrés partiels

La régression linéaire selon le critère des moindres carrés partiels, aussi nommée *Partial Least Square* en anglais (PLS), permet de pallier aux limitations de la MLR en projetant les descripteurs moléculaires dans un nouvel espace [168]. Cette méthode d'apprentissage crée des composantes principales en maximisant la covariance entre les différents sous-ensembles de descripteurs afin de définir un nouveau sous-espace orthogonal. Cette méthode présente l'avantage d'être peu sensible à la colinéarité élevée

entre les descripteurs moléculaires. D'autre part, elle est très utile lorsque le jeu de données comporte plus de variables (descripteurs) que d'observations (molécules) [168]. Cette méthode permet également l'utilisation de noyaux permettant de définir des relations non linéaires entre l'activité et les descripteurs [172]. L'astuce noyaux sera explicitée plus en détails dans le cadre des machines à vastes marges.

c) Machine à vastes marges

Les machines à vastes marges, aussi nommé *Support Vector Machine* (SVM), ont été créées par Vapnik et Chervonenkis en 1964 [173]. Cet algorithme a d'abord été développé pour l'élaboration de classifications linéaires binaires et a été par la suite étendu à des résolutions non linéaires grâce à l'incorporation de l'astuce noyau (*kernel-trick*) [174,175]. D'autres extensions permettent l'utilisation de la SVM pour l'élaboration de modèles de régression et de modèles de classification multiclasses [176,177,167,168]. Le principe inhérent de ces méthodes d'apprentissage consiste à utiliser de vastes marges qui autorisent sans pénalité un petit écart entre la valeur réelle et la valeur prédite, mais qui interdit ou pénalise très fortement un gros écart. Pour cela, la SVM a pour but de trouver la séparation optimale entre deux classes d'individus (molécules) avec pour hypothèse que plus cette séparation est large, plus le modèle de classification est robuste. Dans le cas le plus simple, c'est-à-dire les problèmes séparables linéairement (Figure 13), l'algorithme SVM identifie l'hyperplan qui maximise la distance avec les deux populations de molécules à séparer. Une fois l'hyperplan défini, une SVM va maximiser la largeur (ε) et la courbure (C) de la marge (Figure 13) jusqu'à rentrer en contact avec un individu décrit dans l'espace multidimensionnel. Ceci va permettre de déterminer la fonction de pénalité grâce à laquelle toute molécule nouvellement prédite dans la zone frontière, comprise entre les plans supports, sera fortement pénalisée [178]. Il est à noter, que des marges molles peuvent être employées afin de permettre aux individus mal séparés de franchir la frontière décrite par les plans supports.

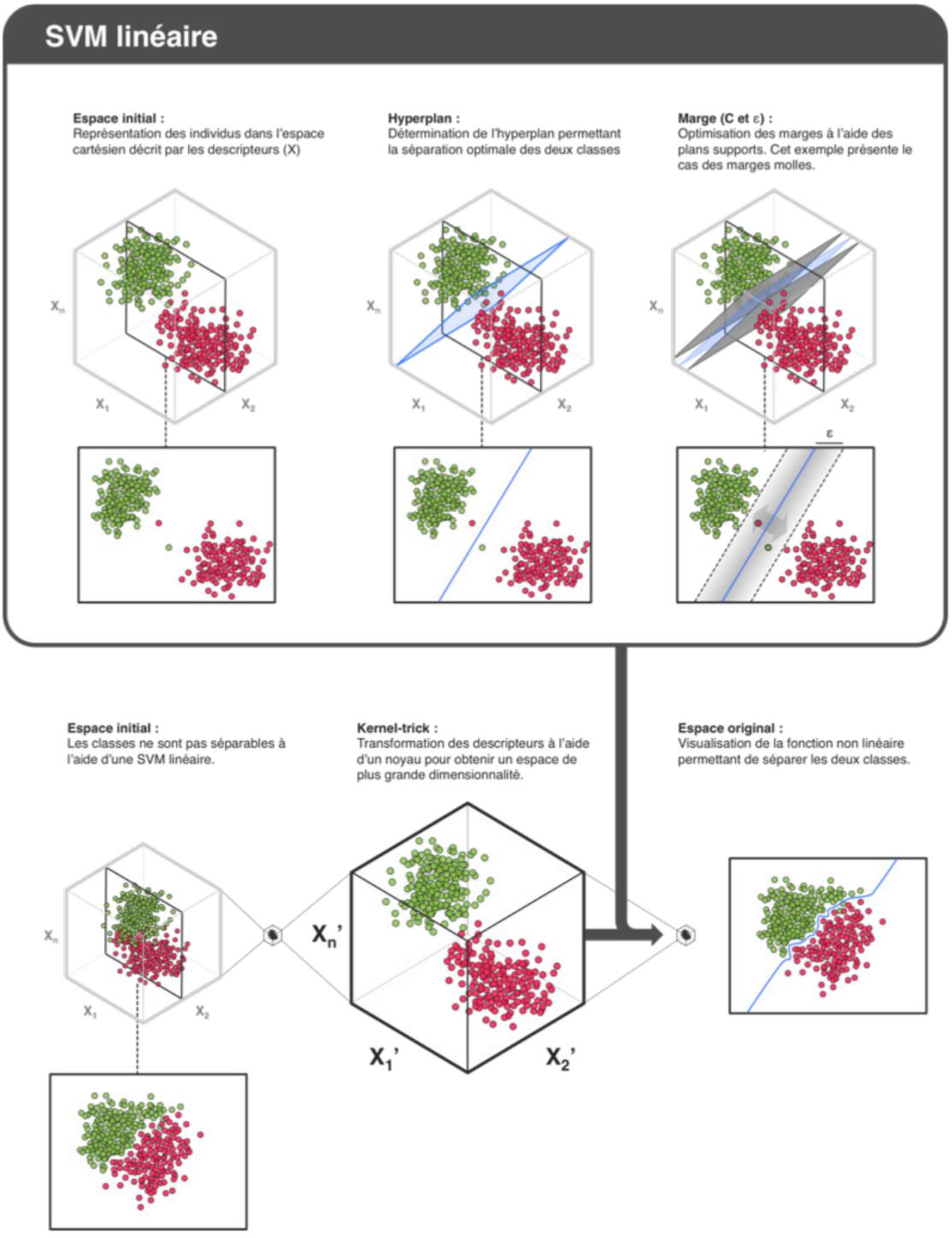


Figure 13 : Représentation schématique des principes fondamentaux de la SVM.

Dans le cas d'un problème non séparable linéairement à l'aide de la SVM, des noyaux peuvent être utilisés pour optimiser la construction de la marge. Le but est de transformer au préalable le problème non linéaire dans l'espace des variables explicatives, en un problème linéaire de plus grandes dimensions (espace de Hilbert) suite à la transformation des descripteurs à l'aide d'un noyau. Un noyau est un algorithme ou une fonction mathématique prenant en entrée une matrice symétrique positive (valeurs des

descripteurs pour un ensemble d'individus), afin de la transformer en une matrice de produit scalaire définissant un espace de redescription (*feature space*). Ainsi, des noyaux linéaires, polynomiaux, gaussiens ou encore sigmoïdaux peuvent être appliqués à cet effet. Suite à cette transformation, le problème peut être résolu en utilisant les approches déjà discutées dans le cadre de la SVM linéaire. Lors de la conversion de l'espace de dimensionnalité supérieure en espace original, la séparation des classes peut représenter une fonction non linéaire dans cet espace comme illustré par la Figure 13 [167,168].

L'utilisation de l'astuce noyau a permis une application beaucoup plus large des algorithmes SVM. Cependant leur utilisation nécessite l'optimisation des paramètres du noyau en plus de la marge (C et ε). Le paramétrage le plus performant sera ensuite utilisé pour construire le modèle final. Cette méthode d'apprentissage présente de nombreux avantages. Elle i) permet de travailler sur des jeux de données disposant d'un grand nombre de descripteurs moléculaires, ii) est dans la majorité des cas efficace lorsque le nombre de descripteurs est nettement supérieur au nombre d'individus étudiés, iii) est polyvalente grâce à l'utilisation de différents noyaux, ce qui permet d'optimiser le modèle final en ajustant la séparation des classes. Néanmoins, cette méthode d'apprentissage présente également des inconvénients : i) elle est sensible à la normalisation des descripteurs ; ii) un sur-apprentissage du modèle peut être observé lorsque l'astuce noyau est utilisée conjointement à un jeu de données ayant un nombre d'individus largement inférieur au nombre de descripteurs ; iii) les SVM sont des approches dites de boîtes noires où aucune interprétation n'est fournie par le modèle.

d) Arbre de décision

Les arbres de décision ont pour but de répartir une population d'individus en classes selon un ensemble de descripteurs. Pour cela, un arbre divise l'ensemble des descripteurs en fonction de la distribution des classes par partitionnement récursif [168]. Par conséquent, un descripteur unique (X) va donner naissance à un nœud, aussi nommé point de division. Les seuils de fractionnement, qui permettent de diviser chaque descripteur, sont choisis de façon à optimiser la discrimination interclasses. Ce partitionnement se poursuit afin de créer des branches jusqu'à ce qu'un critère d'arrêt soit rempli, auquel cas une feuille est créée. Une feuille est un nœud qui contient des individus classés suivant la découpe des descripteurs élaborée par l'arbre de décisions et par conséquent ne donnera pas lieu à la création de nouvelles branches (Figure 14).

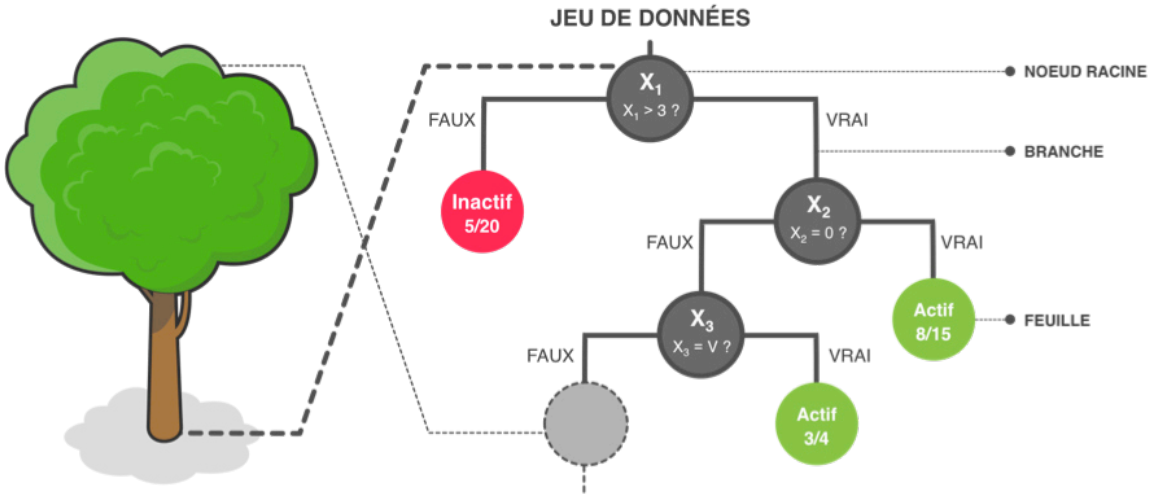


Figure 14 : Représentation schématique d'un arbre de décision.

Les nœuds sont colorés en gris et représentent le descripteur et son point de division lors de la phase d'apprentissage. Les nœuds colorés représentent les feuilles pour lesquelles une classification d'actif ou d'inactif sera faite. Le nombre de molécules actives et le nombre total de molécules sont représentés pour chaque feuille.

Au cours du processus d'apprentissage, le fractionnement d'un descripteur X est réalisé lorsqu'il est possible d'obtenir une discrimination supplémentaire des classes. Lors de cette étape, différents seuils de fractionnement vont être testés puis comparés à l'aide d'un critère de gain, qui s'apparente à l'index de Gini lors d'une classification ou la somme des carrés des écarts dans le cadre d'une régression [179]. Le seuil de fractionnement sélectionné est celui qui apporte la valeur optimale du critère de gain utilisé, cela signifie que le descripteur découpé selon ce seuil permet de faire la discrimination optimale entre les molécules actives et inactives.

Le processus d'apprentissage récursif est arrêté lorsqu'un nœud a atteint le nombre minimal d'adhésion, c'est-à-dire que le nombre de molécules assignées à une feuille n'est plus significatif pour considérer le nœud comme cohérent. Dès lors, la croissance de l'arbre sera stoppée et une feuille finale contenant l'ensemble des individus non classés sera créée. D'autre part, la taille des arbres de décisions est un facteur important, car un arbre trop petit peut être soumis au sous-apprentissage des données, tandis que les arbres de trop grande taille peuvent être soumis au sur-apprentissage des données. Généralement, un arbre sera cultivé jusqu'à ce que les nœuds atteignent cette taille minimale d'adhésion, puis une phase d'élagage sera entreprise. La procédure d'élagage vise à optimiser l'adéquation aux données par rapport à la taille de l'arbre en supprimant les branches n'apportant aucun gain supplémentaire [168]. Plusieurs algorithmes d'arbre de

décision sont communément rencontrés dans la littérature dont les arbres de type CART ou C45.

Ces modèles peuvent être utilisés pour prédire de nouvelles molécules, en utilisant leurs valeurs de descripteurs moléculaires et en les comparant à chaque nœud de l'arbre de décision. Ainsi, les seuils de fractionnement vont être considérés comme des règles mutuelles exclusives, facilitant la classification des molécules en fonction de leurs valeurs de descripteurs. La classe d'une molécule est attribuée en fonction du vote majoritaire dans la feuille sélectionnée. Prenons l'exemple de la Figure 14 : une molécule qui dispose des valeurs de descripteurs $X_1 > 3$, $X_2 = 1$ et $X_3 =$ VRAI sera attribuée à la feuille du troisième niveau de l'arbre de décision, et sera donc classée comme « actif », car la feuille correspondante contient 3 actifs et 1 inactif.

Cette approche présente plusieurs avantages : i) elle est simple à comprendre et à interpréter ; ii) elle ne nécessite pas ou peu de préparation de données et peut être appliquée sur les descripteurs non normalisés ; iii) elle est capable de prendre conjointement en compte des données numériques et catégoriques. Elle possède également des inconvénients, à savoir : i) un arbre de décision est sensible au nombre de descripteurs utilisés et peut fournir un modèle trop complexe ne permettant pas de généraliser les données. Le modèle sera donc enclin au sur-apprentissage. ; ii) Cette méthode d'apprentissage est instable aux petites variations du jeu de données (résultats non répétables).

Cependant, bien que cette approche possède des inconvénients qui peuvent être critiques pour la mise en place d'un modèle de prédiction, elle représente également l'arbre qui cache la forêt. En effet, la méthode des forêts aléatoires utilise plusieurs arbres de décision pour fournir un modèle hautement prédictif ne disposant pas de ces inconvénients.

e) Forêt aléatoire

Les forêts aléatoires, aussi nommées *Random Forests* (RF), peuvent être assimilées à un apprentissage d'ensemble, aussi nommé *ensemble learning*, dont le principe repose sur la création de plusieurs modèles (arbres de décision) afin de les combiner en un seul modèle (modèle d'ensemble), et ceci dans le but d'obtenir de meilleurs prédictions par rapport à un modèle unique [167]. Trois approches peuvent être utilisées dans le cadre de l'*ensemble learning*. Les approches de *bagging* ou de *boosting*

sont construites à l'aide du même algorithme d'apprentissage, mais sur des sous-ensembles de données différents. Alternativement en utilisant l'approche de *stacking*, un modèle pourrait être créé en combinant les prédictions de tous les modèles uniques. Les prédictions des modèles uniques peuvent être combinées par une simple moyenne des valeurs prédites (régression) ou par la détermination d'un vote majoritaire (classification). L'interprétation des modèles d'ensemble est intrinsèquement difficile ; même les algorithmes normalement interprétables, comme les arbres de décision, seront difficilement, voire pratiquement impossibles à comprendre de manière significative, une fois incorporés dans le modèle d'ensemble.

Dans le cadre du *bagging*, la variabilité est introduite dans le processus de modélisation en échantillonnant le jeu de données pour chaque modèle unique. Le jeu de données va donc être séparé de manière aléatoire en deux entités, à savoir le jeu d'apprentissage qui va être utilisé pour entrainer le modèle et le jeu de test qui va être utilisé pour évaluer le modèle. Ainsi, l'approche de type *bagging* est un processus durant lequel plusieurs modèles uniques, ayant une erreur intrinsèque importante, vont être combinés afin d'obtenir un modèle d'ensemble disposant d'une erreur moins élevée. Lorsque les erreurs entre les modèles uniques ne sont pas corrélées, une réduction significative de l'erreur peut être observée en faisant la moyenne des résultats de tous les modèles [180].

Le *boosting* est plus sophistiqué que le *bagging*. Contrairement à ce dernier, où les modèles sont formés de manière indépendante et aléatoire, le *boosting* cherche à réduire l'erreur en se basant sur les molécules mal prédites durant les étapes antérieures. Pour cela, le jeu de données d'un modèle unique va être échantillonné en se basant sur les données mal prédites par le modèle précédent. Cette approche permet de noter chaque modèle en fonction de sa capacité à séparer les classes étudiées. Par exemple dans le cas de la méthode AdaBoost, les modèles uniques sont pondérés (pénalité) de telle sorte que leur contribution dans le modèle d'ensemble sera plus ou moins importante pour la prédiction de nouvelles molécules.

Une forêt aléatoire applique l'approche du *bagging* aux arbres de décision de telle sorte qu'une "forêt d'arbres" est construite et que les réponses obtenues par tous les modèles uniques sont combinées pour obtenir une prédiction finale. Pour cela, une forêt aléatoire va échantillonner aléatoirement le jeu de données en fonction de l'espace chimique couvert (jeu d'apprentissage et jeu de test), mais également en fonction des descripteurs

moléculaires (Figure 15). La forêt aléatoire résultante est une collection moyenne d'arbres décorrélés.

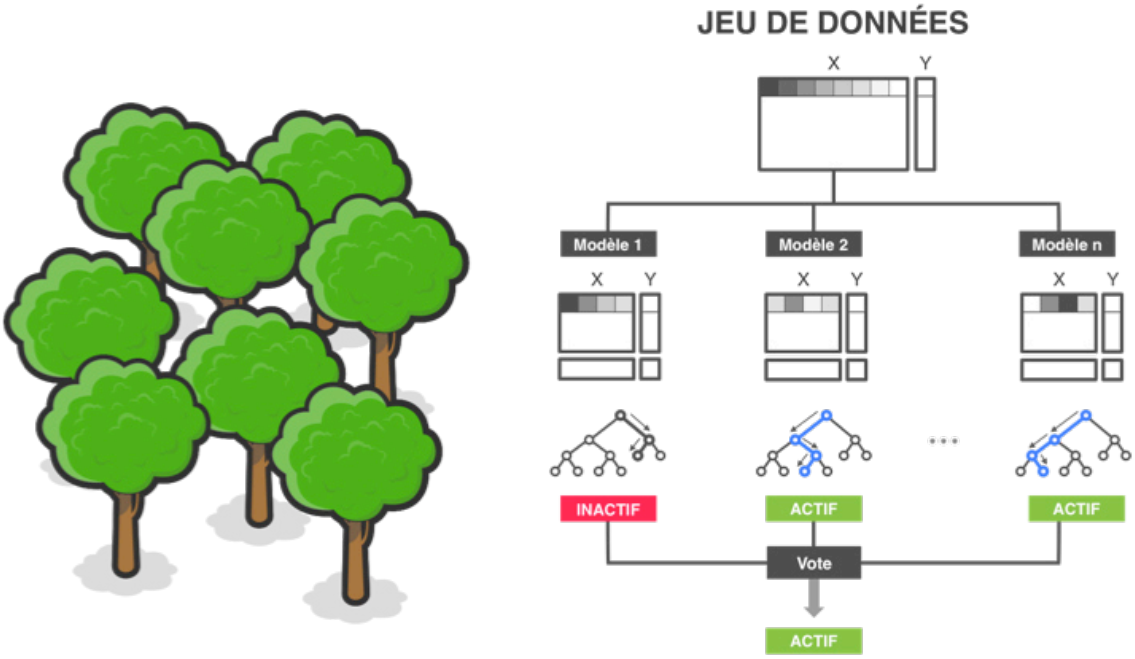


Figure 15 : Schéma de principe d'une forêt aléatoire.

En résumé, cette méthode est plus performante que les arbres de décision seuls. Cependant, comme toutes les méthodes de *bagging*, elle est peu interprétable par rapport à la méthode d'apprentissage de base. En effet, une forêt aléatoire est constituée de nombreux arbres de décision (données et descripteurs) qui sont construits sans élagage. De ce fait, l'interprétation manuelle du chemin parcouru par chaque arbre est impossible.

f) k-plus proches voisins (k-NN)

La méthode d'apprentissage *k*-NN consiste à attribuer une classe à un individu en fonction de ses *k* plus proches voisins dans l'espace multidimensionnel décrit par les descripteurs moléculaires. Cet algorithme détermine les *k* plus proches voisins pour une molécule d'intérêt à l'aide d'un calcul de distance (similarité). Les valeurs d'activité des plus proches voisins sont ensuite utilisées pour attribuer une valeur à la molécule étudiée. L'impact de chaque voisin sur la valeur prédite peut être pondérée à l'aide de la similarité observée entre les individus [141,181].

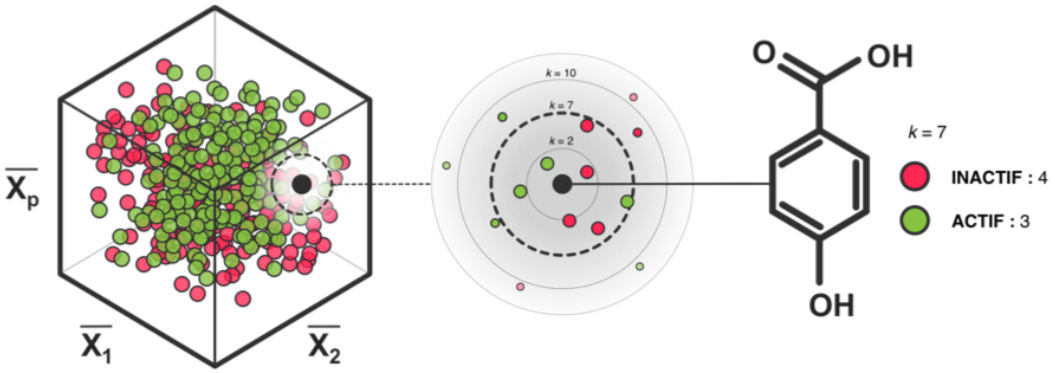


Figure 16 : Exemple du *k*-NN pour un nombre de voisins *k* = 7.

Pour s'assurer que tous les descripteurs soient pris en compte de manière équitable, il est indispensable de les normaliser afin que l'amplitude de certains d'entre eux ne biaise pas injustement d'autres descripteurs dans le calcul de la distance. D'autre part, la valeur du nombre de voisins *k* est sélectionnée en essayant plusieurs seuils, et en estimant à chaque itération les performances du modèle à l'aide d'une validation croisée (Ch1 3.2.4.3). Le nombre de plus proches voisins sélectionnés sera celui pour lequel le modèle disposera des meilleures performances. Les performances doivent cependant être estimées à l'aide de métriques adaptées (Ch1 3.2.4.2).

Dans le cas d'une classification, le *k*-NN identifie les $k \geq 1$ plus proches voisins de la molécule étudiée grâce à la détermination des distances calculées sur la base des descripteurs normalisés. La règle de la majorité est ensuite utilisée pour déterminer la classe de la molécule d'intérêt comme représenté par la Figure 16, où l'acide 4-hydroxybensoique sera considéré comme inactif pour un nombre de voisin $k = 7$. Il est possible d'utiliser une approche pondérée ou non pondérée, comme énoncé par l'Equation 3, afin de limiter l'impact des voisins les plus éloignés présents dans la sphère d'inclusion.

$$y(x) = \sum_{j=1}^{k} w_j y_{\sigma(j)}$$

Equation 3 : Règle de majorité utilisée par un *k*-NN pour une classification binaire [182].

$y(x)$ représente la valeur prédite d'une molécule ; w_j représente le poids du voisin j qui est inversement proportionnel à la distance ; $y_{\sigma(j)}$ est la classe attribuée au voisin j ; et k est le nombre maximal de plus proches voisins. w_j est égal à 1 lorsque le *k*-NN est non pondéré.

L'inconvénient majeur de cette méthode d'apprentissage réside dans le calcul des distances. Ainsi, la détermination des plus proches voisins pour tous les individus du jeu

de données nécessite de calculer une matrice de similarité carrée, qui requiert une puissance de calcul élevée et un temps de prédiction long lorsque le nombre de molécules et/ou le nombre de descripteurs augmentent. Des solutions ont donc été proposées pour améliorer l'efficacité de cette approche telles que la parallélisation sur GPU [182].

En résumé, le *k*-NN est une méthode simple à mettre en œuvre et efficace. Le degré d'interprétabilité de cet algorithme est basé sur la transparence des descripteurs et de la métrique de similarité. Ainsi, il est possible de retrouver les *k* voisins les plus proches d'une molécule et de les présenter à un utilisateur conjointement avec la prédiction.

g) Modèle de consensus

Il est possible que des modèles disposant de performances semblables fournissent des prédictions différentes. Ceci est explicable par le fait que chaque modèle est construit en utilisant un espace chimique précis, un jeu de descripteurs spécifique ou une méthode d'apprentissage qui peut être différente. Afin d'obtenir des modèles avec de meilleures performances ainsi que des prédictions plus robustes, il est possible de créer des modèles de consensus dont l'objectif est alors de combiner les modèles individuels [183–185]. Il existe plusieurs moyens de créer des modèles de consensus. Dans le cadre de cette thèse, nous avons utilisé la méthode du vote majoritaire dans le cas des modèles de classification et la méthode des valeurs moyennées dans le cas des modèles de régression. La mise en place d'un modèle de consensus nécessite néanmoins de pouvoir comparer les modèles en fonction de leur performance. Pour cela, une validation individuelle de chaque modèle est requise et sera présentée ci-après.

3.2.4. Validation

La validation est indispensable et présente différents enjeux pour justifier la qualité d'un modèle (Q)SAR. Dans le cadre de cette partie, nous verrons les phénomènes responsables de la non validité d'un modèle (Q)SAR, ainsi que les métriques de performances et les méthodes de validation utilisées pour appréhender et prévenir chacun de ces phénomènes.

3.2.4.1. Enjeux de la validation d'un modèle (Q)SAR

La validation est une étape importante qui permet de vérifier qu'un modèle est statistiquement valide et performant, c'est-à-dire qu'il est capable de prédire avec fiabilité l'activité biologique étudiée pour un ensemble de molécules. L'élaboration d'un modèle prédictif peut être perturbée par différents phénomènes comme le sous-apprentissage, le

sur-apprentissage ou encore la corrélation aléatoire. Comme nous l'avons énoncé précédemment, ces phénomènes peuvent être induits par différents facteurs, comme par exemple la taille du jeu de données, le nombre de descripteurs moléculaires, ou encore l'utilisation d'une méthode d'apprentissage particulière. Afin de pouvoir identifier les conditions les plus propices à l'obtention d'un modèle performant, il est intéressant de définir l'impact de chacun de ces facteurs sur les phénomènes perturbant la création des modèles de prédiction.

Toutes les méthodes d'apprentissage possèdent une erreur statistique qui dépend du biais et de la variance, comme illustré par la Figure 17. Le biais peut être défini comme une erreur induite par l'utilisation d'une méthode statistique incapable de saisir les tendances sous-jacentes du jeu de données. La variance peut être décrite comme une erreur induite par la sensibilité du modèle au bruit initialement présent dans le jeu de données [186]. Ces erreurs sont toutes deux affectées par la complexité du modèle, qui dépend de la méthode d'apprentissage plus ou moins sophistiquée, du nombre de descripteurs, ou encore du nombre de molécules (taille du jeu de données). Ainsi, un modèle simpliste est susceptible de trop généraliser les tendances sous-jacentes présentes dans le jeu de données, se traduisant par un biais élevé et une variance faible (sous-apprentissage). Un modèle trop complexe a généralement un faible biais et une variance trop élevée, se traduisant par une connaissance parfaite du jeu d'apprentissage (sur-apprentissage), mais une relation structure-activité pas assez généraliste pour être appliquée sur de nouvelles molécules inconnues du modèle.

Le sous-apprentissage (*under-fitting*) signifie que le modèle n'est pas assez sensible et par conséquent ne permet pas de prédire les différences d'activité observer par de faibles modifications structurales. Plusieurs causes peuvent être responsables de ce phénomène, comme la présence d'erreurs dans l'ensemble de données, l'utilisation d'une méthode d'apprentissage non adaptée ou encore une mauvaise sélection des descripteurs pertinents pour modéliser l'activité.

Le sur-apprentissage (*over-fitting*) traduit la sensibilité accrue du modèle pour les tendances sous-jacentes du jeu de données utilisé pour le créer (jeu d'apprentissage). Ceci peut être causé par un mauvais paramétrage de la méthode d'apprentissage, comme par exemple dans le cas d'une forêt aléatoire où le nombre d'arbres va être trop élevé permettant au modèle final d'appréhender correctement tous les cas de figure. Ce phénomène de sur-apprentissage peut également être causé par l'utilisation d'un nombre

de descripteurs nettement supérieur au nombre d'individus. Dans ce cas précis, le modèle va disposer d'informations trop conséquentes lui permettant de prédire convenablement l'ensemble des individus du jeu d'apprentissage, voire même les valeurs aberrantes.

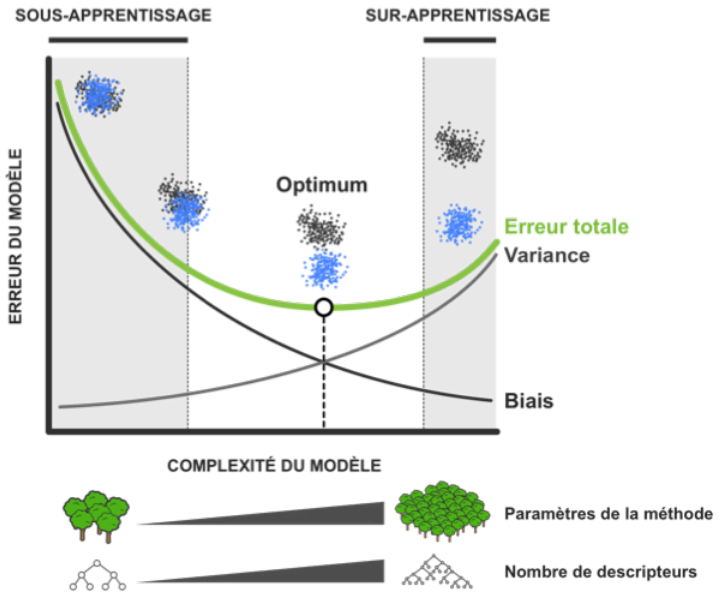


Figure 17 : Représentation de l'erreur d'un modèle (Q)SAR en fonction de sa complexité.

Les nuages de points bleu et gris représentent la séparation de deux classes dans le cadre d'un modèle de classification en fonction de la complexité du modèle.

Ce problème lié au nombre d'individus et au nombre de descripteurs peut être aussi la cause du phénomène de corrélation aléatoire (*chance correlation*) [187,188]. Comme observé par Topliss *et al.*, lorsque le nombre d'individus diminue, pour un nombre de descripteurs constant, la corrélation entre les descripteurs et l'activité est plus facilement observée. De la sorte, des corrélations fortuites vont avoir un effet considérable se traduisant par l'obtention d'un modèle avec de très bonnes performances, mais sans une réelle signification d'un point de vu biologique. On en conclut que le rapport entre le nombre d'individus et le nombre de descripteurs doit être adapté pour éviter ce phénomène. Ainsi, une règle empirique admise par la communauté exige que le nombre d'individus soit au moins trois fois supérieur au nombre de variables afin de prévenir les risque de corrélation aléatoire [188].

En résumé, l'objectif lors de la création d'un modèle est de diminuer le biais et d'augmenter la variance de telle sorte que l'erreur totale du modèle soit minimale (Figure 17). Ainsi, des critères de performance sont nécessaires afin d'estimer la qualité d'un

modèle. De plus, des méthodes de validation doivent être adoptées pour vérifier que le modèle n'est pas enclin aux phénomènes précédemment énoncés.

3.2.4.2. Critères d'évaluation des modèles

Plusieurs critères statistiques permettent de définir les performances d'un modèle. Ces métriques sont différentes dans le cas d'une régression ou d'une classification et il en existe une grande variété. Pour cette raison, nous ne présentons ci-dessous que les métriques utilisées dans le cadre de cette thèse.

a) Modèles de classification

Les performances d'un modèle de classification nous apportent les informations nécessaires permettant de juger la capacité du modèle à séparer deux classes de molécules. Elles sont estimées à partir des informations contenues dans une matrice de confusion (Figure 18) qui contient les comptes de vrais positifs (VP), de vrais négatifs (VN), de faux positifs (FP) et de faux négatifs (FN), déterminés en comparant les classes prédites par le modèle aux classes expérimentales du jeu de données.

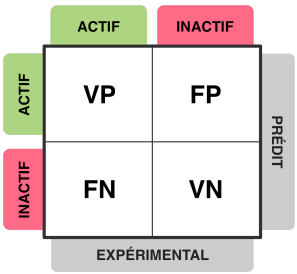


Figure 18 : Exemple de matrice de confusion.

Cette matrice permet d'avoir accès à plusieurs critères de performance. Les premiers critères sont la sensibilité, la spécificité et la justesse. La sensibilité (Equation 5) traduit la capacité du modèle à prédire correctement les molécules actives, tandis que la spécificité (Equation 6) représente la capacité du modèle à prédire correctement les molécules inactives. La justesse (Equation 4) est un paramètre qui rend compte de la capacité du modèle à classer correctement les molécules actives et inactives. Il est également intéressant de savoir à quel point le modèle se trompe pour la prédiction d'une classe spécifique. La précision est le critère statistique qui traduit la justesse du modèle pour une classe, comme par exemple l'Equation 7 qui correspond à la précision du modèle pour la classe active. Chacun de ces paramètres fournit des valeurs comprises entre 0 et 1, avec une valeur de 1 lorsque les performances du modèle sont excellentes.

Critère de performance	Equation
Equation 4 : Justesse (Acc)	$\frac{VP + VN}{VP + VN + FP + FN}$
Equation 5 : Sensibilité (Sens)	$\frac{VP}{VP + FN}$
Equation 6 : Spécificité (Spe)	$\frac{VN}{VN + FP}$
Equation 7 : Précision (Pre)	$\frac{VP}{VP + FP}$
Equation 8 : Coefficient de corrélation de Matthews (MCC)	$\frac{VP * VN - FP * FN}{\sqrt{(VP + FP)(VP + FN)(VN + FP)(VN + FN)}}$

Table 2 : Critères de performance d'une classification.

Dans le cadre d'une classification, un jeu de données peut comporter des biais, comme par exemple la présence de classes déséquilibrées. Dans ce cas, un modèle peut disposer de performances indiquant qu'il est juste, même s'il est peu précis sur la classe minoritaire. De plus, la taille du jeu de données va impacter l'espace chimique couvert par le modèle de classification. De la sorte, un modèle peut disposer d'une justesse élevée, mais ce dernier peut ne pas être pertinent dû au faible espace chimique qu'il couvre. Ainsi, la tendance actuelle est d'utiliser des métriques combinant plusieurs informations contenues dans le modèle. La première métrique est issue de la courbe ROC (*Receiver Operating Characteristics*), qui est une représentation visuelle couramment utilisée pour illustrer le succès et l'erreur d'un modèle de classification. Cette courbe consiste à représenter le taux de vrais positifs (sensibilité) en fonction du taux de faux positifs (1 - spécificité). A partir de cette courbe, il est possible de définir l'aire sous la courbe (AUC) qui incorpore les paramètres de sensibilité et de spécificité. Ainsi, plus l'AUC est élevée, plus le modèle de classification est précis. L'avantage de ce critère est qu'il permet de prendre en compte la taille du jeu de données employé pour la création du modèle et la précision du modèle pour chacune des classes. Cependant, ce critère est sensible à l'équilibre des classes. Le coefficient de corrélation de Matthews est une deuxième

métrique permettant de définir la qualité d'un modèle de classification dichotomique. Comme représenté par l'Equation 8, ce critère de performance prend en considération l'ensemble des informations transmises par la matrice de confusion. Il correspond à un coefficient de corrélation entre les classes prédites et expérimentales et il varie entre -1 et 1. Un coefficient de -1 indique un modèle totalement erroné qui se trompe dans tous les cas, tandis qu'un coefficient de 1 indique un modèle parfait qui ne se trompe jamais. Ce coefficient présente l'avantage de prend en compte la taille du jeu de données et il est également peu sensible au déséquilibre des classes. Cependant, l'inconvénient de ce critère de qualité est qu'il est plus restrictif que ceux énoncés précédemment.

b) Modèles de régression

Les performances d'un modèle de régression sont déterminées en comparant les valeurs prédites par le modèle aux valeurs expérimentales du jeu de données. Il existe plusieurs critères de qualité permettant de définir les performances d'une régression (Table 3).

Critère de performance	Equation
Equation 9 : Coefficient de détermination (R^2)	$1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y_i})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$
Equation 10 : Coefficient de détermination ajusté (R^2_{adj})	$\frac{\{(n-1) * R^2\} - p}{n - 1 - p}$
Equation 11 : Erreur moyenne absolue (MAE)	$\frac{\sum_{i=1}^{n}\lvert y_i - \hat{y_i}\rvert}{n}$
Equation 12 : Erreur quadratique moyenne (RMSE)	$\sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y_i})^2}{n}}$

Table 3 : Critères de performance d'une régression.

Avec $\hat{y_i}$ la valeur prédite de l'activité pour la molécule i ; y_i la valeur expérimentale de l'activité pour la molécule i ; $\bar{y}$ la moyenne des valeurs expérimentales d'activité ; n le nombre de molécules dans le jeu de données considéré ; p le nombre de descripteurs.

Le coefficient de détermination (Equation 9) et l'erreur quadratique moyenne (Equation 12) sont les deux critères les plus utilisés. Le R^2 représente la qualité de l'ajustement du modèle de régression. Il varie entre 0 et 1, avec un coefficient de 1 lorsque la corrélation entre les prédictions et les valeurs observées est maximale. Ce critère peut être complété à l'aide du coefficient de détermination ajusté (R^2_{adj}) qui prend en compte le nombre de

descripteurs (*p*) et le nombre de molécules (*n*) (Equation 10). Le RSME représente l'erreur interne du modèle. Plus sa valeur est faible, plus le modèle est performant. L'information transmise par le RMSE peut être analysée conjointement à l'erreur moyenne absolue (MAE) qui permet de refléter la dispersion de l'erreur dans le modèle (Equation 11) [189]. Le Q^2 est un autre critère qui est spécifique à la validation interne décrite ci-après.

c) Critère générique

Le score de Dixon [185] (nommé M_{score} dans ce projet de thèse) est un critère applicable aussi bien pour les modèles de régression que les modèles de classification. L'objectif de ce score est de déterminer le pouvoir prédictif du modèle et s'il présente les caractéristiques du sur-apprentissage. Pour cela, le score cherche à exprimer la différence entre les performances du modèle sur le jeu d'apprentissage et sur le jeu de test, comme représenté par l'Equation 13.

$$M_{score} = P_{test} * (1 - |P_{apprentissage} - P_{test}|)$$

Equation 13 : Score de Dixon.

Avec P_{test} la performance du modèle sur le jeu de test et $P_{apprentissage}$ la performance du modèle sur le jeu d'apprentissage.

Un score compris entre 0,6 et 1 indique que le modèle n'est pas enclin au sur-apprentissage et qu'il est hautement prédictif. Ce score est très utile pour comparer plusieurs modèles de prédictions afin de ne sélectionner que les plus pertinents. Le calcul de ce dernier nécessite cependant une étape de validation interne et une étape de validation externe.

3.2.4.3. Validation interne

La validation interne d'un modèle (Q)SAR se fait à partir du jeu d'apprentissage [190]. La première étape de cette validation consiste à déterminer la précision du modèle sur l'ensemble du jeu d'apprentissage. La deuxième étape est une validation croisée (*cross-validation*) qui permet d'appréhender la qualité et la robustesse du modèle et cela en simulant la situation où le modèle doit prédire un jeu de données sur lequel il n'a pas été entrainé. Durant cette étape de validation croisée, le jeu d'apprentissage va être divisé en deux sous-ensembles, à savoir un jeu d'étalonnage et un jeu de validation. Le jeu d'étalonnage permet de construire le modèle, tandis que le jeu de validation est utilisé pour vérifier si le modèle prédit correctement les molécules non utilisées dans le processus d'apprentissage.

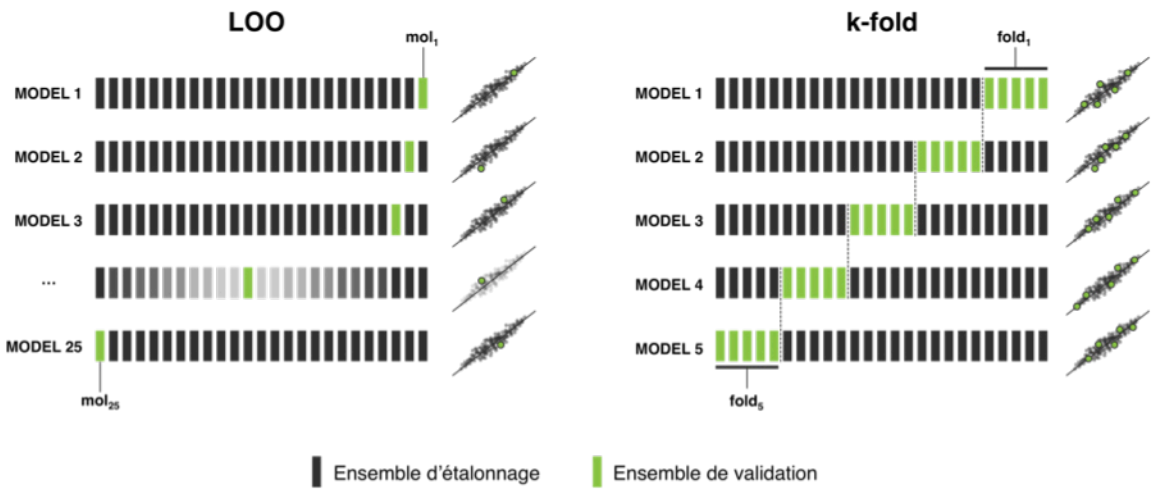


Figure 19 : Validation croisée LOO et LMO *k-fold* dans le cadre d'un modèle de régression et pour un jeu de données de 25 molécules.

La validation croisée peut être réalisée à l'aide de différentes techniques. La première technique se base sur l'omission individuelle de chaque molécule du jeu d'apprentissage, aussi nommée *Leave-One-Out* (LOO). Comme illustré par la Figure 19, cette procédure itérative va exclure temporairement une molécule du jeu d'apprentissage afin de constituer le jeu de validation et toutes les autres molécules (*n*-1) vont constituer le jeu d'étalonnage. Cette étape est répétée jusqu'à temps que chaque molécule ait constituée un jeu de validation. Cette technique est simple à mettre en œuvre, mais elle présente certains inconvénients. En effet, si nous prenons l'exemple d'un jeu d'apprentissage qui contient 25 molécules (Figure 19), la validation croisée va devoir définir 25 modèles temporaires, ce qui n'est pas efficace car le temps de validation augmente proportionnellement au nombre de molécules présentes dans le jeu d'apprentissage. D'autre part, l'utilisation du jeu d'étalonnage de taille *n*-1 permet d'obtenir des modèles temporaires très robustes, mais ceci n'est pas représentatif du jeu de données initial [191]. Par conséquent, la validation croisée LOO ne permet pas d'apprécier la pertinence du modèle sur un ensemble de données qui lui est inconnu. De plus, l'utilisation d'une seule molécule en tant que jeu de validation est très sensible à la présence de valeurs aberrantes.

Pour cette raison, la deuxième technique consiste à omettre plusieurs molécules du jeu d'apprentissage et elle est ainsi nommée *Leave-Many-Out* (LMO). Il existe plusieurs façons de mettre en œuvre une validation croisée LMO, cependant dans le cadre de cette thèse nous avons utilisé l'approche *k-fold*, qui consiste à diviser le jeu d'apprentissage en *k* sous-ensembles. Comme illustrée par la Figure 19, cette approche *k-fold* est itérative et va exclure à chaque étape un des *k* sous-ensembles pour l'utiliser en tant que jeu de

validation, puis étalonner le modèle sur les sous-ensembles non exclus (*k*-1). Comme dans le cas de la validation croisée LOO, cette étape est répétée jusqu'à temps que chaque sous-ensemble ait constitué un jeu de validation. Cette technique est simple à mettre en œuvre et permet de réduire considérablement le temps nécessaire pour valider le modèle. En effet, dans l'exemple de la Figure 19, le jeu d'apprentissage de 25 molécules est divisé en 5 sous-ensembles et ne nécessite que 5 modèles temporaires pour effectuer la validation croisée. Cette technique est donc plus rapide que la validation croisée LOO et permet de valider le modèle à l'aide de sous-ensembles représentatifs du jeu de données initial. Plusieurs optimisations de la validation LMO ont été proposées par la communauté, comme par exemple l'utilisation d'une stratification des *k* sous-ensembles ou encore une répétition de la validation croisée. La stratification consiste à s'assurer que tous les sous-ensembles soient représentatifs du jeu d'apprentissage, c'est-à-dire que la répartition de l'activité (Y) est similaire à celle observée dans le jeu d'apprentissage [192]. La répétition consiste à répéter plusieurs fois l'étape de validation croisée, ce qui permet de réduire la variabilité du modèle [193]. Dans le cadre de la validation croisée, le paramètre Q^2 peut être déterminé suivant l'Equation 14.

$$Q^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

Equation 14 : Paramètre de validation croisée Q^2.

Avec $\hat{y}_i$ la valeur prédite de l'activité pour la molécule i à chaque itération (jeu de validation) ; y_i la valeur expérimentale de l'activité pour la molécule i ; $\bar{y}$ la moyenne des valeurs expérimentales d'activité.

On tient alors le raisonnement que si des modèles réalisés avec un plus petit nombre d'individus sont corrects il est probable qu'un modèle réalisé avec l'ensemble des individus sera aussi bon, voire meilleur. Cependant, la différence entre le R^2 et le Q^2 ne doit pas excéder 0,3. Un R^2 nettement supérieur au Q^2 signifie que le modèle est très prédictif sur le jeu d'apprentissage et peu robuste pour la prédiction de molécules externes. Par conséquent, la validation croisée permet d'identifier les phénomènes de sur-apprentissage du modèle.

En résumé, la validation croisée est idéale pour juger la qualité et la robustesse d'un modèle dès les premières étapes de sa création. Cependant, l'inconvénient majeur de la validation interne est le manque de prévisibilité du modèle lorsqu'il est appliqué à un nouveau jeu de données.

3.2.4.4. Randomisation de l'activité

La randomisation de l'activité, aussi nommée *Y-scrambling*, consiste à mélanger de façon aléatoire les valeurs d'activité pour toutes les molécules du jeu de données. Les valeurs expérimentales mélangées et les valeurs prédites sont ensuite comparées. Si les performances du modèle sont médiocres cela signifie que le modèle à une réelle signification biologique et qu'il n'est pas enclin au phénomène de corrélation aléatoire (Ch1 3.2.4.1).

3.2.4.5. Validation externe

Comme énoncé précédemment, la validation interne ne prend en compte que les molécules appartenant au même ensemble de données (jeu d'apprentissage). Par conséquent, on ne peut pas juger de la capacité du modèle développé à prédire des molécules inconnues [194]. L'objectif de la validation externe est d'estimer le pouvoir prédictif du modèle sur un ensemble de données non inclus dans le jeu d'apprentissage. Ainsi, la prédiction d'un jeu de test permet d'estimer l'applicabilité du modèle sur de nouveaux ensembles de molécules. Dans de nombreux cas, lorsqu'un jeu de test externe au modèle n'est pas disponible, le jeu de données initial est au préalable divisé en deux sous-ensembles, à savoir le jeu d'apprentissage et le jeu de test. Le jeu d'apprentissage sera utilisé pour la création du modèle et l'étape de validation interne, tandis que le jeu de test sera employé pour l'étape de validation externe à la fin du processus. La découpe de l'ensemble de données en jeu d'apprentissage et jeu de test doit répondre à plusieurs exigences.

3.2.4.6. Sélection des jeux d'apprentissage et de test

Le choix du jeu d'apprentissage et du jeu de test peut être réalisé de différentes manières. La seule exigence est que les deux sous-ensembles soient représentatifs de l'espace chimique décrit par le jeu de données initial [195–198]. Idéalement, la division du jeu de données doit être effectuée de telle sorte qu'un individu du jeu de test se trouve à proximité d'un individu du jeu d'apprentissage dans l'espace descriptif du modèle. Les approches suivantes sont régulièrement utilisées pour sélectionner ces sous-ensembles :

i) **Sélection aléatoire :** la division du jeu de données se fait par une simple découpe aléatoire.

ii) **Sélection basée sur l'activité (Y) :** les valeurs d'activité biologique sont utilisées pour former des groupes individus. La sélection se fait ensuite de façon aléatoire dans chacun des groupes.

iii) **Sélection basée sur les descripteurs (X) :** les valeurs de descripteurs moléculaires sont utilisées pour calculer la similarité des molécules (distance). Des méthodes de regroupements, comme par exemple les *k*-moyennes, sont utilisées pour définir des groupes de molécules en fonction de leur similarité. La sélection se fait ensuite de façon aléatoire dans chacun des groupes définis.

3.2.4.7. Domaine d'applicabilité

Le domaine d'applicabilité (DA) est une région théorique de l'espace chimique couvert par un modèle (Q)SAR en termes de descripteurs et de réponse biologique. Le besoin de définir un domaine d'applicabilité est considéré comme une étape indispensable pour justifier la validité d'un modèle (Q)SAR. Cependant, sa mise en place reste de nos jours un sujet de controverse et de recherche au sein de la communauté scientifique. Comme énoncé précédemment, un modèle (Q)SAR établit une relation entre la structure chimique des composés décrite par des descripteurs moléculaires et la propriété biologique associée. En théorie, un modèle peut être utilisé pour prédire les propriétés de nouvelles molécules lorsque les informations structurales sont disponibles. Cependant, lors de la prédiction d'un nouveau jeu de données, une question de taille se pose, à savoir : « La molécule nouvellement prédite est-elle présente dans l'espace chimique connu du modèle ? ».

Le DA a pour objectif de fixer les contraintes du jeu d'apprentissage utilisé pour définir le modèle. Ces contraintes peuvent-êtres employées pour déterminer si une nouvelle molécule est considérée comme étant en dedans ou en dehors du DA, et de manière plus concrète s'il est possible de faire une interpolation ou une extrapolation des valeurs prédites. Seules les prédictions des molécules présentes dans le DA vont être considérées (interpolation). Les prédictions des molécules en dehors du DA sont *a priori* incorrectes, car le comportement du modèle est inconnu. En résumé, le DA a pour but de quantifier la déviation d'un modèle (Q)SAR, qui est par définition chimiquement limité, et de borner l'utilisation du modèle en fonction de la similarité entre les molécules du jeu d'apprentissage et les molécules nouvellement prédites.

Afin de répondre correctement à cette problématique, il est indispensable de prendre en compte les points clefs nécessaires à la mise en place du DA. i) Le premier consiste à identifier des sous espaces chimiques denses pour lesquels le modèle est censé être fiable. ii) Le deuxième est de déterminer le degré de généralisation du modèle, c'est-à-dire, de fixer un ou plusieurs seuils permettant de définir les limites du DA. Cette étape est cruciale, car le choix d'un seuil trop restrictif induirait un modèle extrêmement fiable, mais ne prenant en compte que des structures chimiques très similaires à celles présentes dans le jeu d'apprentissage. iii) Le DA doit être, dans la mesure du possible, caractérisé.

De nombreux travaux ont été menés afin de proposer de nouvelles approches ou d'améliorer les techniques existantes pour optimiser la détermination des domaines d'applicabilité. L'ensemble des méthodes couramment rencontrées et décrites dans la littérature ne doivent pas être considérées comme définitives et restent un point de départ pour le développement de nouvelles approches toujours plus fiables afin d'améliorer la prise de décision suite à la prédiction (Q)SAR. Il existe donc différentes approches permettant de définir un DA (Figure 20), plus ou moins complexes à mettre en œuvre, dont les plus populaires sont présentées et discutées ci-dessous.

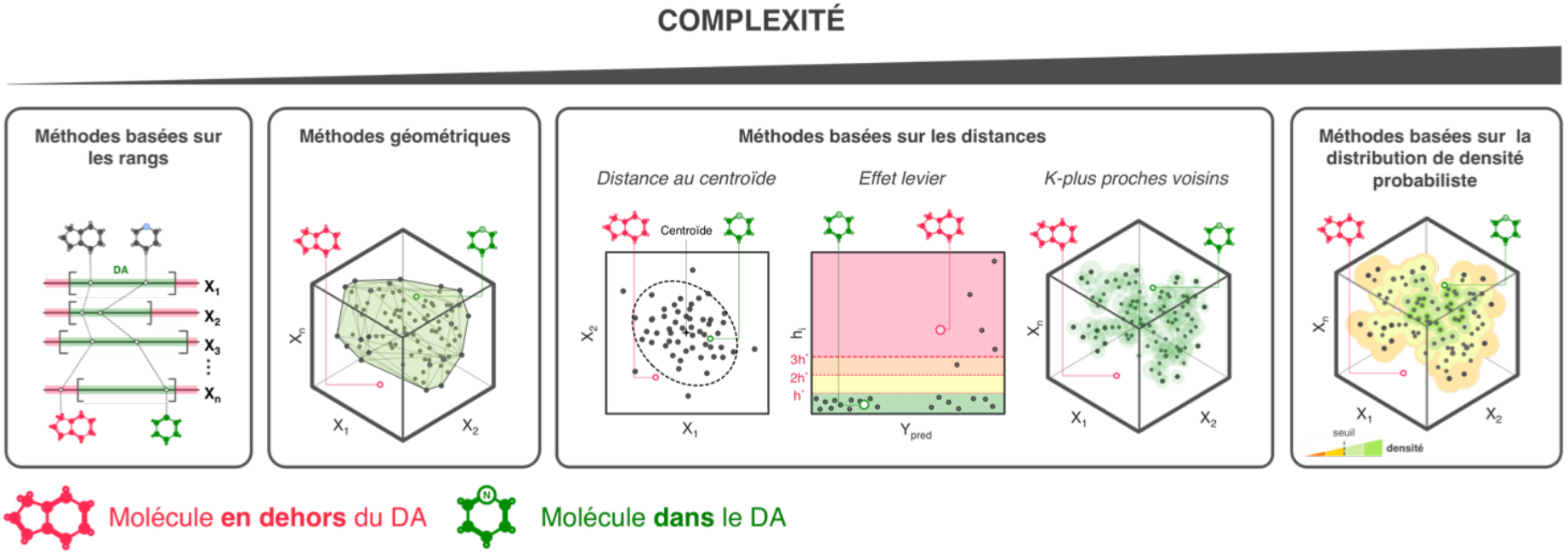


Figure 20 : Représentation schématique des méthodes utilisées pour définir un DA.

Les approches permettant d'élaborer un DA sont ordonnées en fonction de leur principe (rang, géométrie, distance, densité) et de la complexité de la méthode utilisée. Les descripteurs moléculaires sont représentés par la notation X_n.

a) Méthodes basées sur les rangs

Ces méthodes se basent sur les gammes de valeurs des descripteurs utilisés par le modèle. Elles consistent à définir des seuils maximal et minimal déterminés à partir du jeu d'apprentissage pour chaque descripteur du modèle. Ainsi, toute molécule ayant des valeurs de descripteurs comprises entre ces seuils sera considérée comme faisant partie du DA. Les boîtes de délimitation, *Bounding box* ou *PCA Bounding box,* sont les approches les plus connues de ce type. Ces approches simplistes et rapides à mettre en œuvre possèdent cependant des inconvénients majeurs. En effet, elles ne prennent pas en compte la colinéarité des descripteurs (*Bounding Box* uniquement) et les régions dépeuplées du domaine d'applicabilité [199–201]. En résumé, les seuils utilisés sont peu restrictifs et susceptibles d'accepter dans le domaine d'interpolation des molécules très diverses induisant un risque important de prioriser des molécules mal prédites.

b) Méthodes géométriques

Ces méthodes caractérisent le DA en définissant un espace convexe contenant l'ensemble des individus du jeu d'apprentissage. L'enveloppe convexe est générée à partir de l'espace à *n*-dimensions correspondant aux *n* descripteurs utilisés par le modèle [202]. Une nouvelle molécule peut être considérée comme étant en dedans (interpolation) ou en dehors (extrapolation) de cette enveloppe convexe. L'obtention d'une enveloppe convexe représentative est affectée par l'augmentation du nombre de descripteurs. Par ailleurs, le DA prend en compte l'ensemble des individus du jeu d'apprentissage, ce qui signifie que cette méthode ne peut en aucun cas identifier les éventuelles régions vides et présente alors les mêmes limitations que les méthodes basées sur les boites de délimitations.

c) Méthodes basées sur les distances

Ces approches définissent l'espace d'interpolation en calculant un ou plusieurs seuils basés sur les distances des molécules présentent dans le jeu d'apprentissage. Elles sont les approches les plus communément rencontrées dans la littérature. Plusieurs méthodes peuvent être élaborées et sont différenciables selon le type de distance utilisée :

- Distances au centroïde :

Plusieurs distances sont utilisées dans le cadre de ces approches dont les plus populaires sont : la distance Euclidienne, la distance de Mahalanobis, la distance City Block ou Manhattan, ou encore la distance de Tanimoto aussi nommée distance de Jaccard. Un seuil correspondant à la distance limite au centroïde du jeu d'apprentissage va permettre de définir si une nouvelle molécule est assimilable au domaine d'interpolation. Ces approches sont sensibles aux dimensionnalités élevées du modèle, car la présence de descripteurs non pertinents ou de descripteurs redondants (corrélés) peut induire des distances calculées non significatives [203,204]. Cependant, la distance de Mahalanobis, basée sur le calcul d'une matrice de covariance, présente l'avantage de prendre en compte l'influence des éventuels descripteurs corrélés. D'autre part, l'absence de règles strictes dans la littérature pour définir les seuils peut conduire à des résultats ambigus.

- Effet de levier ou *leverage* [191] :

La méthode du levier, initialement prévue pour détecter les points d'influence (*outliers* extrêmes) dans les modèles de régression, est considérée comme la méthode de référence par l'OCDE pour définir un DA [205]. L'effet de levier mesure l'influence h_i de chaque molécule i sur les estimations obtenues par le modèle de régression. L'effet levier d'une molécule i dans l'espace des descripteurs du modèle est défini selon l'équation :

$$h_i = x_i^T (X^T X)^{-1} x_i \ \ (i = 1, \dots, n)$$

Equation 15 : Détermination de l'influence du h_i.

Lorsque x_i représente un vecteur contenant les valeurs de descripteurs pour une molécule, X est la matrice de dimension $n \times k$ (k descripteurs et n molécules) et T la matrice transposée. Un seuil limite h^* est ensuite définit comme égal à 3 fois la moyenne des h_i obtenus pour les molécules du jeu d'apprentissage. Une nouvelle molécule disposant d'une valeur h_i supérieure à h^* sera considérée comme en dehors du DA. Cette méthode purement statistique ne permet en aucun cas de répondre à la caractérisation du domaine d'interpolation et, via le calcul du h_i , s'éloigne des informations structurales comprises dans le DA. De plus, la comparaison d'une molécule en fonction du seuil global h^* implique que cette méthode est peu sensible et comme dans le cas des méthodes précédentes ne permet pas de considérer les régions de faible densité du DA. D'autre part, cette approche n'est applicable qu'aux modèles de régression et n'est donc pas ou peu utilisable pour les modèles de classification (non

universelle). Ces observations sont applicables aux méthodes dites de distance au modèle ou *DModX* [206], initialement prévues pour détecter les points aberrants (*outliers* modérés) dans les modèles de régression PLS.

- K-plus proches voisins :

La méthode des k-plus proches voisins définit le DA en évaluant la distance/similarité entre les molécules du jeu d'apprentissage et les molécules du jeu de test, afin de déterminer un ou plusieurs seuils d'acceptabilité [207]. Une nouvelle molécule sera considérée comme prédite de façon fiable si la distance avec ses plus proches voisins du jeu d'apprentissage est inférieure au(x) seuil(s) établi(s). Cette méthode exclusivement basée sur les distances inter-individus est fortement impactée par les problèmes de dimensionnalité énoncés pour les approches de distance au centroïde. D'autre part, la sélection du seuil pour établir la règle d'acceptabilité, n'est pas une notion strictement définie dans la littérature. En effet, plusieurs travaux proposent d'incorporer à ce seuil des notions de densité locale ou de fiabilité locale, afin d'obtenir des limites uniques pour chaque individu du jeu d'apprentissage [208]. Par conséquent, contrairement à toutes les méthodes précédemment énoncées, cette approche permet d'atteindre une granulométrie plus fine du DA et permet de prendre en considération la notion de densité. Ceci rend possible l'identification des sous-espaces chimiques denses et la caractérisation de l'espace d'interpolation en fonction des plus proches voisins observés pour une molécule nouvellement prédite. De plus, cette approche simple et universelle peut-être appliquée à des modèles de régression mais également de classification. Les travaux de Roberto Todescini ont très largement contribués à la valorisation de cette méthode basée sur l'utilisation des approches k-NN afin de définir un DA [209–211].

d) Méthodes basées sur la distribution de densité probabiliste [201]

De nos jours cette méthode est considérée comme la plus performante pour estimer le DA d'un modèle (Q)SAR, car elle permet de mettre en lumière des régions de haute densité et des régions dépeuplées du domaine d'interpolation. Le GTM développé par Gaspar, Baskin, Marcou, Horvath et Varnek est le meilleur exemple des approches basées sur la densité de probabilité [212–214].

e) Autres méthodes

D'autres méthodes sont rencontrées dans la littérature comme les approches basées sur les arbres de décision ou les forêts aléatoires, les approches par étapes ou

stepwise, les approches basées sur des regroupements intelligents de type *k-means* ou encore les approches basées sur les noyaux (*kernel*) propres à certaines méthodes d'apprentissage. Ces méthodes restent des approches exotiques et pour cette raison ne seront pas décrites.

4. Modèles de prédiction ADME-Tox dans le processus de découverte de nouveaux médicaments : vue d'ensemble et limitations actuelles

En 2017, des scientifiques de plusieurs entreprises pharmaceutiques (Eli Lilly, Roche, Pfizer, AstraZeneca, etc.) ont conjointement publié un article qui présente l'implication des approches de prédiction *in silico* ADME-Tox dans le processus de découverte de nouveaux principes actifs [215]. Cet article représente selon nous une vue d'ensemble pertinente et récente sur l'importance des modèles (Q)SAR pour la conception de nouveaux médicaments. Les auteurs de cet article expliquent que les approches *in silico* n'ont pas pour but de remplacer les tests expérimentaux, mais sont utilisées pour accélérer la prise de décision lors des projets de recherche et développement. Ainsi, des outils de prédiction ADME-Tox peuvent guider les chimistes afin d'orienter les prochaines étapes de conception, et ceci dans le but d'accélérer le processus d'optimisation des *leads* tout en réduisant le nombre de synthèses nécessaires à l'identification de candidat médicaments. De plus, l'utilisation de ces modèles offre de nouvelle opportunité pour évaluer le profil ADME-Tox des candidats potentiels, et ceci tout en minimisant les activités éthiquement contestées telles que l'expérimentation animale ou humaine. Les auteurs de cet article montrent à l'aide de plusieurs exemples que l'utilisation de ces approches virtuelles est aujourd'hui incontournable dans l'industrie pharmaceutique et représente une réelle alternative pour réduire les coûts et le temps nécessaires à la découverte de nouvelles molécules bioactives.

Par conséquent, il n'est pas étonnant de voir que le nombre d'outils dédiés à la prédiction des profils ADME-Tox est en constante évolution depuis plusieurs années (ANNEXE A). Cependant, 73% des experts dans ce domaine jugent que les prédictions de ces outils sont certes utiles mais pourraient être améliorées [215]. Bien que ce domaine scientifique soit mature, l'innovation peut encore contribuer à l'amélioration des outils déjà existants. Plusieurs axes d'amélioration sont énoncés comme étant critiques pour l'incorporation

plus aisée de ces approches dans le processus de découverte de nouveaux médicaments.

Le premier axe concerne les jeux de données utilisés pour la création des modèles de prédiction. La création de jeux de données de grande taille et de haute qualité nécessite des efforts considérables afin de vérifier l'ensemble des informations. En effet, les données *in vitro* et *in vivo* extraites de la littérature sont déterminées à l'aide de tests expérimentaux très divers qui disposent d'une variabilité importante. Ainsi, un compromis est donc souvent fait entre la quantité et la qualité des jeux de données afin de produire des modèles fiables et robustes dans la mesure du possible. Par conséquent, parmi l'ensemble des outils actuellement proposés, il n'est pas rare de voir que les modèles sont construits sur des jeux de données de taille restreinte (100 à 300 molécules) et ne couvrent donc qu'une infime partie de l'espace chimique (ANNEXE A).

Le deuxième axe correspond aux méthodes d'apprentissage et aux descripteurs employés pour créer les modèles de prédiction. Le choix des descripteurs moléculaires et de l'algorithme dépend des modélisateurs, de l'expertise interne au laboratoire, ou encore de l'activité modélisée. Les seules exigences selon les principes de l'OCDE sont que i) les descripteurs moléculaires soient explicatifs, ii) que le nombre de descripteurs moléculaires utilisés dans le modèle soit le plus faible possible, et iii) que le modèle possède des performances suffisantes pour être considéré comme prédictif et utile. Cependant, comme énoncé par Lombardo *et al.*, il est observé que des modèles de classification utilisant des descripteurs peu explicatifs donnent souvent des prédictions plus fiables pour modéliser les propriétés ADME-Tox sur des jeux de données couvrant un vaste espace chimique [215]. De plus, il est suggéré que l'utilisation de modèles de consensus, qui combinent plusieurs modèles, représentent un moyen sûr d'estimer la pertinence d'une prédiction et ainsi de réduire les risques de se tromper. Par conséquent, les modèles ADME-Tox actuels peuvent être considérés comme des "boites noires" de meilleure précision mais dont l'interprétation est compliquée. Ainsi, l'interprétation des résultats issus de ces outils dépend largement de la perception des scientifiques au sein des équipes de recherche, qui dans la majorité des cas jugent que les résultats manquent de clarté sur la fiabilité des prédictions ainsi que sur les limites du modèle [215].

Le troisième axe concerne les mesures de confiance et le domaine d'applicabilité des modèles, qui sont des éléments indispensables pour juger l'utilité d'une prédiction. Les mesures de confiance ont pour but d'estimer la précision d'une valeur prédite. Elles

peuvent être représentées sous la forme d'une probabilité d'activité pour un modèle de classification, ou encore une erreur standard sur la prédiction pour un modèle de régression. Le domaine d'applicabilité a pour but de vérifier que la molécule prédite est présente dans l'espace chimique connu du modèle. L'information qu'il transmet à l'utilisateur, doit permettre de savoir si l'utilisation de la prédiction fait l'objet d'une interpolation ou d'une extrapolation. L'objectif est de transmettre aux chercheurs l'ensemble des informations leurs permettant de juger la capacité du modèle à prédire correctement leurs molécules. Bien que ces informations soient d'une importance cruciale, la plupart des outils ADME-Tox disponibles ne les transmettent pas lors de la prédiction de nouvelles séries moléculaires.

Le quatrième axe est l'interprétation des résultats. Comme énoncé par Lombardo *et al.*, l'interprétation des modèles "boites noires" peut s'avérer complexe. Ainsi, des efforts doivent être entrepris pour créer des approches complémentaires axées sur l'interprétation et la généralisation des idées transmises par les modèles. Afin de répondre à cette problématique, l'utilisation de carte d'activité permet de représenter l'activité modélisée sur la structure chimique des molécules étudiées et cela en fonction de leurs motifs structuraux. Ceci donne une information visuelle aux chimistes leur permettant d'identifier rapidement les motifs favorisant ou défavorisant la propriété modélisée. D'autres approches pourraient constituer un cinquième axe intitulé interprétation et valorisation des approches (Q)SAR. Le but de ces approches est d'utiliser les prédictions (Q)SAR conjointement aux *Match Molecular Pairs* (MMP) afin de proposer des analogues structuraux des molécules imaginées et synthétisées par un chimiste. Une MMP est une paire de composés qui possèdent des différences structurales se traduisant par une différence en terme d'activité. Cette modification structurale peut être représentée par une transformation chimique applicable à d'autres molécules possédant la même sous-structure initiale. Il est alors possible de trouver des transformations qui contribuent favorablement et défavorablement à l'activité modélisée. L'objectif de ces approches est d'enrichir localement l'espace chimique afin d'en extraire des règles génériques permettant de proposer aux chimistes médicinaux une table de RSA avec les transformations potentiellement intéressantes dans le cadre d'un projet spécifique.

5. Conclusion

Les acteurs travaillant à l'élaboration de nouveaux médicaments doivent faire face à des défis de taille durant toutes les étapes de découverte d'une molécule bioactive.

L'automatisation des tâches répétitives a considérablement accru la productivité scientifique lors des différentes étapes de ce processus. La CADD a sans aucun doute grandement contribué à l'automatisation de la conception de nouveaux principes actifs, et ceci du début des projets de recherche jusqu'à la sélection des candidats médicaments. De nos jours, les autorités réclament plus de tests expérimentaux permettant de prouver l'innocuité, l'efficacité ou encore le bon profil médicamenteux d'un candidat médicament afin d'atteindre l'AMM. Le profil ADME-Tox des candidats médicaments est donc une composante indissociable de la réussite d'un projet de recherche. Cependant, les tests permettant de définir ces propriétés sont coûteux, disposent d'un faible débit et sont effectués tardivement dans le processus de recherche et de développement. Afin de réduire l'impact des profils ADME-Tox défavorables sur la réussite d'un projet, plusieurs laboratoires ont mis en œuvre des stratégies basées sur la modélisation QSAR, afin d'identifier et filtrer les composés les plus prometteurs dès l'étape d'optimisation des *leads*. L'incorporation de cette technologie pour la détermination précoce des propriétés ADME-Tox a été un franc succès, comme en atteste l'exemple proposé par AstraZeneca. Cependant, la mise en œuvre des approches (Q)SAR pour cette utilisation comporte encore plusieurs verrous, dont le premier concerne les données expérimentales ADME-Tox.

Chapitre 2 : Construction des jeux de données ADME-Tox pour l'élaboration de modèles (Q)SAR

Lors de cette thèse nous avons cherché à valoriser les sources de données publiques, dans le but de proposer des modèles de prédiction créés sur la base de données libres et accessibles à tous. Notre objectif a été de rechercher et collecter des données expérimentales pour plusieurs propriétés ADME-Tox, afin de les préparer pour l'élaboration de modèles (Q)SAR. Cette recherche de données nous a conduit à identifier les sources de données potentielles, les extraire et définir un protocole de préparation et de standardisation des informations à notre disposition.

1. Sources de données publiques

Il existe une grande diversité de sources de données publiques qui peuvent être sous la forme de jeux de données publiés spécifiques à une propriété, ou encore sous la forme d'une base de données contenant des points de mesures pour une multitude d'activités biologiques. Chacune de ces sources possède des avantages et des inconvénients qui lui est propre.

1.1. Jeux de données publiques

Les jeux de données définis pour des propriétés spécifiques présentent des avantages par rapport à l'extraction de bases de données. En effet, les jeux de données sont susceptibles d'avoir déjà été utilisés pour la construction de modèles (Q)SAR et par conséquent, des efforts d'évaluation de la qualité et de vérification des données peuvent déjà avoir été entrepris [216]. Si les jeux de données mis à disposition ont été révisés et si la procédure d'élaboration de l'ensemble de données à été publiée, un certain nombre d'obstacles à la création d'un jeu de données fiable et homogène a été franchi. Toutefois, il n'est pas rare de constater qu'un ensemble de données contient des mesures ADME-Tox provenant de différentes sources, chacune avec son propre niveau d'informations sur le protocole expérimental suivi. Les jeux de données les plus souvent rencontrés dans la littérature possèdent un minimum d'information sur les protocoles expérimentaux, mais également un manque d'information au sujet des références à partir desquelles les valeurs expérimentales ont été extraites. Cette absence d'information rend compliquée la vérification des mesures expérimentales et a pour conséquence une réduction de la qualité globale du jeu de données. Il est possible de trouver des informations plus riches à partir d'ensembles de données plus petits. Cependant, les modèles (Q)SAR générés à

partir de ces petits jeux de données couvriront un espace chimique plus restreint. Une autre possibilité consiste à combiner plusieurs sources en vérifiant la qualité et l'homogénéité des mesures afin de constituer un vaste jeu de données homogène. Par ailleurs, les jeux de données publiés peuvent disposer de différents formats (csv, xls, pdf, sdf, etc.) ce qui peut rendre compliqué leur extraction. De plus, l'information sur la structure chimique des composés n'est pas toujours explicite. Dans certains cas, la structure moléculaire est partagée de façon non ambiguë sous la forme d'une information structurale ou d'un identifiant unique du composé, comme par exemple un SMILES ou un numéro CAS. Ceci peut être directement exploité pour la création d'un jeu de données. Cependant, il arrive parfois que l'information sur la structure soit renseignée sous la forme d'un nom générique, ce qui nécessite une étape intermédiaire de recherche et de vérification de la structure moléculaire.

1.2. Bases de données publiques

Une base de données contient des informations multiples et structurées au sujet de la structure chimique des composés, de leurs propriétés biologiques, mais également dans certains cas des informations sur les protocoles expérimentaux qui ont permis d'effectuer la mesure. Les bases de données présentent l'avantage de proposer des informations correctement formatées sous la forme de table et contenant une information distincte par colonne (aussi nommée champ). Plus le niveau d'information est élevé, plus il est aisé de sélectionner les mesures afin de créer un jeu de données *a priori* homogène. L'ensemble des informations stockées dans les bases de données provient d'une multitude de sources extraites manuellement par une intervention humaine, ou automatiquement à l'aide de robots. De ce fait, le nombre de mesures est plus conséquent, mais la présence d'erreurs est plus fréquente et l'homogénéité des données est moins importante. Ainsi, l'utilisation de ces mesures pour l'élaboration d'un modèle (Q)SAR nécessite au préalable des étapes importantes de nettoyage et de vérification des données.

2. Identification des sources de données ADME-Tox

Identifier de nouvelles sources est aujourd'hui fondamental dans une approche de centralisation des données. L'objectif est alors de collecter, stocker, uniformiser et exploiter les données provenant de sources multiples afin de les valoriser en enrichissant les données internes au laboratoire. L'intérêt est donc de transformer ces données hétérogènes en valeurs exploitables. Cette création de valeurs est tout d'abord réalisée par l'identification des sources qui apportent des données pouvant améliorer, faire croître, ou encore réinventer de vastes jeux de données homogènes pour la création de modèles (Q)SAR. Cependant, la rareté de ces données ADME-Tox rend leur extraction complexe. Ainsi, les sources proposant ce type de données ne sont pas clairement identifiées de nos jours. Afin de répondre à cette problématique d'identification des sources ADME-Tox, nous avons proposé en 2017 dans la revue *Molecular Informatics* une étude ayant pour but de référencer plusieurs bases de données pouvant être exploitées dans le cadre des approches (Q)SAR.

Comprehensive Network Map of ADME-Tox Databases

Baptiste Canault,[a] Stéphane Bourg,[a] Philippe Vayer,[b] and Pascal Bonnet*[a]

[a] B. Canault, S. Bourg, P. Bonnet, Institut de Chimie Organique et Analytique (ICOA), Université d'Orléans et CNRS, UMR7311, BP 6759, 45067 Orléans, France E-mail: pascal.bonnet@univ-orleans.fr

[b] P. Vayer, Technologie Servier, 25-27 rue Eugène Vignat, BP 11749, 45007 Orléans cedex 1 (France)

Supporting information for this article is available on the WWW under https://doi.org/10.1002/minf.201700029

Abstract:

In the last decade, many statistical-based approaches have been developed to improve poor pharmacokinetics (PK) and to reduce toxicity of lead compounds, which are one of the main causes of high failure rate in drug development. Predictive QSAR models are not always very efficient due to the low number of available biological data and the differences in the experimental protocols. Fortunately, the number of available databases continues to grow every year. However, it remains a challenge to determine the source and the quality of the original data. The main goal is to identify the relevant databases required to generate the most robust predictive models. In this study, an interactive network of databases was proposed to easily find online data sources related to ADME-Tox parameters data. In this map, relevant information regarding scope of application, data availability and data redundancy can be obtained for each data source. To illustrate the usage of data mining from the network, a dataset on plasma protein binding is selected based on various sources such as DrugBank, PubChem and ChEMBL databases. A total of 2,606 unique molecules with experimental values of PPB were extracted and can constitute a consistent dataset for QSAR modeling.

Keywords: ADME-Tox · Network · Database · PPB

1 Introduction

The pharmaceutical industry is facing important challenges through every stage of the drug discovery process. Nowadays, computer-assisted drug design plays a key role in decision-making from the start of R&D projects until selection of clinical candidate.[1] A drug needs to be safe, effective, and have appropriate pharmacokinetic properties (PK). Recent studies indicate that inappropriate PK properties for drug candidates are one of

the main causes for late-stage failures especially in phase I clinical trials.[2] The prediction of absorption, distribution, metabolism, excretion (ADME) and toxicity (Tox) properties is a research field very investigated early in lead optimization. Accordingly, prediction models were developed for human PK to predict a wide range of chemical structures and to guide medicinal chemists in both design and optimization of improved chemical leads such as quantitative structure-activity/property relationship (QSAR or QSPR).[3,4] These computational models are based on experimental data and seek to cover a widest possible chemical space with a reliable prediction. The quantity and the quality of available data used to train models play a crucial role in the coverage of chemical space and performance of the model.[5,6] Nevertheless, some homogeneous *in vitro* or *in vivo* ADME-Tox data are subject to commercial licenses and usually concern private chemical collections.[7] However, while some ADME-Tox data have been already published they are still sparse throughout a number of databases, which creates difficulties in constituting large datasets in order to perform robust and reliable computational models.[8] Today, a multitude of online life sciences databases have emerged and could also contribute to the elaboration of larger datasets than those already published. Nevertheless, numerous concerns are raised when scientists want to collect relevant ADME-Tox data from online databases.[9]

Most databases integrate a wide variety of numerical data and cannot be easily assigned to a specific area of research. Although a minority of databases is domain-specific, many of them provide some data applicable to multidisciplinary domains of research. Considering that the ADME-Tox information represent a very small part of data commonly found on the web, it is therefore not surprising that databases are not clearly designated as providing data related to PK properties. As an example, the Database Commons catalog (http://databasecommons.org) and the Nucleic Acids Research online Database Collection[10] do not take into account ADME-Tox category to classify databases. Consequently, it remains more difficult to locate easily databases for a specific application related to pharmacokinetics.

All databases accessible online are not necessarily available. Although many resources are freely available, some databases are commercial or only accessible online without any possibility to download the data. The data availability can be a serious impediment to collect data, standardize data and develop predictive models.

Information on molecular structures or experimental data are present in many databases. Therefore, the risks of data redundancy and errors propagation from supplier to client databases are growing.[11] Understanding the relationships between online resources could be crucial in order to save time and to locate the most useful database for specific needs. The need to have such global network of interconnected life sciences-related databases and more precisely a global ADME-Tox network has been previously discussed[12], but it has never been implemented and proposed to the scientific community.

While some databases contain highly curated data, some only provide crude data without careful analysis and critical assessment. Consequently, data quality can strongly varied between databases created from different sources. Nevertheless, information on data curation and data quality is not always clearly defined and it remains difficult to determine whether or not collected data may contain erroneous measurements.

In the present work, a comprehensive network map of ADME-Tox data sources was proposed to offer a global understanding of the relationships between them and to easily retrieve data for a specific need. Particular attention was given to the description of data content and availability. A network analysis was proposed to determine the level of connections between databases and characterize their role of supplier or client in the proposed network. Finally, in order to demonstrate the applicability of the network, we proposed a case study with the selection of relevant ADME-Tox databases in order to constitute a large dataset including plasma protein binding (PPB) property. While not performed in the current study, this dataset could be used to build QSAR models for ADME-Tox modeling.

2 Materials and methods

2.1 Creation of ADME-Tox Network

The coherent map was performed with the R package visNetwork[13]. This package is intuitive and fast for the visualization of dynamic network composed of several thousand objects. To build ADME-Tox network, this tool requires set of nodes (*N*) and edges (*E*) representing all the databases and their connections, respectively.

The set *N* was generated with a manual search extracted from VLS3D (http://www.vls3d.com) and Click2Drug (https://www.click2drug.org), which are two comprehensive lists of 150 databases or tools related to ADME-Tox applications. We

described each database by a validated URL, by a field corresponding to data availability (yes or no) and by their corresponding reference (DOI or PMID). A classification based on the Nucleic Acids Research online Database Collection[10] was added to the *N* set, as category or sub-category nodes, and represents the scope of database according to their content. This classification was completed with "ADME-Tox" category and five sub-categories "Absorption", "Distribution", "Metabolism", "Elimination" and "Toxicity".

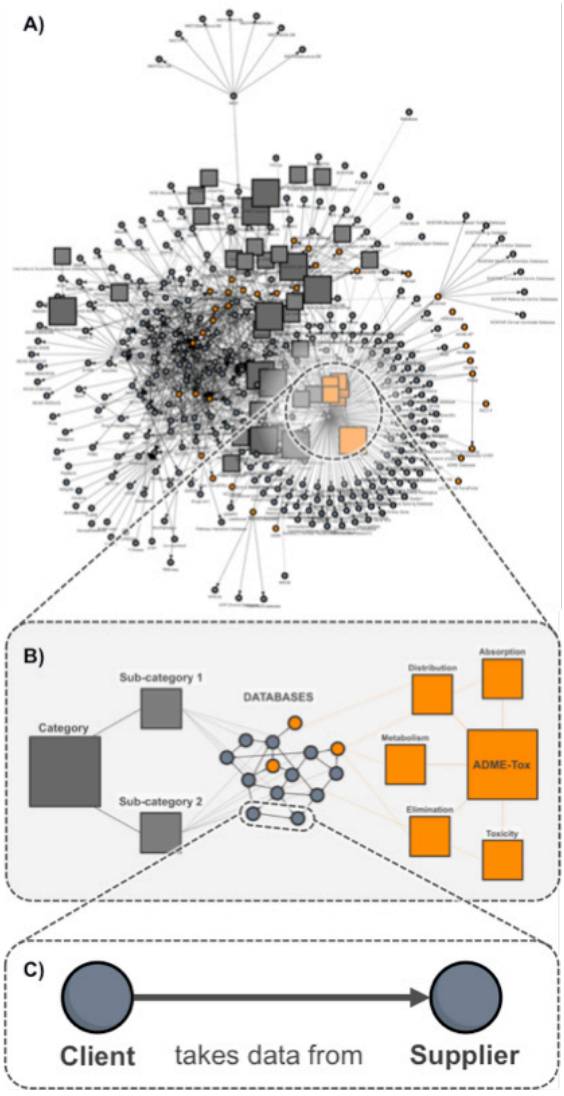


Figure 1. ADME-Tox network. (A) Complete network representation with defined connections between databases manually extracted from published articles. (B) Description of ADME-Tox network features. The "square" and "dot" differentiate categories and databases respectively; node size discerns categories from sub-categories, as well as orange color to distinguish ADME-Tox databases from other nodes. (C) Orientation of edges in the ADME-Tox network.

Information about origin of data sources was manually extracted from articles of published database, or from database websites when available. This information constitutes the edges between nodes (set *E)* containing links of reference source (client or supplier) between databases. Moreover, we added connections between databases and category or sub-category nodes that help to characterize multidisciplinary content of data resource directly on the network map. In some cases, the analysis of the origin of the data has allowed us to identify new databases (not included in VLS3D or Click2Drug lists), which were then added to the set of nodes (set *E*).

2.2 Network analysis

After the construction of the network, we removed the databases, which are not connected to any categories or sub-categories. They correspond to unexploitable resources in the network because of their single edge (network boundary) and consequently these databases are not relevant to perform analysis. A sub-network was obtained, which contains only databases with an exhaustive description of their data content. From this reduced map, we took into account all the links present between databases and the orientation of the edges in order to characterize the role, as supplier or client, of each connected ADME-Tox data sources. The edges between databases and categories were not counted since it does not provide any relevant information and would overestimate the number of links.

For each database, we calculated the number of nodes corresponding to the sum of the number of suppliers and clients providing or requesting information respectively. Finally, a database was considered as “Supplier”, if the number of suppliers is higher than the number of clients, otherwise it was defined as “Client”.

2.3 Dataset preparation

To prepare and clean chemical structures, each database uses its own structure preparation protocol. Consequently, it is possible to have different structures from diverse databases corresponding to the same compound.[14] To address this issue, we have standardized all chemical structures extracted from selected databases with the following protocol: mixture, salts and solvent were removed, charges were neutralized and molecular structures were cleaned using ChemAxon[15]. The standardization was complete by the protonation of the structures at pH 7.4. Standard InChI and standard

InChIKey were calculated from standardized chemical structures. When multiple values of a property were present in a dataset, we selected the maximal value.

3 Results

3.1 Presentation of ADME-Tox network

The complete network, presented in Figure 1.A, contains 11 categories, 29 sub-categories and 373 databases connected by 865 edges. Different features were applied to facilitate understanding of the network (Figure 1.B). Grey squares represent categories and sub-categories from the online Database Collection while orange squares and dots represent the newly added ADME-tox data (Figure 1.B). The links are oriented depending of the original source and represent the flow of diverse data between “client” and “supplier” databases (Figure 1.C). The network was implemented in a dynamic web page containing some widgets to help the user in the selection of subsets of nodes using database or category name as presented in supporting information.

3.2 Analysis of ADME-Tox databases connections

The network analysis was performed using sub-network resulting from data preparation. A total of 267 databases, which are not connected to any categories or sub-categories, were removed. The reduced network and all information included in the database related to ADME-Tox properties are provided in supporting information.

Determine the most active databases in term of data exchanges and highlight databases according to their global roles are two key factors to locate the most useful data sources. In this study, the number of connections and orientation of the links are used to evaluate the global role of each ADME-Tox database. The number of “Client” and “Supplier” of each database and their respective connections are shown in Figure 2. The majority of the most connected databases are those proposing downloadable datasets. Accordingly, for the ADME-Tox category, three groups are now identified in the network such as “Supplier”, “Client” and “Undefined”.

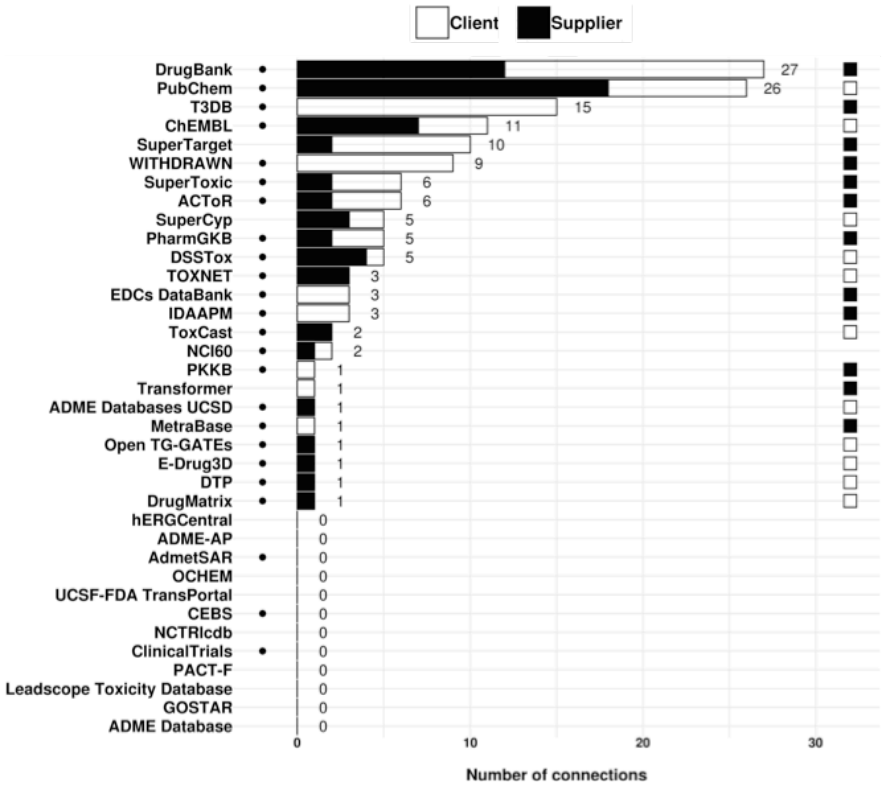


Figure 2. Client and Supplier databases used in the network. Bar chart corresponds to the total number of connections per database, where white and black colors represent the number of suppliers and clients, respectively. The squares represent the main groups ("Supplier", "Client") of the databases in the network. The black filled circles indicate whether a database provides a downloadable dataset. The empty circles represent undownloadable data from the database.

3.3 Case study: Preparation of dataset of plasma protein binding

Drugs present in the blood circulation are in reversible association with plasma proteins. The ratio unbound/(unbound + bound) characterizes the fraction of free drug available to induce the therapeutic effect. The unbound fraction of the drug is also the portion that may be metabolized and/or eliminated. As a reservoir, the bound fraction is released to maintain the equilibrium and therefore is related to the metabolic half-life of a drug in the body. Therefore, PPB is a key property of drug distribution, which can also impact pharmacological properties of the drugs.

Many questions necessarily arise during searching data for a specific ADME-Tox application, e.g., where are the databases that contain data available to my needs? Are the data easily accessible? Are the data redundant and/or reliable? In this section, we will

show how the network map of ADME-Tox databases can help to locate and collect several data to constitute a large dataset related to PPB.

We used the reduced network to identify 16 potential sources of PPB data in the "Distribution" sub-category. After careful analysis, only 7 databases contained PPB data. AdmetSAR[16], PharmGKB[17] and E-Drug3D[18] databases were excluded because their PPB data is not accessible as a downloadable dataset. ClinicalTrials[19] database was discarded due to the low number of accessible assays with known chemical structures. Only data from ChEMBL[20], PubChem[21] and DrugBank[22] databases were extracted and prepared. A total of 3,436 molecules with PPB values were retrieved.

In order to control the chemical structures redundancy, we used the Venn diagram shown in Figure3.A, which represents the overlap between datasets using the standard InChIKey. Only 37 molecules are common to all datasets and ChEMBL covers larger parts of PubChem with 754 overlapped molecules. Furthermore, DrugBank has 76 molecules in common with the other two databases and 756 unique molecules not present in the other two databases. To see how DrugBank is different from PubChem and ChEMBL, we compared PPB distribution of each dataset as depicted in Figure 3.B. Whereas ChEMBL, PubChem and DrugBank are left-skewed for the PPB distribution (medians around 95%), DrugBank covers a wider range of PPB values since 50% of the values are between 50 and 98% of PPB. Nevertheless, the representation of the 37 common molecules based on their identical InChIKey shows that DrugBank values are different from the two other databases on the boxplots (white circle).

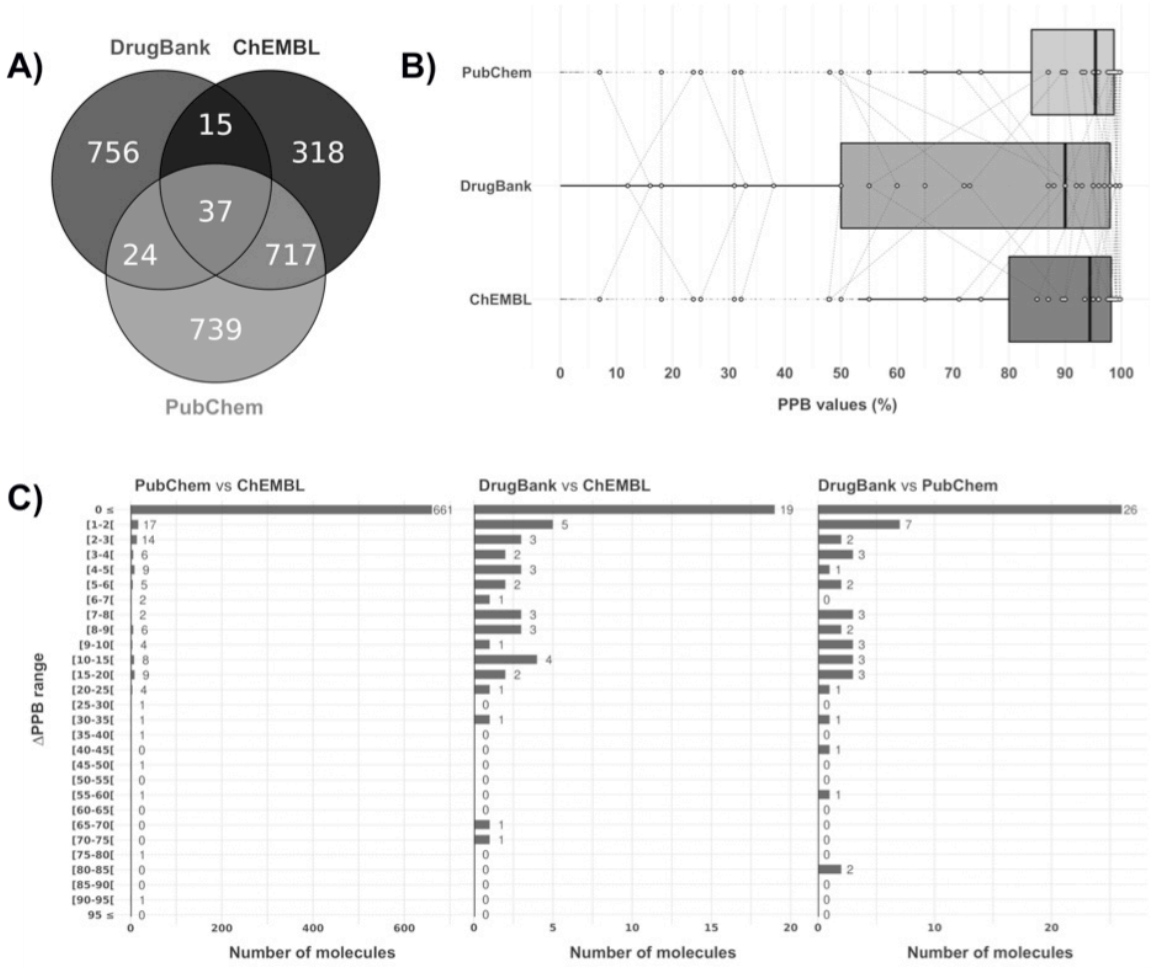


Figure 3. Analysis of PPB datasets extracted from ChEMBL, DrugBank and PubChem. (A) Venn diagram comparing molecules in each dataset. (B) Box plots of PPB distribution with representation of 37 common molecules (white circle) in the 3 databases. (C) Difference (ΔPPB) of binned PPB values between ChEMBL, DrugBank and PubChem.

The data reliability is another important issue to provide quality data for the development of reliable models. Figure 3.C depicts the PPB difference between each dataset using 28 bins. Among the 754 molecules in common between PubChem and ChEMBL, PPB values are approximately 87% identical and 10% with a difference between 1 and 10% of PPB and only 3% are different with a range of difference of binned PPB (ΔPPB) higher than 10%. According to the results of the overlap and PPB difference between datasets, ChEMBL and PubChem are roughly equivalent in term of PPB measurements whereas DrugBank is significantly different from ChEMBL and PubChem with only 37% and 42% of identical PPB values respectively.

4 Discussion

Nowadays, scientists are faced with the increase of dataflow widespread in various databases. It becomes difficult to locate relevant ADME-Tox data across the multitude of available life science databases. To address this issue, we propose a comprehensive and interactive network using 373 connected databases (Figure 1.A). On this map, nodes of categories and sub-categories were added to help users to easily identify relevant databases. In this regard, information about data availability was added for each database and this information could save time during data collection in a research project.

We focused our work on a sub-network to allow a new understanding of connections between commonly used databases and attempts to offer an overview of the available PK data resources. As we noted above, the most used databases propose downloadable datasets and three groups of databases are observed in Figure 2. The "Client" group is composed of DrugBank[22], T3DB[23], SuperTarget[24], WITHDRAWN[25], ACToR[26], PharmGKB[17], SuperToxic[27], EDCs DataBank[28], IDAAPM[29], PKKB[30], Transformer[31], and finally MetraBase[32]. Most of them are public databases containing multidisciplinary accessible data from diverse sources. Data contained in these databases are generally curated as far as possible, which make them suitable to build a dataset for predictive modeling. The "Supplier" group contained a majority of databases that are supported by governmental agencies, such as PubChem[21], ChEMBL[20], DSSTox[33], TOXNET[34], ToxCast[35], Open TG-GATEs[36], DTP[37], or even DrugMatrix[38]. Most of them contains *in vitro* and/or *in vivo* assays, as well as information about experimental protocols, which is important to create homogeneous datasets.[39] The third group contained databases with undefined role in the network. This is explained by either a perfect balance in the flow of data exchanges of "Supplier" and "Client" like NCI-60 from DTP or a total absence of connections. Concerning this last point, some databases are public and downloadable but are surprisingly not used as data provider for other databases, such as AdmetSAR[16], CEBS[40] and ClinicalTrials[19], others are public but with an online restricted access like OCHEM,[41] UCSF-FDA TransPortal[42], hERGCentral[43] or ADME-AP[44], and finally some are private with restricted data access like GOSTAR[45], Leadscope[46], ADME Database[47] or PACT-F[48]. It should be noted that the most important commercial databases like Cloe Knowledge[49] and Pharmapendium[50] are not present in the network because they were not present in VLS3D and Click2Drug lists. Consequently, this network represents the most common database in the field of life sciences and gives us the possibility to explore all the gathered data in a unique interactive map.

To know if the present work was useful to find and collect relevant data, we proposed a case study involving the collection of a PPB dataset. The datasets were extracted from ChEMBL, PubChem and DrugBank that have been previously identified as major supplier databases (PubChem and ChEMBL) or major client database (DrugBank). While the number of common molecules between DrugBank and the other two databases is small (Figure 3.A), this case study shows that ChEMBL and PubChem contain a large number of redundant data, despite few different PPB values (Figure 3.C). This observation is in accordance with the known complementarity information between ChEMBL and PubChem[51], which is represented on the network by a balanced edge (reduced network Supporting Information). Therefore, the network orientation of links between these two databases provides information on the data redundancy. Regarding to data reliability, DrugBank provides different PPB values for 76 common structures with the other databases and the majority of ΔPPB are ranging between 0 and 20%. By combining various databases into a unique dataset, care should be taken on the pretreatment of the data, and a careful cleaning process is required. Moreover, 756 molecules remain exclusive to DrugBank. The analysis of the data shows that DrugBank provides complementary data to PubChem and ChEMBL. Thus, a total of 2,606 unique molecules containing experimental values, extracted from these three databases, could be used to predict PPB property whereas the majority of QSAR models already published are based on datasets with a number of molecules below 1,000 compounds.[52–59]

5 Conclusions

In this study, a comprehensive map of ADME-Tox databases extracted from a network of life science databases is proposed. A case study is presented using data extraction of PPB values which resulted in 2,606 molecules with corresponding experimental values. There are several advantages in using this network. First, the present work combines and describes the most commonly used databases in the field of life science and especially in pharmacokinetics. Furthermore, the interactive network map combines all information necessary to understand connections between databases and can be used to select the available resources for a specific application. Finally, the network can be useful in analyzing data redundancy by visualizing the dataflow between data sources. This comprehensive network map is a tool that can find useful application in drug discovery projects and especially in the ADME-Tox research field.

Acknowledgements

This work was also supported by Servier. Authors whish to thank the Région Centre Val de Loire for financial supports.

References

[1] G. Sliwoski, S. Kothiwale, J. Meiler, E. W. Lowe, *Pharmacol. Rev.* **2013**, *66*, 334–395.

[2] M. J. Waring, J. Arrowsmith, A. R. Leach, P. D. Leeson, S. Mandrell, R. M. Owen, G. Pairaudeau, W. D. Pennie, S. D. Pickett, J. Wang, et al., *Nat. Rev. Drug Discov.* **2015**, *14*, 475–486.

[3] J. Hodgson, *Nat. Biotechnol.* **2001**, *19*, 722–726.

[4] A. Boobis, U. Gundert-Remy, P. Kremers, P. Macheras, O. Pelkonen, *Eur. J. Pharm. Sci. Off. J. Eur. Fed. Pharm. Sci.* **2002**, *17*, 183–193.

[5] J. Kirchmair, A. H. Göller, D. Lang, J. Kunze, B. Testa, I. D. Wilson, R. C. Glen, G. Schneider, *Nat. Rev. Drug Discov.* **2015**, *14*, 387–404.

[6] A. Cherkasov, E. N. Muratov, D. Fourches, A. Varnek, I. I. Baskin, M. Cronin, J. Dearden, P. Gramatica, Y. C. Martin, R. Todeschini, et al., *J. Med. Chem.* **2014**, *57*, 4977–5010.

[7] I. V. Tetko, O. Engkvist, U. Koch, J.-L. Reymond, H. Chen, *Mol. Inform.* **2016**, *35*, 615–621.

[8] J. Bajorath, *F1000Research* **2015**, *4*, DOI 10.12688/f1000research.6653.1.

[9] R. Guha, D.-T. Nguyen, N. Southall, A. Jadhav, *Curr. Protoc. Chem. Biol.* **2012**, *4*, 193–209.

[10] D. J. Rigden, X. M. Fernández-Suárez, M. Y. Galperin, *Nucleic Acids Res.* **2016**, *44*, D1-6.

[11] S. Philippi, J. Köhler, *Nat. Rev. Genet.* **2006**, *7*, 482–488.

[12] H. Van de Waterbeemd, M. De Groot, *SAR QSAR Environ. Res.* **2002**, *13*, 391–401.

[13] A. B. V. (vis js library in htmlwidgets/lib, http://visjs.org, http://www.almende.com/home), B. T. (R interface), *visNetwork: Network Visualization Using "Vis.js" Library*, **2016**.

[14] A. Hersey, J. Chambers, L. Bellis, A. Patrícia Bento, A. Gaulton, J. P. Overington, *Drug Discov. Today Technol.* **2015**, *14*, 17–24.

[15] JChem 15.2.9, ChemAxon, **2015**, http://www.chemaxon.com.

[16] F. Cheng, W. Li, Y. Zhou, J. Shen, Z. Wu, G. Liu, P. W. Lee, Y. Tang, *J. Chem. Inf. Model.* **2012**, *52*, 3099–3105.

[17] M. Hewett, D. E. Oliver, D. L. Rubin, K. L. Easton, J. M. Stuart, R. B. Altman, T. E. Klein, *Nucleic Acids Res.* **2002**, *30*, 163–165.

[18] E. Pihan, L. Colliandre, J.-F. Guichou, D. Douguet, *Bioinforma. Oxf. Engl.* **2012**, *28*, 1540–1541.

[19] NLM, NIH, "ClinicalTrials," can be found under https://clinicaltrials.gov, **2017**.

[20] A. Gaulton, L. J. Bellis, A. P. Bento, J. Chambers, M. Davies, A. Hersey, Y. Light, S. McGlinchey, D. Michalovich, B. Al-Lazikani, et al., *Nucleic Acids Res.* **2012**, *40*, D1100-1107.

[21] S. Kim, P. A. Thiessen, E. E. Bolton, J. Chen, G. Fu, A. Gindulyte, L. Han, J. He, S. He, B. A. Shoemaker, et al., *Nucleic Acids Res.* **2016**, *44*, D1202-1213.

[22] D. S. Wishart, C. Knox, A. C. Guo, S. Shrivastava, M. Hassanali, P. Stothard, Z. Chang, J. Woolsey, *Nucleic Acids Res.* **2006**, *34*, D668-672.

[23] E. Lim, A. Pon, Y. Djoumbou, C. Knox, S. Shrivastava, A. C. Guo, V. Neveu, D. S. Wishart, *Nucleic Acids Res.* **2010**, *38*, D781-786.

[24] N. Hecker, J. Ahmed, J. von Eichborn, M. Dunkel, K. Macha, A. Eckert, M. K. Gilson, P. E. Bourne, R. Preissner, *Nucleic Acids Res.* **2012**, *40*, D1113-1117.

[25] V. B. Siramshetty, J. Nickel, C. Omieczynski, B.-O. Gohlke, M. N. Drwal, R. Preissner, *Nucleic Acids Res.* **2016**, *44*, D1080-1086.

[26] R. Judson, A. Richard, D. Dix, K. Houck, F. Elloumi, M. Martin, T. Cathey, T. R. Transue, R. Spencer, M. Wolf, *Toxicol. Appl. Pharmacol.* **2008**, *233*, 7–13.

[27] U. Schmidt, S. Struck, B. Gruening, J. Hossbach, I. S. Jaeger, R. Parol, U. Lindequist, E. Teuscher, R. Preissner, *Nucleic Acids Res.* **2009**, *37*, D295-299.

[28] D. Montes-Grajales, J. Olivero-Verbel, *Toxicology* **2015**, *327*, 87–94.

[29] A. Legehar, H. Xhaard, L. Ghemtio, *J. Cheminformatics* **2016**, *8*, 33.

[30] D. Cao, J. Wang, R. Zhou, Y. Li, H. Yu, T. Hou, *J. Chem. Inf. Model.* **2012**, *52*, 1132–1137.

[31] M. F. Hoffmann, S. C. Preissner, J. Nickel, M. Dunkel, R. Preissner, S. Preissner, *Nucleic Acids Res.* **2014**, *42*, D1113-1117.

[32] L. Mak, D. Marcus, A. Howlett, G. Yarova, G. Duchateau, W. Klaffke, A. Bender, R. C. Glen, *J. Cheminformatics* **2015**, *7*, 31.

[33] US EPA, "DSSTox," can be found under https://www.epa.gov/chemical-research/distributed-structure-searchable-toxicity-dsstox-database, **2017**.

[34] G. C. Fonger, D. Stroup, P. L. Thomas, P. Wexler, *Toxicol. Ind. Health* **2000**, *16*, 4–6.

[35] US EPA, "ToxCast," can be found under https://www.epa.gov/chemical-research/toxicity-forecasting, **2017**.

[36] Y. Igarashi, N. Nakatsu, T. Yamashita, A. Ono, Y. Ohno, T. Urushidani, H. Yamada, *Nucleic Acids Res.* **2015**, *43*, D921-927.

[37] NIH, “Developmental Therapeutics Program (DTP),” can be found under https://dtp.cancer.gov, **2017**.

[38] B. Ganter, S. Tugendreich, C. I. Pearson, E. Ayanoglu, S. Baumhueter, K. A. Bostian, L. Brady, L. J. Browne, J. T. Calvin, G.-J. Day, et al., *J. Biotechnol.* **2005**, *119*, 219–244.

[39] X. Fu, A. Wojak, D. Neagu, M. Ridley, K. Travis, *J. Cheminformatics* **2011**, *3*, 24.

[40] NIH, “Chemical Effects in Biological Systems (CEBS),” can be found under http://tools.niehs.nih.gov/cebs3/ui/, **n.d.**

[41] I. Sushko, S. Novotarskyi, R. Körner, A. K. Pandey, M. Rupp, W. Teetz, S. Brandmaier, A. Abdelaziz, V. V. Prokopenko, V. Y. Tanchuk, et al., *J. Comput. Aided Mol. Des.* **2011**, *25*, 533–554.

[42] K. M. Morrissey, C. C. Wen, S. J. Johns, L. Zhang, S.-M. Huang, K. M. Giacomini, *Clin. Pharmacol. Ther.* **2012**, *92*, 545–546.

[43] F. Du, H. Yu, B. Zou, J. Babcock, S. Long, M. Li, *Assay Drug Dev. Technol.* **2011**, *9*, 580–588.

[44] L. Z. Sun, Z. L. Ji, X. Chen, J. F. Wang, Y. Z. Chen, *Bioinforma. Oxf. Engl.* **2002**, *18*, 1699–1700.

[45] S. A. R. P. Jagarlapudi, K. V. R. Kishan, *Methods Mol. Biol. Clifton NJ* **2009**, *575*, 159–172.

[46] C. Yang, R. D. Benz, M. A. Cheeseman, *Curr. Opin. Drug Discov. Devel.* **2006**, *9*, 124–133.

[47] T. Fujitsu Kyushu Systems, “ADME Database,” can be found under http://www.fujitsu.com/jp/group/kyushu/en/solutions/industry/lifescience/admedatabase, **2017**.

[48] PharmaInformatic, “PACT-F,” can be found under http://www.pharmainformatic.com/html/pact-f.html, **2017**.

[49] Cyprotex, “CloeKnowledge,” can be found under http://www.cloegateway.com, **2017**.

[50] Elsevier, “PharmaPendium,” can be found under https://www.elsevier.com/solutions/pharmapendium-clinical-data, **2017**.

[51] A. P. Bento, A. Gaulton, A. Hersey, L. J. Bellis, J. Chambers, M. Davies, F. A. Krüger, Y. Light, L. Mak, S. McGlinchey, et al., *Nucleic Acids Res.* **2014**, *42*, D1083-1090.

[52] X.-W. Zhu, A. Sedykh, H. Zhu, S.-S. Liu, A. Tropsha, *Pharm. Res.* **2013**, *30*, 1790–1798.

[53] E. M. del Amo, L. Ghemtio, H. Xhaard, M. Yliperttula, A. Urtti, H. Kidron, *PloS One* **2013**, *8*, e74758.

[54] M. Lobell, V. Sivarajah, *Mol. Divers.* **2003**, *7*, 69–87.

[55] J. Wang, G. Krudy, X.-Q. Xie, C. Wu, G. Holland, *J. Chem. Inf. Model.* **2006**, *46*, 2674–2683.

[56] S. Weaver, M. P. Gleeson, *J. Mol. Graph. Model.* **2008**, *26*, 1315–1326.

[57] H. Li, Z. Chen, X. Xu, X. Sui, T. Guo, W. Liu, J. Zhang, *Biopharm. Drug Dispos.* **2011**, *32*, 333–342.

[58] K. Yamazaki, M. Kanaoka, *J. Pharm. Sci.* **2004**, *93*, 1480–1494.

[59] J. R. Votano, M. Parham, L. M. Hall, L. H. Hall, L. B. Kier, S. Oloff, A. Tropsha, *J. Med. Chem.* **2006**, *49*, 7169–7181.

Dans le cadre de cet article, un réseau prenant en considération les relations existantes entre les sources a été proposé afin de pouvoir visualiser les échanges de données entre les différentes bases et ainsi prioriser la sélection et l'extraction de données disponibles (Figure 21.a). Un travail de recensement a également été entrepris afin de proposer aux chercheurs un inventaire permettant de sélectionner les sources de données en fonction de leur contenu (Figure 21.b).

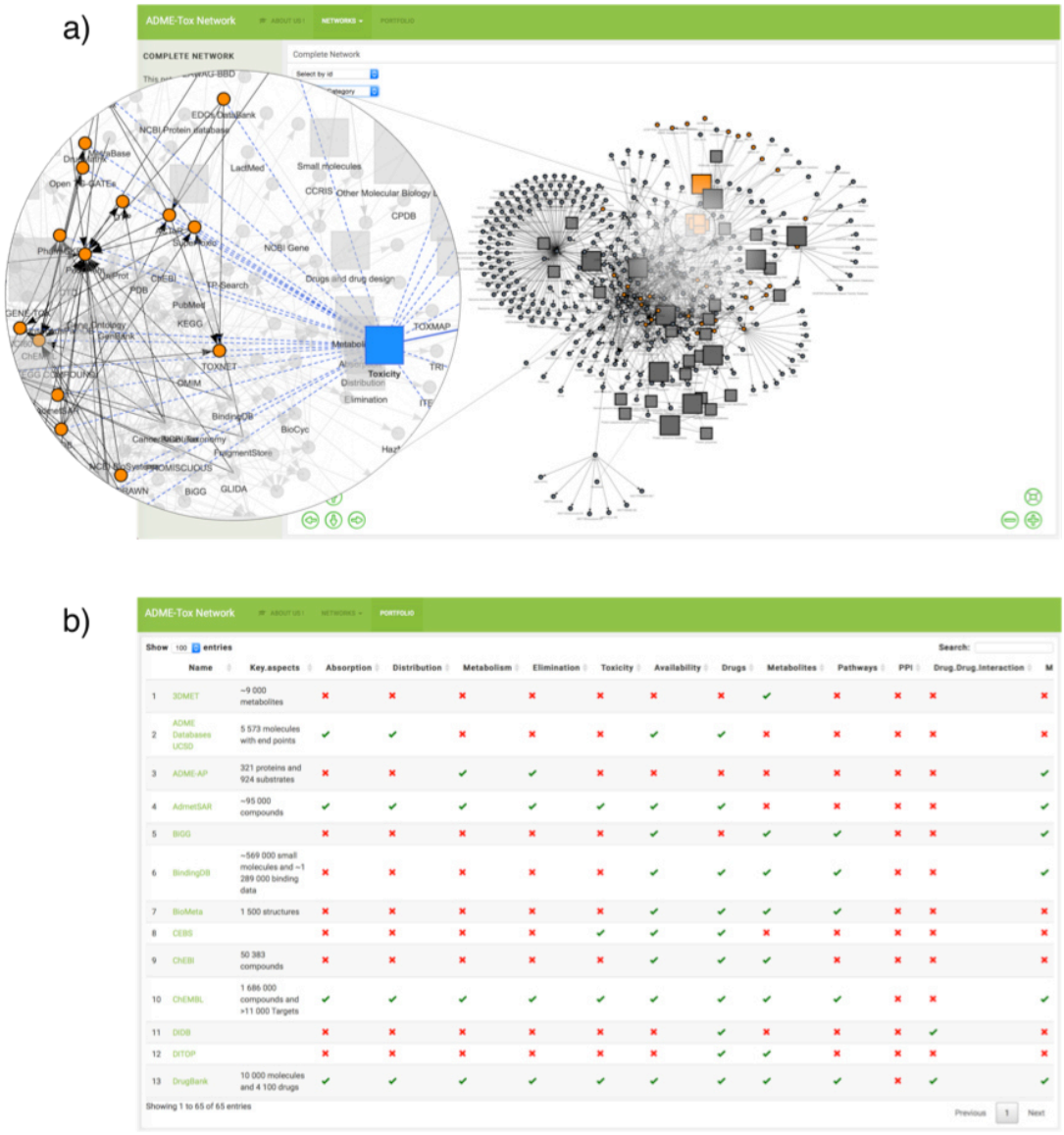


Figure 21 : Illustrations des outils interactifs développés pour l'identification et la sélection des sources de données ADME-Tox.

a) Représentation du réseau interactif de bases de données. Ce réseau permet de visualiser les connexions entre bases, mais également de sélectionner plusieurs bases en fonction de leur catégorie. b) Représentation du recensement des sources de données ADME-Tox. Cette liste propose plusieurs colonnes concernant le contenu des bases de données référencées.

Ce réseau a joué un rôle important dans le cadre de cette thèse afin d'identifier les sources les plus sollicitées par leurs environnements et les sources « mères » qui proposent des données peu ou non modifiées. Ceci permet de pallier les éventuelles modifications et intégrations d'erreurs par les traitements de données opérés par les sources intermédiaires comme illustré par la Figure 22.

a)

Toxicity Profile	
Toxicity Values	LD50: 1010 mg/kg (Oral, Rat) (11) LD50: 358 mg/kg (Intravenous, Rat) (11) LD50: 880 mg/kg (Intraperitoneal, Rat) (11)

b)

Toxicity

Organism	Test Type	Route	Reported Dose (Normalized Dose)	Effect	Source
mammal (species unspecified)	LD50	intraperitoneal	> 1250mg/kg (1250mg/kg)		Archives of Environmental Contamination and Toxicology. Vol. 6, Pg. 279, 1977.
mammal (species unspecified)	LD50	oral	4gm/kg (4000mg/kg)		Archives of Environmental Contamination and Toxicology. Vol. 6, Pg. 279, 1977.
mouse	LD50	intraperitoneal	880mg/kg (880mg/kg)		Bulletin of Environmental Contamination and Toxicology. Vol. 8, Pg. 245, 1972.
rat	LD50	intravenous	358mg/kg (358mg/kg)	LUNGS, THORAX, OR RESPIRATION: DYSPNEA BEHAVIORAL: SOMNOLENCE (GENERAL DEPRESSED ACTIVITY) GASTROINTESTINAL: "HYPERMOTILITY, DIARRHEA"	Food and Cosmetics Toxicology. Vol. 12, Pg. 63, 1974.
rat	LD50	oral	1010mg/kg (1010mg/kg)		Toxicology and Applied Pharmacology. Vol. 60, Pg. 33, 1981.

Figure 22 : Exemple d'une modification non désirée des données ChemIDplus[217] par T3DB[218] pour le 2,2',3,3',4-pentachlorobiphenyl.

La mesure transmise par T3DB (a) n'est pas identique à celle transmise par ChemIDplus (b) au sujet de l'espèce (représenté en rouge).

Grâce à ce travail nous avons pu sélectionner les sources de données à extraire en priorité afin de créer les jeux de données ADME-Tox.

3. Extraction et uniformisation des données

Une fois les données extraites des différentes sources, leur vérification et leur uniformisation sont indispensables afin de créer des jeux de données homogènes. La vérification consiste à contrôler l'intégrité des données (structures moléculaires, mesures, unités, etc.) par rapport à la référence des données. L'uniformisation des données consiste à utiliser des informations à partir desquelles il est possible de sélectionner des points de mesures similaires, c'est-à-dire ayant été déterminés selon des protocoles expérimentaux équivalents. Cette remarque est d'autant plus importante au sujet des propriétés ADME-Tox, car les valeurs de ces dernières dépendent majoritairement des conditions opératoires utilisées pour les mesurer. En effet, ces propriétés physiologiques sont déterminées à l'aide de tests *in vitro* ou *in vivo* et sont donc influencées par les conditions opératoires comme par exemple la voie d'administration, la dose administrée, la température, le temps d'incubation, le pH, l'organisme, les fluides biologiques, les

organes, les lignées cellulaires ou encore la méthode d'analyse utilisés pour effectuer la mesure, etc. L'ensemble de ces conditions opératoires constitue des données annexes, aussi appelées métadonnées. Bien que ces métadonnées soient importantes, elles ne sont pratiquement jamais explicitement renseignées dans lesbases de données.

Les données ADME-Tox étant glanées à partir de plusieurs articles, il n'est pas rare de rencontrer des dénominations différentes de la propriété. En effet, contrairement aux valeurs d'IC_{50} ou d'autres mesures de ce type, pour lesquelles les scientifiques utilisent un terme générique pour partager leurs résultats à la communauté, les termes faisant référence aux propriétés ADME-Tox sont peu uniformisés. Ainsi, plusieurs noms peuvent être rencontrés pour une même propriété ADME-Tox. Cette observation peut également être faite au sujet des unités [219]. Plusieurs unités peuvent être rencontrées pour l'ensemble des propriétés ADME-Tox, comme par exemple des mL/min/kg ou des L/h pour la clairance. Ces particularités empêchent l'utilisation d'outils informatiques automatiques pour parcourir, uniformiser ou sélectionner les données les plus adaptées pour la problématique étudiée. Lors de cette thèse, nous avons dû faire face à plusieurs obstacles de ce type pour lesquels nous avons développés différentes stratégies dans le but de créer des jeux de données homogènes. Afin d'illustrer nos propos, nous allons présenter l'extraction de quelques sources de données que nous avons eu l'occasion d'explorer.

3.1. ADMET Xtractor : analyseur de texte pour la ChEMBL

La ChEMBL est une base de données créée par l'Institut Européen de Bioinformatique (EBI) qui a été initialement conçue afin de regrouper des millions de données au sujet de l'activité ou de l'affinité biologique de petites molécules pour plusieurs milliers de protéines [220,221]. En 2014, la ChEMBL a fait évoluer son architecture afin de pouvoir intégrer des données ADME-Tox, et a proposé dans la foulée le service ADME SARfari intégralement dédié à la prédiction de ces propriétés. Depuis 4 ans, cette base de données fournit des efforts considérables afin d'améliorer l'exactitude et la qualité des informations transmises au sujet des données ADME-Tox [222].

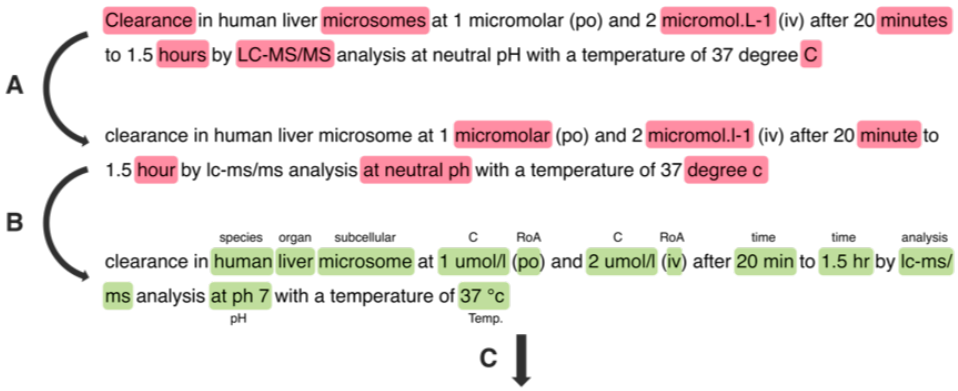
La ChEMBL peut être téléchargée sous différents formats afin d'être directement intégrée dans des systèmes de gestion internes au laboratoire. Cependant, au fil des années, le schéma relationnel de la base de données évolue rendant obsolète les requêtes SQL préalablement établies, ce qui pose un souci majeur concernant la maintenance du système. De plus, comme énoncé par Papadatos *et al.* en 2015, nous avons également

observé qu'une propriété ADME-Tox pouvait être nommée de diverses manières dans cette base, ce qui rend compliquée l'identification des mesures pour la propriété souhaitée [219]. Afin de pallier ces limitations, nous avons choisi d'utiliser le service en ligne proposé par la ChEMBL (https://www.ebi.ac.uk/chembldb), qui permet d'effectuer une requête selon les termes spécifiques à une propriété ADME-Tox. Ce service en ligne nous donne ensuite l'ensemble des essais biologiques contenant des termes similaires à ceux de la requête effectuée. Il est ensuite possible d'appliquer des filtres successifs afin de ne sélectionner que les noms de propriété ADME-Tox qui nous intéressent et de la sorte rejeter toutes les dénominations qui nous sont inconnues. Un exemple de la procédure que nous avons suivie pour extraire les données de la ChEMBL est présenté en ANNEXE B. Pour la suite des explications, nous garderons l'exemple des données extraites pour la fraction libre dans le plasma.

Comme illustré par la Table 4, seul le champ PREF_NAME de la base de données fait référence aux conditions opératoires énoncées précédemment (organes, fluides, tissus, espèces, etc.). Ainsi, l'information transmise par ce champ est corrompue car il contient des indications différentes, ce qui rend impossible la sélection des données selon ces caractéristiques. De plus, ces informations ne coïncident pas obligatoirement avec les descriptions des mesures. Nous remarquons également que le champ DESCRIPTION apporte des informations complémentaires (voie d'administration, concentration, espèce, organe, fluide, etc.) pouvant être utilisées afin d'affiner notre jugement lors de la création du jeu de données. Ces informations sont contenues dans des phrases et par conséquent, ne sont pas directement exploitables.

Afin d'extraire ces informations automatiquement, nous avons mis en place un analyseur de texte nommé « ADMET Xtractor » dont le principe repose sur trois étapes (Figure 23): i) la description de la mesure va être uniformisée à l'aide d'un outil de traitement de langage naturel (spaCy [223]), ii) plusieurs expressions régulières vont permettre d'uniformiser les notations scientifiques, et iii) plusieurs dictionnaires d'équivalences basés sur les documentations de la FDA [224] ainsi que des expressions régulières sont ensuite utilisés pour identifier les informations contenues dans le texte. Cet outil nous offre la possibilité d'extraire des métadonnées énoncées dans la description, comme la voie d'administration (*RoA*), la concentration (*C*), la température (*Temp*), le temps (*Time*), la méthode d'analyse (*Analysis*), le pH (*pH*), l'espèce (*Species*), l'organe (*Organ*), le fluide (*Fluid*), le tissue (*Tissue*) ou encore la fraction sous-cellulaire (*SubCell*). Cet outil est évolutif et permet d'ajouter de nouveaux dictionnaires d'équivalences pour rechercher

encore plus d'informations. L'inconvénient est qu'il ne prend pas en compte la syntaxe d'une phrase, c'est-à-dire qu'il n'est pas capable d'identifier les relations existantes entre les informations qu'il extrait. Par exemple dans la Figure 23, les deux concentrations sont chacune liée à une voie d'administration spécifique, mais cette relation n'est pas prise en considération par l'outil. Néanmoins, ADMET Xtractor permet d'extraire rapidement l'ensemble des critères souhaités. Il a été utilisé comme une assistance permettant d'accélérer l'exploration des données lors de cette thèse. Un exemple d'application est représenté par la Table 5. Il est à noter que les unités des paramètres extraits peuvent être différentes. Pour cette raison, nous avons créé un convertisseur universel d'unité permettant de convertir des distances, des volumes, des masses, des temps et des températures. Ceci a été intégré à ADMET Xtractor afin d'obtenir des informations standardisées selon les mêmes unités (Table 5). Un fois l'ensemble des métadonnées extraites, il est possible d'explorer l'ensemble de données afin de sélectionner les mesures correspondantes à la propriété ADME-Tox visée.



Species	Organ	SubCell	C	RoA	Time	Analysis	pH	Temp
Human	Liver	Microsome	1 umol/l, 2 umol/l	Oral, Intravenous	20 min, 1.5 hr	LC, MS, MS	7	37°C

Figure 23 : Illustration du protocole suivi par ADMET Xtractor.

La phrase illustrée a été construite afin de montrer un cas typique pouvant être rencontré dans une base de données avec des notations particulières. L'étape A consiste à uniformiser la phrase à l'aide de spaCy. L'étape B a pour objectif d'uniformiser les notations scientifiques comme les unités. L'étape C utilise les dictionnaires d'équivalences internes à ADMET Xtractor pour identifier et récupérer les informations présentes dans la description. En résulte une table avec l'ensemble des informations que l'analyseur a identifié à partir de la description.

CMPD_CHEMBLID	STANDARD_TYPE	RELATION	STANDARD_VALUE	STANDARD_UNITS	DESCRIPTION	ORGANISM	PREF_NAME
CHEMBL401796	Fu	=	0.602		Fraction unbound in Beagle dog liver microsomes at 1 uM	Canis lupus familiaris	Liver
CHEMBL480531	Fu	<	0.01		Fraction unbound in human plasma	Homo sapiens	Plasma
CHEMBL1098000	Fu	=	0.0254		Fraction unbound in baboon plasma		ADMET
CHEMBL452664	Fu	=	0.1		Fraction unbound in rat plasma at 3 mg/kg, iv and 10 mg/kg, po	Rattus norvegicus	Plasma
CHEMBL599	Fu	=	0.0018		Fraction unbound in human plasma at 15 mg, po	Homo sapiens	Plasma
CHEMBL246284	Fu	=	0.026		Fraction unbound ligand in mouse serum at 10 uM by equilibrium dialysis	Mus musculus	Plasma
CHEMBL27	Fu	=	0.01		Fraction unbound ligand in rat adrenal gland after 20 hrs	Rattus norvegicus	Adrenal gland
CHEMBL270083	Fu	=	0.839		Fraction unbound drug in Wistar rat assessed as excretion in urine at 50 mg/kg, iv	Rattus norvegicus	Rattus norvegicus

Table 4 : Exemple de données extraites de la ChEMBL pour la fraction libre.

CMPD_CHEMBLID : identifiant du composé ; STANDARD_TYPE : nom standardisé de la propriété ; RELATION : opérateur lié à la valeur de la propriété ; STANDARD_VALUE : valeur standardisée de la propriété ; STANDARD_UNITS : unité de la mesure ; DESCRIPTION : description de la mesure ; ORGANISM : organisme sur lequel la mesure a été faite ; PREF_NAME : information sur l'échantillon utilisé pour effectuer la mesure.

DESCRIPTION	RoA	Analysis	C	Fluid	Organ	Species	Tissue	SubCell	Temp	Time	pH
Fraction unbound in Beagle dog liver microsomes at 1 uM			1e-06 mol/l		Liver	Dog		Microsome			
Fraction unbound in human plasma				Plasma		Human					
Fraction unbound in baboon plasma				Plasma		Monkey					
Fraction unbound in rat plasma at 3 mg/kg, iv and 10 mg/kg, po	Oral, Intravenous		3e-06 g/g, 1e-05 g/g	Plasma		Rat					
Fraction unbound in human plasma at 15 mg, po	Oral			Plasma		Human					
Fraction unbound (%) ligand in mouse serum at 10 uM by equilibrium dialysis		Dialysis, Equilibrium	1e-05 mol/l	Serum		Mouse					
Fraction unbound ligand in rat adrenal gland after 20 hrs						Rat	Adrenal			1200.0min	
Fraction unbound drug in Wistar rat assessed as excretion in urine at 50 mg/kg, iv	Intravenous		5e-05 g/g	Urine		Rat					

Table 5 : Exemple de métadonnées extraites suite à l'utilisation de l'outil ADMET Xtractor.

Les descriptions proviennent de l'exemple présenté dans la Table 4. L'utilisation d'ADMET Xtractor permet de voir que plusieurs mesures ne correspondent pas à la fraction libre dans le plasma.

Dans le cadre de notre exemple, nous souhaitons modéliser la fraction libre d'un composé dans le plasma chez l'humain. Nous avons priorisé dans un premier temps les données humaines pour lesquelles l'information sur la structure moléculaire des composés et la valeur numérique de la propriété sont disponibles. Un second filtre est appliqué afin de ne sélectionner que les composés qui disposent d'une valeur exacte de la propriété avec un champ RELATION ayant une valeur « = » (Table 4). En effet, les composés qui possèdent des opérateurs supérieurs ou inférieurs n'ont pas de valeurs fixes, ce qui rend ces mesures inexploitables pour la création d'un modèle de régression, et pour lesquelles aucune valeur exacte ne peut être utilisée lors de l'élaboration d'un modèle de classification. Au total, 2090 mesures de la fraction libre respectent nos filtres sur les 5122 mesures initiales publiées. L'ensemble des métadonnées correctement identifiées peuvent être utilisées pour explorer les données. Comme représenté par la Table 6, le nom de la propriété « Fu » utilisée dans la ChEMBL peut faire référence à plusieurs notions de fraction libre.

STANDARD_TYPE	RELATION	Species	Fluid	Organ	SubCell	Count
Fu	=	Human				610
					Hepatocyte	41
				Brain		47
				Intestine	Microsome	7
				Kidney	Microsome	7
				Liver		4
					Microsome	47
			Blood			10
			Plasma			1221
				Liver		4
					Microsome	4
			Serum			81
				Kidney	Microsome	1
				Liver	Microsome	6

Table 6 : Comptage du nombre de données en fonction des métadonnées extraites par ADMET Xtractor.

En vert sont représentées les informations utilisées pour sélectionnes les données de fraction libre dans le plasma. Les informations concernant les tissus biologiques n'ont pas été représentés, car pour les données humaines aucune correspondance n'a été identifiée par ADMET Xtractor.

Ainsi, les informations complémentaires nous montrent que plusieurs mesures concernent des organes spécifiques comme le cerveau, les intestins, le foie ou encore les reins. Ces mesures sont effectuées afin de déterminer la distribution du principe actif entre le sang et divers organes, afin de caractériser la fixation tissulaire du médicament. En d'autres termes, ces mesures ne traduisent pas la fraction libre dans le plasma (Fu_p) que nous souhaitons modéliser. Concernant l'information sur les fluides biologiques, nous remarquons que 763 mesures n'ont pas d'information, 10 mesures ont été obtenues à

partir du sang, 1229 à partir du plasma et 88 à partir du sérum. Afin d'extraire un sous-ensemble de données *a priori* homogène, nous avons sélectionné les 1221 mesures qui font référence à la fraction libre dans le plasma et qui ne disposent d'aucune indication sur l'organe. Au total, 271 publications ont été exploitées par la ChEMBL pour obtenir ces mesures. Nous avons entrepris une étape de vérification des données dans les articles de référence a été entreprise. Cette vérification nous a permis de voir que 2 % des mesures n'avaient pas été déterminés sur le bon organisme, qu'uniquement 1 % des valeurs de Fu_p étaient erronées et que moins de 1 % des structures moléculaires ne correspondaient pas à celles présentées dans les publications de référence. Au final, 1196 mesures de la fraction libre dans le plasma ont été obtenues et vérifiées afin de constituer un jeu de données utilisable lors d'une approche de modélisation (Q)SAR.

Nous avons suivi la même approche pour l'extraction d'autres propriétés à partir de la ChEMBL. De plus, la méthodologie que nous avons présentée au sujet de l'outil ADMET Xtractor a été modifié dans le but d'extraire aussi la base de données DrugBank [225], car les mesures qu'elle propose sont présentées sous la forme de phrases. Cette solution a permis l'extraction des données Plasma Protein Binding (PPB), mais des efforts sont encore à fournir pour les autres propriétés ADME-Tox, pour lesquelles une extraction nécessite la compréhension de la syntaxe.

3.2. WebScraper : extracteur de bases de données en ligne

D'autres bases de données proposent des mesures expérimentales correctement formatées pour des propriétés ADME-Tox précises et donnent en plus la possibilité de les visualiser en ligne. C'est le cas par exemple de ChemIDplus [217]. Cette base de données est une composante de la base TOXNET [226] proposée par l'institut national de la santé américain (NIH), dont le rôle est de centraliser les informations chimiques de plus de 400 000 molécules. A partir du service en ligne de cette base de données il est possible à l'aide du numéro CAS d'accéder à diverses informations comme la toxicité (LD_{50}) ou encore plusieurs propriétés physico-chimiques provenant de la base de données PhysProp [227]. Cependant, l'extraction de ces données pour une collection de molécules peut s'avérer complexe et nécessite d'effectuer une requête manuelle pour chacune d'entre elles. Il existe une version téléchargeable de cette base, mais elle ne contient pas les propriétés d'intérêt. Afin de pallier à ce problème, nous avons créé l'outil WebScraper dont le rôle est d'extraire les données proposées par les bases accessibles en ligne. Cet outil a été développé à l'aide du langage de programmation Python et des librairies *requests*, *BeautifulSoup*, *lxml*, et *re*. Son fonctionnement est simple et repose sur trois

étapes : i) la base de données est interrogée pour une molécule souhaitée, et ceci en adéquation avec sa politique de partage de données, ii) si une réponse est obtenue la base nous fournit les mesures correspondantes, iii) les informations contenues dans le fichier rapatrié sont extraites sous la forme d'une table pouvant être facilement exploitée. WebScraper nous a donné la possibilité d'extraire plus de 16 000 mesures de la LD_{50} à partir de la base de données ChemIDplus en à peine deux jours.

Il existe plusieurs bases de ce type pour lesquelles un protocole d'extraction, en accord avec la politique de partage de chaque base, a été établi et intégré à WebScraper. Par exemple la base de données PubChem [228], supportée par le NIH, propose plus de 236 millions de mesures pour environ 3 millions de composés. Les données ADME-Tox contenues dans cette base ne représentent qu'une infime partie des mesures proposées. Par conséquent, nous n'avons pas choisi de télécharger et d'intégrer la base de données PubChem en interne. Notre solution a été de rechercher au préalable l'ensemble des essais correspondant aux propriétés ADME-Tox visées, dans le but d'extraire uniquement les données dont nous avions réellement besoin, à l'aide de WebScraper. D'autres protocoles de ce type ont été développés pour la base de données PhysProp ou encore des catalogues chimiques tels que Sigma-Aldrich ou Tocris pour lesquels des données de solubilité dans plusieurs solvants peuvent être extraites.

3.3. WebChem : extracteur d'information chimique en ligne

Lors de notre exploration des bases de données comme PubChem et ChEMBL, nous avons remarqué que la plupart des jeux de données ADME-Tox récemment publiés et de taille conséquente n'ont pas été intégrés depuis l'année 2016. De ce fait, un travail supplémentaire a été entrepris dans le but d'identifier et d'extraire l'ensemble des jeux de données publiés pour chacune des propriétés ADME-Tox souhaitées. Les mesures provenant des jeux de données publiés sont *a priori* homogènes. Nous avons tout de même vérifié, dans la mesure du possible, les données extraites de ces sources afin de déterminer les éventuelles erreurs concernant la propriété modélisée et la structure chimique qui lui était associée. Néanmoins comme nous l'avons énoncé précédemment (Ch2 1.1), les structures moléculaires des composés ne sont pas toujours clairement renseignées. C'est le cas par exemple du jeu de données proposé par Obach *et al.* [229].

Ce jeu de données contient des mesures ADME-Tox pour 670 médicaments présentés sous la forme d'une table directement intégrée dans le corps de la publication. La première difficulté pour l'utilisation de ces données a été d'extraire les mesures provenant

de la publication sous le format pdf. Pour cela, nous avons utilisé un outil en ligne permettant de convertir les tableaux d'un fichier pdf sous un format excel (http://pdftoxls.com). Une fois les données extraites, nous avons vérifié manuellement que les mesures concordaient avec les données publiées. L'ensemble des paramètres ADME-Tox présentés dans la publication ne sont pas directement exploitables pour la création d'un modèle (Q)SAR. En effet, l'information sur la structure chimique n'est pas renseignée et seul le nom générique de la molécule est répertorié comme illustré par la Table 7.

ChemName	Fu	Unit	SMILES	Search
Abacavir	50.0	%	N=c1[nH]c(=NC2CC2)c2ncn(C3C=CC(CO)C3)c2[nH]1	Check
Acebutolol	74.0	%	CCCC(=O)Nc1ccc(OCC(O)CNC(C)C)c(C(C)=O)c1	OK
Acecainide	90.0	%	CCN(CC)CCNC(=O)c1ccc(NC(C)=O)cc1	OK
Acetaminophen	52.0	%	[2H]c1c([2H])c(NC(C)=O)c([2H])c([2H])c1O	OK
Acetazolamide	4.0	%	CC(=O)N=c1[nH]nc(S(N)(=O)=O)s1	OK

Table 7 : Exemple des données de fraction libre présentées par Obach *et al.* traitées par l'outil WebChem.

Le nom générique de la molécule est renseigné dans la colonne ChemName et a été utilisé pour effectuer la recherche de structure moléculaire à l'aide de l'outil WebChem. Les colonnes SMILES et Search ont été ajoutées par l'outil. Le terme *Check* de la colonne Search informe l'utilisateur que les SMILES obtenus à partir des différentes bases de données ne sont pas concordants et qu'une vérification manuelle est nécessaire pour choisir la structure chimique désirée.

Nous avons créé l'outil WebChem dans le but d'accéder à la structure chimique des composés. Cet outil a été développé en Python selon une procédure similaire à celle présentée dans le cadre de WebScraper. Il permet d'interroger simultanément plusieurs bases de données, comme par exemple ChemSpider [230], PubChem, Wikipédia [231], ou encore le service d'identification universel du NIH (CIR) [232]. Les structures chimiques obtenues à partir de ces différentes sources sont ensuite standardisées selon notre protocole interne (Ch3 1.1.2), puis comparées afin de pallier les éventuelles erreurs d'identification. Si les structures chimiques extraites ne sont pas concordantes, WebChem nous informe qu'une vérification manuelle doit-être effectuée (le terme *Check* est ajouté à la colonne *Search*) comme illustré par la Table 7. Cet outil peut paraître simpliste et son utilité peut être contestée, mais il nous a permis d'optimiser notre temps afin d'identifier les structures chimiques pour plusieurs jeux de données.

4. ADMET db : base de données interne

Les données extraites et vérifiées à partir de plusieurs sources de données ont été préparées, de façon à ce que toutes les mesures d'une propriété ADME-Tox soient homogènes avec des conditions opératoires similaires, des unités uniformisées et que les

structures chimiques des composés associés soient correctement identifiées. Au final, des données ont été préparées pour 49 propriétés ADME-Tox (Table 8) et constituent notre base de données ADMET db.

ENDPOINT	DESCRIPTION	DATA	ADMET
LogS_H2O	Solubility in Water	10150	PC
LogS_DMSO	Solubility in DMSO	5155	PC
LogS_EtOH	Solubiliity in EtOH	1276	PC
LogPwo	Water/Octanol partitioning coeficient	8200	PC
LogD7.4	LogP at pH 7.4	7484	PC
LogPapp	Apparent permeability in Caco-2 cells	2621	A
ASBT_I	ASBT hTP inhibitors	150	A
ASBT_S	ASBT hTP substrates	100	A
BCRP_I	BCRP hTP inhibitors	382	A
BCRP_S	BCRP hTP substrates	146	A
MDR1_I	MDR1 hTP inhibitors	4781	A
MDR1_S	MDR1 hTP substrates	1817	A
MRP1_I	MRP1 hTP inhibitors	418	A
MRP1_S	MRP1 hTP substrates	168	A
MRP2_I	MRP2 hTP inhibitors	96	A
MRP2_S	MRP2 hTP substrates	188	A
OCT1_I	OCT1 hTP inhibitors	199	A
OCT1_S	OCT1 hTP substrates	78	A
PEPT1_I	PEPT1 hTP inhibitors	80	A
PEPT1_S	PEPT1 hTP substrates	158	A
HIA	Human Intestinal Absorption	2636	A
F	Oral bioavailability	1769	A
BBB	Blood-Brain Barrier	5950	A/D
LogBB	Blood-Brain permeation	1329	A/D
Vd	Volume of distribution	952	D
Vdss	Volume of distribution at steady state	2635	D
Fu_plasma	Fraction unbound in human plasma	5004	D
CYP1A2_I	CYP 1A2 inhibition	13239	M/T
CYP2C9_I	CYP 2C9 inhibition	12881	M/T
CYP2C19_I	CYP 2C19 inhibition	13427	M/T
CYP3A4_I	CYP 3A4 inhibition	12997	M/T
CYP2D6_I	CYP 2D6 inhibition	13897	M/T
CYP2C9_S	CYP 2C9 substrates	673	M
CYP2D6_S	CYP 2D6 substrates	673	M
CYP3A4_S	CYP 3A4 substrates	674	M
CL_h	Hepatic clearance	309	E
CL_r	Renal clearance	309	E
CL_tot	Total clearance	2504	E
CL_p	Plasma clearance	758	E
t1/2	half-life	1947	E
AMES	Ames Toxicity	6512	T
hERG	hERG channel Toxicity	5984	T
Carcino.	Carcinogenecity	280	T
DILI	Drug Induced-Liver Injury (Hepatotoxicity)	1773	T
LD50_R_iv	Acute IV toxicity in rat	1305	T
LD50_R_po	Acute Oral toxicity in rat	5178	T
LD50_M_iv	Acute IV toxicity in mouse	3549	T
LD50_M_po	Acute Oral toxicity in mouse	5508	T
TD50	Carcinogenecity	1197	T

Table 8 : Nombre de données extraites par propriété ADME-Tox.

Cette table présente les propriétés (ENDPOINT) pour lesquelles plusieurs données ont été extraites et vérifiées. Le nombre total de données obtenues par propriété est représenté dans la colonne DATA. La catégorie de la propriété ADME-Tox est renseignée dans la colonne ADMET (PC : propriétés physico-chimiques ; A : Absorption ; D : Distribution ; M : Métabolisme ; E : Elimination ; T : Toxicité).

Le nombre de mesures présenté dans la Table 8 correspond au nombre total de mesures extraites pour une propriété ADME-Tox. Les jeux de données disposent donc de valeurs multiples pour certaines molécules, qu'il sera indispensable de traiter dans le cadre de la plateforme MetaPredict.

5. Conclusion et perspectives

Nous avons réussi à collecter et uniformiser 169 496 points de mesures expérimentales pour 49 propriétés ADME-Tox. Comparée aux autre outils disponibles en ligne, notre base de données incorpore plus de mesures ADME-Tox uniformisées pouvant être utilisées pour la création de modèles (Q)SAR. Notre analyse des différentes sources de données publiques et notre expertise acquise lors des étapes indispensables d'extraction et de préparation de données, nous poussent à croire qu'il est possible d'améliorer l'offre actuellement disponible.

En effet, les nouvelles bases de données ne cherchent pas à créer une valeur ajoutée, mais à acquérir dans un temps restreint les données déjà existantes. A titre d'exemple, nous avons eu l'occasion de voir la création et l'évolution de la base de données IDAAPM [233] durant la période de cette thèse. Cette base de données a été créée en 2016 et avait pour principale mission de proposer une large collection de données en rapport avec les propriétés ADME-Tox. Cependant, nous avons remarqué que les données contenues dans cette base provenaient dans la majorité des cas de la DrugBank et étaient par conséquent non utilisables en l'état. Comme mentionné précédemment, les données de la DrugBank sont présentées sous la forme de descriptions, ce qui nécessite une étape d'extraction manuelle. Ces nouvelles bases proposent peu de données nouvelles, ne permettent pas de pallier le problème des métadonnées et ne permettent pas d'explorer de nouvelles régions de l'espace chimique. Selon nous, une donnée innovante est la mesure d'une propriété ADME-Tox pour une molécule précise qui n'est pas proposée par une autre base de données. L'innovation est donc considérée comme le référencement d'une nouvelle donnée pouvant permettre d'explorer une nouvelle région de l'espace chimique. De plus, il est important de noter que les bases de données comme la ChEMBL n'ont pas intégré les jeux de données les plus récents, tout comme plusieurs études cliniques qui ont été publiées depuis plusieurs décennies.

La perspective envisagée est de proposer à l'avenir une base de données capable de centraliser des mesures ADME-Tox ainsi que leurs métadonnées. Selon nous, les métadonnées à extraire doivent couvrir plusieurs aspects comme i) l'erreur sur la mesure, ii) les conditions opératoires utilisées pour obtenir la valeur de la propriété (Ch2 3.1), et pour finir iii) les aspects démographiques dans le cas d'une étude *in vivo*. Concernant ce dernier point, les données ADME-Tox sont influencées par les aspects démographiques comme par exemple le nombre d'individus sur lesquels la mesure a été effectuée, l'état

de santé de la population des individus étudiés, leur sexe, leur poids ou encore leur âge. Bien que ces informations soient d'une importance cruciale afin de constituer un ensemble de données de haute qualité, elles ne sont jamais extraites et intégrées dans les bases de données accessibles. Par exemple, il n'est pas rare que les jeux de données utilisés pour la clairance hépatique (CL_h) prennent en compte des mesures déterminées chez des individus sains, mais aussi chez des individus atteints d'insuffisance hépatique. Ainsi, l'élaboration d'un modèle peut être compromise en raison des tendances divergentes observées entre ces deux populations d'individus.

Afin d'alimenter notre base de données, plusieurs approches peuvent être envisagées. La première approche consiste à créer un outil capable d'extraire automatiquement les données ADME-Tox dans les articles souhaités et la structure chimique associée. Cependant, bien que les technologiques liées à l'étude du langage naturel et à l'extraction de texte progressent, leur application au domaine de la chimie reste limitée [234,235]. En effet, les données présentées dans les publications peuvent être sous différentes formes, comme par exemple dans le cas extrême, une table présentant plusieurs propriétés ADME-Tox avec une représentation moléculaire partagée sous la forme d'une structure de Markush. Ce cas de figure, pourtant si pratique pour la compréhension de tous, est presque impossible à résoudre par les systèmes d'extraction de données actuels. Pour cette raison, la deuxième approche consiste à extraire manuellement les publications contenant des données ADME-Tox. Il serait envisageable de prioriser les données provenant de journaux spécialisés comme par exemple le journal *Clinical Pharmacokinetics*. Cependant, toutes les références extraites ne proposent pas des mesures ADME-Tox. Ainsi, nous pensons qu'il est possible de créer un outil capable d'évaluer rapidement la qualité d'une publication en se basant sur les principes énoncés par ClinPK [236]. Une fois les publications à extraire identifiées, nous pouvons envisager des approches similaires à celles présentées dans le cadre d'ADMET Xtractor pour faciliter le travail d'extraction. L'utilisation de cet outil combiné à une vérification manuelle peut permettre d'enrichir rapidement la base de données proposée. Nous pensons qu'une base de données contenant l'ensemble de ces informations peut apporter de nombreuses solutions pour la création plus aisée de jeux de données homogènes et spécifiques, afin d'améliorer la précision et la robustesse des modèles ADME-Tox.

Pour conclure, notre objectif a été de rechercher des mesures provenant de bases de données libres et de jeux de données publiés. Notre but était de construire une collection de données consistante pour plusieurs propriétés ADME-Tox. Ceci peut faciliter la

création de vastes jeux de données utilisables pour l'élaboration de modèles globaux, qui sont indispensables dans le cadre de criblages virtuels à haut débit. L'idée sous-jacente est alors d'avoir une base de données interne qui dans un second temps peut nous offrir la possibilité de proposer des modèles locaux pour les séries congénériques étudiées par les chimistes de notre laboratoire. Ainsi, les données de la base ADMET db peuvent être utilisées pour l'élaboration de modèles (Q)SAR. Cependant, la création d'un modèle de prédiction pour chacune des propriétés ADME-Tox peut s'avérer fastidieuse. Pour cette raison, nous avons choisi de développer la plateforme MetaPredict pour faciliter la création automatique de modèles ADME-Tox.

Chapitre 3 : MetaPredict – plateforme automatique de création de modèles (Q)SAR pour la prédiction ADME-Tox

L'automatisation des tâches répétitives dans le processus laborieux de découverte de nouveaux médicaments a largement contribué à l'augmentation de la productivité scientifique (Ch1 1.3). Cette automatisation apporte plusieurs avantages tels que i) une capacité d'exploration élevée, ii) une qualité de recherche accrue en réduisant les erreurs, ii) un gain de temps significatif, et iii) une réduction des coûts. Dans cette ère où de grandes quantités de données sont produites chaque jour et où les ressources informatiques ne cessent d'évoluer, l'introduction d'apprentissage automatique, comme les approches (Q)SAR, a amélioré le processus de découverte de médicaments. De nombreuses applications réussies ont été rapportées dans la littérature. Elles attestent de l'importance de ces approches combinées aux méthodes traditionnelles pour relever les défis actuels de la chimie médicinale [237,238].

Les propriétés ADME-Tox indispensables lors d'un projet de *drug discovery* sont nombreuses. L'élaboration d'un modèle de prédiction pour chacune d'entre elles peut être un travail fastidieux et chronophage. Ceci implique la détermination des descripteurs moléculaires et des algorithmes les plus adaptés pour modéliser une propriété, mais également la création et la validation du modèle. Sachant que la construction d'un modèle (Q)SAR comporte des étapes répétitives et bien caractérisées, il est possible de les automatiser. L'automatisation de ces étapes critiques permet i) d'échantillonner un plus grand nombre de paramètres afin d'identifier les conditions optimales permettant d'obtenir un modèle de haute qualité, ii) de construire et valider rapidement des modèles pour diverses applications, et pour finir iii) de déployer rapidement les modèles créés au sein des équipes de recherche. D'autre part, l'accroissement des données ADME-Tox disponibles pour de nouvelles molécules offre de nouvelles perspectives pour construire des modèles couvrant des domaines d'applicabilité plus larges. Ainsi, l'utilisation d'une procédure standardisée et automatique facilite l'incorporation de nouvelles données aux modèles existants, dans le but d'obtenir des outils de prédiction toujours plus adaptés à nos besoins [239].

Au cours de la dernière décennie, plusieurs groupes de recherche ont essayé d'automatiser le processus de modélisation (Q)SAR. Plus récemment, Dixon *et al.* (Schrödinger) ont développé AutoQSAR, une application d'apprentissage automatique

visant à faciliter la création de modèles de prédiction [185]. Bien qu'AutoQSAR soit novateur, cet outil présente néanmoins plusieurs inconvénients : i) c'est un outil propriétaire qui nécessite une licence commerciale ii) Les étapes effectuées lors de la création du modèle ne peuvent pas être optimisées pour une application particulière. Ceci nous rend dépendant de la technologie proposée par AutoQSAR. iii) cet outil n'effectue pas toutes les étapes indispensables à la création d'un modèle, comme par exemple la normalisation des structures chimiques ou encore la préparation du jeu de données avant son utilisation. Une alternative intéressante pour la modélisation (Q)SAR entièrement automatisée est la plateforme OCoHEM, qui prend en considération l'ensemble des étapes indispensables à la création d'un modèle [240]. Cependant, son service en ligne rend son utilisation impossible pour une application sur des données privées et sensibles. Cox *et al.* ont proposé une autre alternative, nommée QSAR Workbench [241], qui permet d'exploiter les modèles construits à l'aide de Pipeline Pilot. Malheureusement, ce logiciel n'est pas librement accessible pour la majorité de la communauté scientifique [242]. Par ailleurs, l'utilisation des environnements de travail, comme Pipeline Pilot ou KNIME, présente des désavantages pour l'exploitation et la maintenance des modèles au long terme. Pour pallier l'ensemble de ces limitations, nous avons fait le choix de développer MetaPredict, notre plateforme automatique de création de modèles (Q)SAR pour la prédiction des paramètres ADME-Tox.

Dans le cadre de ce chapitre nous verrons dans un premier temps les étapes développées dans le cadre de la plateforme MetaPredict, puis dans un deuxième temps la stratégie mise en place pour l'élaboration de modèles de consensus, et pour finir la valorisation de la plateforme MetaPredict à l'aide d'outil dédiés à la compréhension des résultats pour les chimistes et la création d'un site internet visant à faciliter l'utilisation des modèles ADME-Tox produits grâce à notre méthodologie.

1. Développement de la plateforme MetaPredict

La plateforme MetaPredict a été conçue sur la base d'outils libres pour créer de manière **efficace** et fiable des modèles de prédiction locaux ou globaux. Lors de son développement, une attention particulière a été portée sur le respect des recommandations de l'OCDE et des bonnes pratiques (Q)SAR énoncées dans la littérature. Cette plateforme a pour objectif de créer automatiquement des modèles valides en échantillonnant plusieurs paramètres et d'apporter des outils facilitant la compréhension des résultats par les chimistes. Une vue d'ensemble des étapes

effectuées par la plateforme MetaPredict est présentée **Figure** 24. Chacune de ces étapes sera présentée plus en détails par la suite.

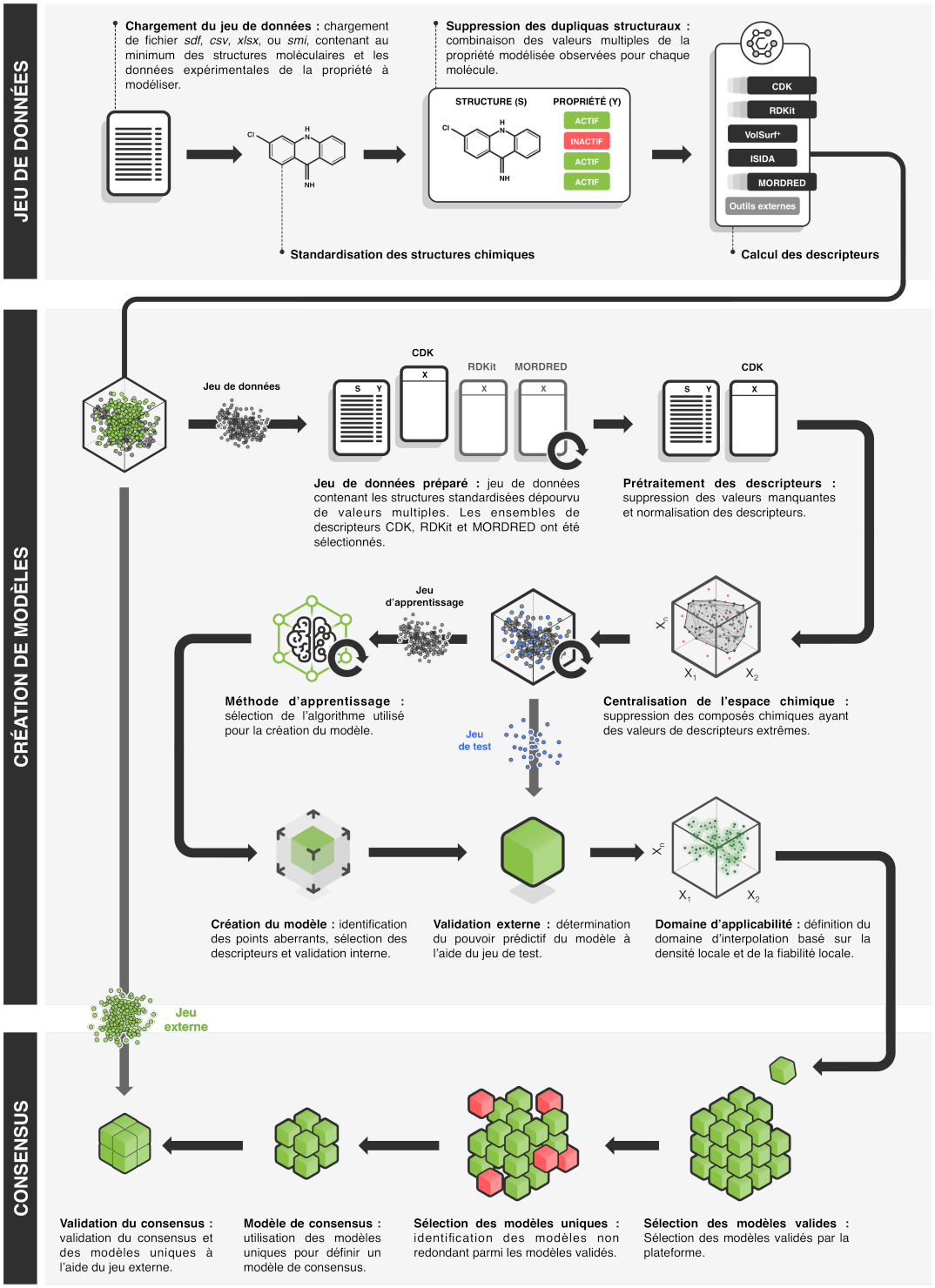


Figure 24 : Vue d'ensemble des étapes réalisées par la plateforme MetaPredict.

Les étapes principales de la plateforme sont présentées avec une brève description. Les étapes concernant les ensembles de descripteurs, la découpe du jeu de données ainsi que le choix de la méthode d'apprentissage sont itératives et sont représentées à l'aide d'une flèche cyclique.

Dans un premier temps, la plateforme charge le fichier contenant le jeu de données (au format *sdf*, *smi*, *csv* ou *xlsx*), qui doit au minimum disposer des structures moléculaires et les mesures expérimentales de la propriété à modéliser. Ce fichier peut également contenir des descripteurs moléculaires provenant d'autres outils non supportés par MetaPredict, afin de les utiliser dans le processus de modélisation. Une fois le jeu de données introduit dans la plateforme, une étape de préparation est effectuée pour i) standardiser les structures chimiques, ii) supprimer les valeurs multiples, iii) discrétiser les valeurs continues d'une propriété pour un modèle de classification et iv) calculer les descripteurs ou les empreintes moléculaires. Chaque ensemble de descripteurs ou d'empreintes va être employé individuellement pour l'étape de création de modèles (Q)SAR. Durant cette étape de création de modèles, plusieurs découpes de l'ensemble de données en jeu d'apprentissage et jeu de test, ainsi que de plusieurs algorithmes vont être explorées. Des modèles vont être validés et vont pouvoir être utilisés pour la création d'un modèle consensus.

Les étapes mises en place dans le cadre de la plateforme MetaPredict seront présentées individuellement. Chacune de ces étapes sera décrite à l'aide d'une explication de la méthodologie développée, puis illustrée à l'aide d'une application concrète sur les données de solubilité aqueuse. Le jeu de données de solubilité aqueuse a été choisi pour illustrer les étapes de la plateforme, car pour cette propriété nous disposons d'un nombre important de mesures et de nombreux modèles linéaires ont déjà été développés. Par la suite, nous présenterons d'autres propriétés peu modélisées comme la fraction libre dans le plasma (Fu_p).

1.1. Préparation du jeu de données

1.1.1. Jeu de données initial de solubilité aqueuse

La solubilité aqueuse est une propriété physico-chimique importante à optimiser lors de la conception de nouveaux médicaments, car elle est un des facteurs clés qui influencent la biodisponibilité. Il existe différentes manières de mesurer la solubilité aqueuse d'un composé [243]. Les données que nous avons exploitées font référence à la solubilité aqueuse intrinsèque qui traduit la concentration maximale d'un composé pouvant être dissout dans 100 mL d'eau. Seules les mesures obtenues à une température comprise entre 20 et 25 °C ont été exploitées. Ainsi, une concentration élevée indique une solubilité aqueuse importante du composé. Les mesures de solubilité aqueuse que nous avons exploitées ont été extraites de cinq sources différentes, à savoir : i) le jeu de

données publié par Delanay [244], ii) le jeu de données publié par Wang *et al.* [245], iii) le jeu de données publié par Huuskonen [246], iv) la base de données ADME/T database [247] et v) la base de données PhysProp [227]. Les valeurs extraites ont été exprimées en mol/L puis transformées selon le logarithme décimal afin d'obtenir une propriété LogS normalement distribuée (Figure 25).

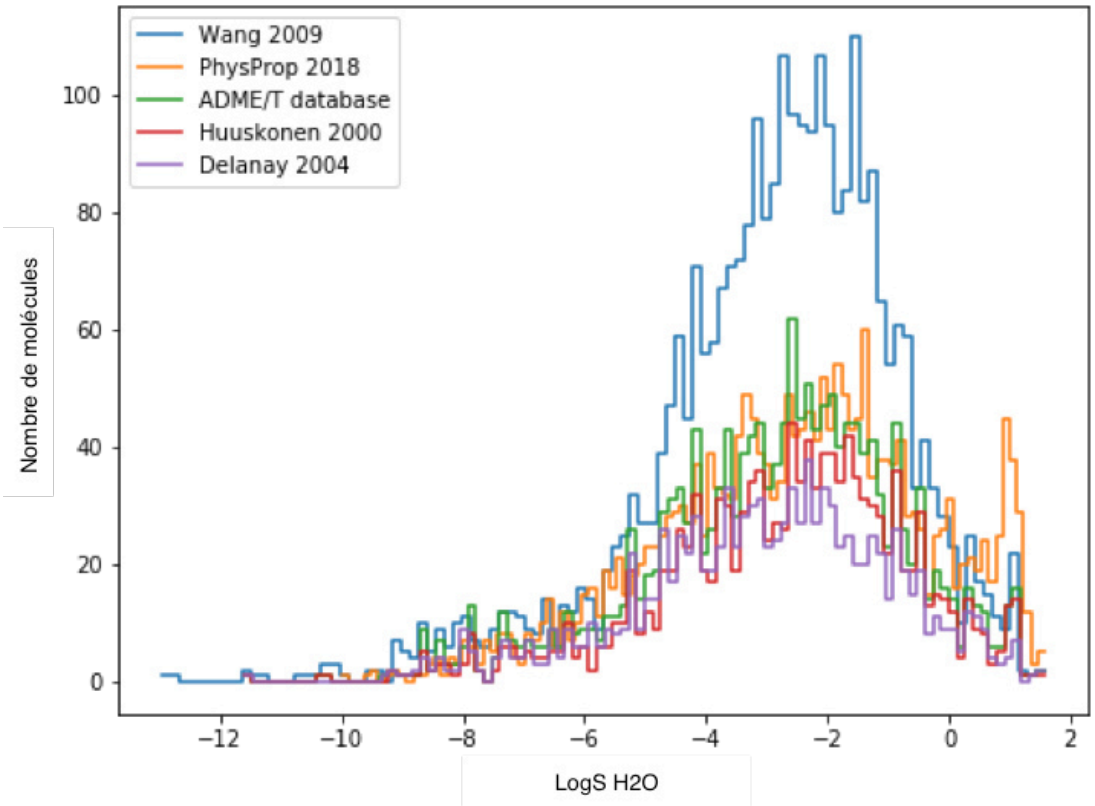


Figure 25 : Distribution du LogS pour les cinq jeux de données extraits.

Le but de la préparation du jeu de données est de vérifier manuellement que les données provenant de sources différentes peuvent être combinées pour la création d'un modèle de prédiction. Par conséquent, la suppression des dupliquas structuraux n'est pas effectuée lors de cette étape et sera réalisée ultérieurement par la plateforme.

Afin de vérifier que les données peuvent être combiné, nous avons calculé le coefficient de corrélation des mesures expérimentales de LogS pour les molécules communes entre les différents jeux de données. Le but est d'identifier les jeux de données suspects qui présentent un faible coefficient de corrélation avec ses congénères. La matrice de corrélation obtenue lors de la comparaison des cinq jeux de données est présentée Table 9.

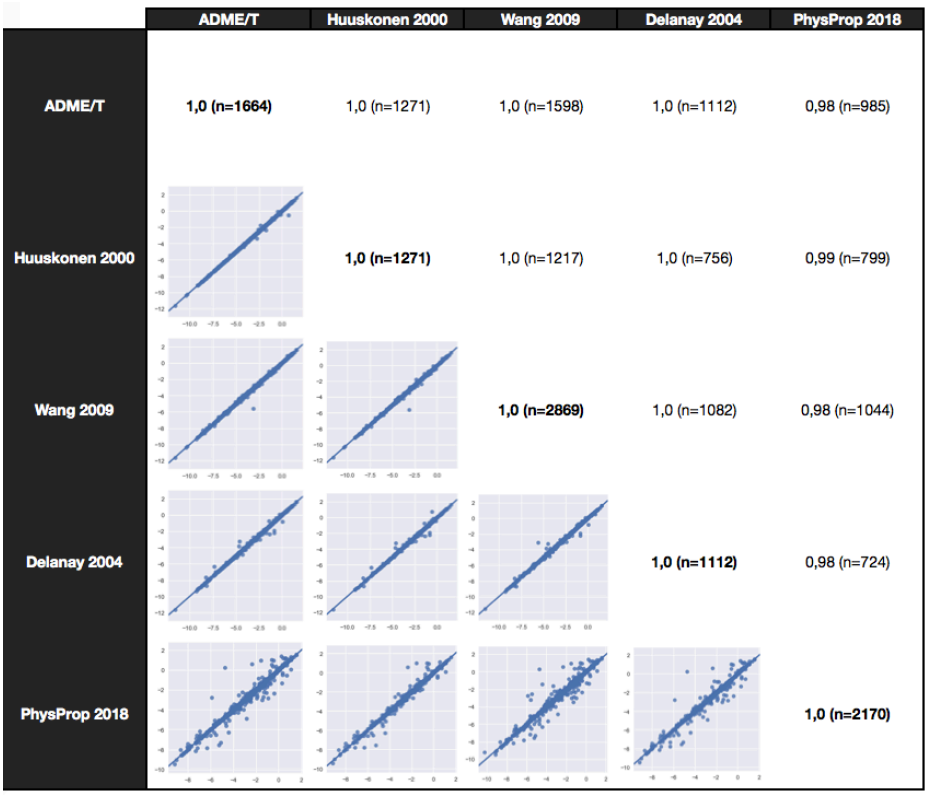


Table 9 : Matrice de corrélation obtenue en comparant les mesures expérimentales des cinq jeux de données LogS.

Chaque jeu de données a été préparé à l'aide d'une standardisation des structures chimiques (Ch3 1.1.2), suivie d'une suppression des dupliquas structuraux et des valeurs multiples (Ch3 0). Un coefficient de corrélation a ensuite été calculé à partir des mesures expérimentales des molécules communes (*n*) entre deux jeux de données. Le coefficient de corrélation est donc présenté avec le nombre de molécules communes (*n*) entre parenthèses.

Aucun coefficient de corrélation atypique n'est observé. Nous en concluons que l'utilisation de tous les jeux de données est possible pour la modélisation du LogS, car les coefficients de corrélation sont au voisinage de 1. On remarque néanmoins, que la base de données PhysProp 2018 extraite à l'aide de notre outil *WebScraper* apporte des mesures de LogS différentes des autres jeux de données. Cette observation est confirmée par la visualisation des régressions, qui nous montre également que les jeux de données ADME/T db, Delanay 2004 et Huuskonen 2000 sont hautement corrélés au jeu de données Wang 2009. Ceci est expliqué par le fait que Wang *et al.* ont proposé un modèle de régression basé en partie sur les données provenant des trois jeux de données mentionnés précédemment. Ainsi, cette analyse nous permet de vérifier que l'intégrité des données a été respectée. Au total, 10 150 mesures du LogS associées à des structures chimiques vont constituer le jeu de données qui va être préparé par la plateforme.

1.1.2. Standardisation des structures chimiques

Les structures chimiques extraites à partir de plusieurs sources de données peuvent ne pas être homogènes. Ainsi, des facteurs tels que l'ionisation des molécules, la forme tautomérique, ou encore les différences de représentations chimiques de certains groupes fonctionnels peuvent sévèrement affecter les performances d'un modèle de prédiction [248]. Afin de pallier ces inconvénients, l'étape de standardisation est indispensable pour représenter toutes les molécules selon le même référentiel (Figure 26).

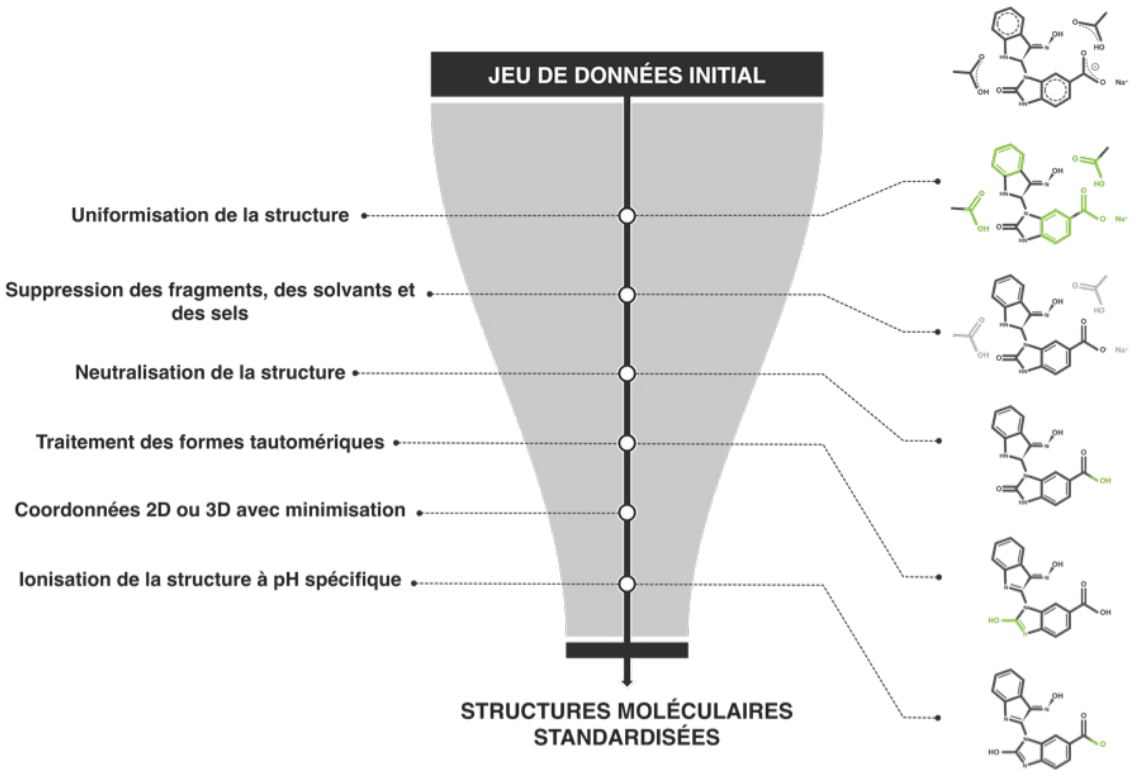


Figure 26 : Représentation des étapes effectuées lors de la standardisation des structures moléculaires.

La standardisation débute par l'uniformisation des représentations chimiques à l'aide de l'outil RDKit [130]. Pour cela, la structure moléculaire (1D, 2D ou 3D) est dans un premier temps convertie en SMILES isomérique. Le SMILES isomérique a été choisi, car il prend en considération la chiralité, l'état de protonation, les isotopes ainsi que l'isomérie d'un composé. Cette structure 1D est ensuite transformée en un objet moléculaire, ce qui permet de représenter toutes les molécules selon les mêmes normes. Cette conversion permet également de vérifier les structures chimiques et d'identifier les structures erronées afin de les supprimer du jeu de données. Les structures valides sont ensuite standardisées à l'aide de MolVS [249], en supprimant les fragments, les molécules de

solvant et les contre-ions présents dans les sels. Ceci est indispensable pour deux raisons : i) la plupart des outils dédiés au calcul des descripteurs ne sont pas équipés pour le traitement de substances chimiques ; ii) la présence de plusieurs entités moléculaires pour une même substance rend impossible l'identification et le traitement des doublons. En règle générale, un jeu de données doit contenir des structures uniques responsables de la mesure qui leur est associée. Ainsi, pour les substances qui contiennent plusieurs molécules, nous avons choisi de les rejeter, car il est impossible de déterminer quelle structure chimique est responsable de la valeur expérimentale obtenue de la propriété modélisée. Par ailleurs, une molécule peut avoir différentes formes tautomériques. Sachant que la tautomérie est dépendante de l'état d'ionisation d'une molécule, une neutralisation de la structure est réalisée avant l'étape d'énumération des tautomères. Seul le tautomère majoritaire est sélectionné pour représenter le composé. A cette étape de la standardisation, l'objet moléculaire ne dispose pas encore de coordonnées. Il est alors possible de définir des coordonnées 2D ou 3D. Dans le cas de coordonnées 3D, la structure est minimisée à l'aide du champ de forces MMFF94s [250]. La standardisation se finalise par l'ionisation des structures chimiques à l'aide de ChemAxon et ceci à un pH spécifique (par défaut 7,4) [251]. Cette étape de standardisation a été inspirée de l'outil de préparation de bases de données moléculaires VSPrep développé au sein de notre laboratoire [252].

Le jeu de données LogS a été préparé selon ce protocole. Suite à l'étape de standardisation, 26 structures chimiques erronées ont été identifiées et rejetées. Le jeu de données qui en résulte contient 10 124 mesures de LogS comprises entre -12,95 et 1,58 unité logarithmique. Nous souhaitons mettre en œuvre des modèles (Q)SAR 2D. Pour cette raison, seules les coordonnées 2D ont été définies pour les structures des composés restants. L'ionisation des structures n'a pas été effectuée dans le cadre de cet exemple, dans le but d'être dans les mêmes conditions que les modèles déjà publiés. Les structures standardisées associées aux mesures de LogS sont ensuite utilisées lors de l'étape de suppression des dupliquas structuraux.

1.1.3. Suppression des dupliquas structuraux

L'élaboration d'un modèle de prédiction suppose que chaque composé du jeu de données est unique. Cependant, il arrive parfois que des dupliquas structuraux soient présents dans l'ensemble de données, notamment lorsque plusieurs mesures sont disponibles pour un même composé. Ces mesures peuvent être identiques ou différentes. Plusieurs phénomènes peuvent être à l'origine de mesures différentes : i) soit la mesure

est fausse et a été introduite par inadvertance lors de l'enregistrement de la donnée, ii) soit l'étape de normalisation effectuée précédemment a engendré la création de plusieurs dupliquas structuraux (même composé actif dans différentes substances), ou iii) soit ce sont des réplicas expérimentaux correspondant à plusieurs mesures pour un même composé.

Afin de pouvoir combiner les mesures multiples d'un même composé, nous avons mis en place une méthodologie différente en fonction de la propriété modélisée. Pour des valeurs qualitatives de la propriété modélisée, si l'une d'entre elles ne concorde pas avec les autres, le composé est automatiquement supprimé. Pour des valeurs quantitatives de la propriété modélisée, la moyenne des valeurs est adoptée lorsque leur fluctuation est raisonnable. La fluctuation des valeurs est estimée à l'aide de l'Equation 16. Elle permet de comparer la variation des valeurs observées pour un composé sur la gamme de mesures de la propriété modélisée. Lorsque la fluctuation est supérieure à 5 %, les mesures correspondantes sont considérées comme suspectes et le composé est supprimé. Ceci limite l'intégration de mesures erronées dans le jeu de données exploité par le modèle.

$$f = \frac{y_{i,max} - y_{i,min}}{y_{max} - y_{min}}$$

Equation 16 : Estimation de la fluctuation des mesures pour l'identification de données suspectes.

Avec $y_{i,max}$ la valeur maximale des mesures de la propriété pour le composé i ; $y_{i,min}$ la valeur minimale des mesures de la propriété pour le composé i ; y_{max} la valeur maximale de la propriété observée dans le jeu de données ; y_{min} la valeur minimale de la propriété observée dans le jeu de données.

L'analyse des dupliquas structuraux dans le jeu de données LogS a permis d'identifier 1859 composés disposants de valeurs multiples. Pour chacun de ces composés, l'étendue des valeurs multiples a été comparée à celle observée dans le jeu de données (Figure 27.A). Les mesures de LogS ont été combinées à l'aide de la moyenne arithmétique pour 1701 composés, qui disposaient d'une fluctuation raisonnable, c'est-à-dire inférieure à 5 % de la gamme des mesures de LogS (Figure 27.B). 158 composés ont été supprimés en raison d'une fluctuation des mesures trop importante (Figure 27.C).

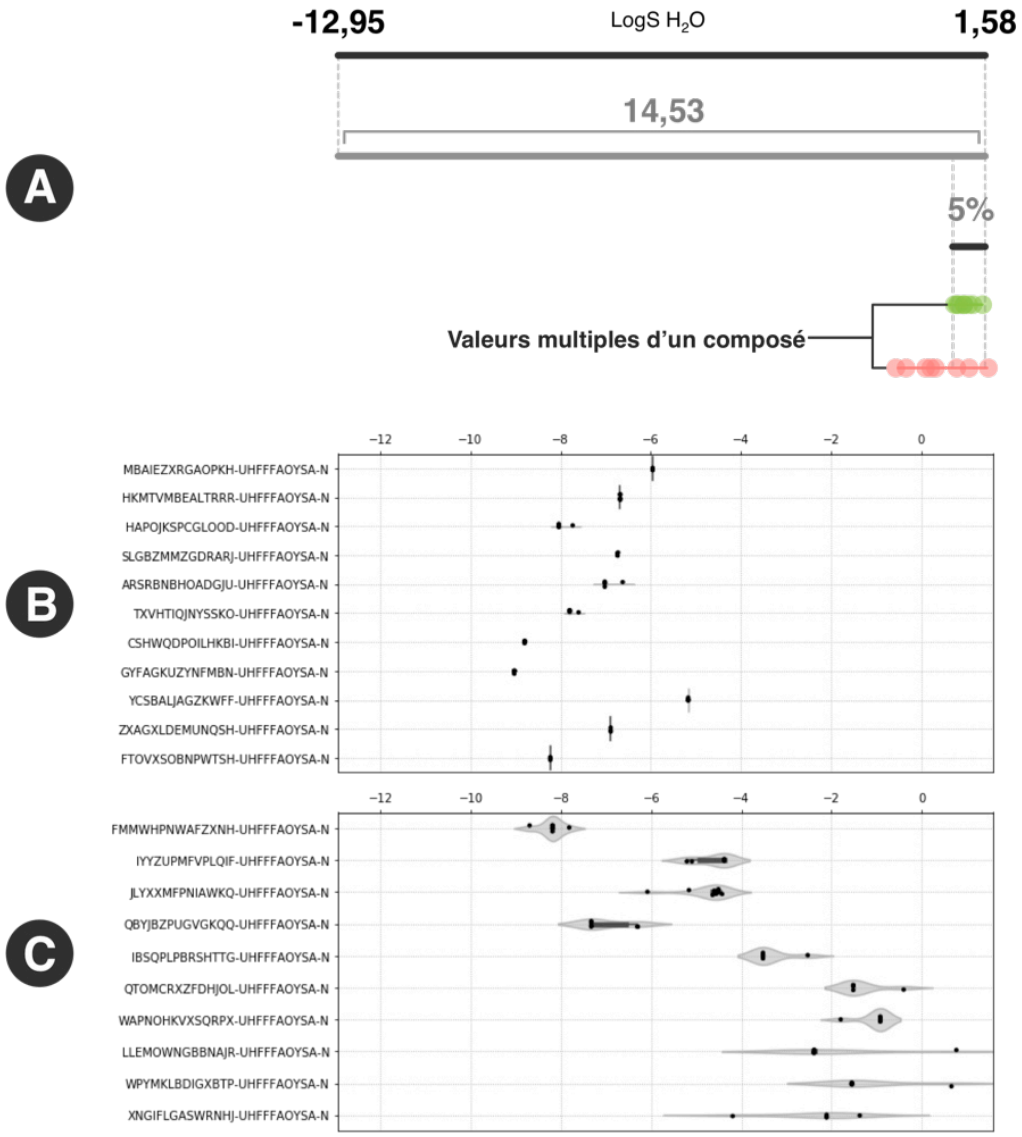


Figure 27 : Comparaison et suppression des valeurs multiples.

A) Représentation schématique de la méthodologie suivie et de l'application du seuil lié à la fluctuation des mesures de LogS. B) Représentation des valeurs multiples pour 10 composés ayant une fluctuation raisonnable. C) Représentation des valeurs multiples pour 10 composés ayant une fluctuation anormale.

Au final, le jeu de données contient 3 873 composés disposant chacun d'une valeur unique de LogS. Le calcul des descripteurs moléculaires peut à présent être effectué pour décrire les structures des composés étudiés.

1.1.4. Détermination des descripteurs ou des empreintes moléculaires

La plateforme MetaPredict supporte le calcul des descripteurs provenant des outils CDK, RDKit, MORDRED, ISIDA et VolSurf+. La liste complète des descripteurs calculés par la plateforme est présentée dans la Table 10.

OUTIL	DESCRIPTEUR	DIMENSION	NOMBRE
CDK	Constitutionnel	1D	17
	Topologique	2D	193
	Electronique	2D	33
	Géométrique	3D	20
	BCUT	3D	6
	WHIM	3D	17
RDKit	Constitutionnel	1D	111
	Topologique	2D	15
	Electronique	2D	16
	Type MOE	2D	58
	FCFP	2D	2048
	ECFP	2D	2048
	FCFC	2D	2048
	ECFC	2D	2048
	MACCS	2D	167
	HASH-AP	2D	2048
	HASH-TT	2D	2048
	RDK	2D	2048
MORDRED	Constitutionnel	1D	221
	Topologique	2D	389
	Electronique	2D	16
	BCUT	2D	24
	Auto-correlation	2D	606
	Géométrique	3D	214
	Type MOE	2D	355
ISIDA	IA25	2D	*
	IA25AP	2D	*
	IA25PCF	2D	*
	IB27	2D	*
	IAB26	2D	*
	IIA24	2D	*
	IIB24	2D	*
	IIAB24	2D	*
VolSurf+	Hybride	2D/3D	128

Table 10 : Liste des descripteurs calculés par la plateforme MetaPredict.

Cette table présente les outils intégrés à notre plate-forme avec le type, la dimension et le nombre de descripteurs. (*) Le nombre de descripteurs ISIDA dépend du nombre de fragments identifiés dans le jeu de données exploité.

Nous avons choisi de calculer les descripteurs moléculaires issus des outils CDK (247 descripteurs), RDKit (200 descripteurs), et MORDRED (1825 descripteurs) pour l'étude de la solubilité aqueuse. Les descripteurs calculés à l'aide d'un outil spécifique vont être considérés comme un ensemble indépendant. La plateforme va utiliser individuellement ces ensembles lors de l'élaboration de modèle de prédiction, afin de rechercher celui qui apporte les modèles de plus hautes performances.

1.2. Élaboration des modèles (Q)SAR

La création d'un modèle de prédiction comporte plusieurs étapes relatives i) au prétraitement du jeu de données, ii) à la sélection des variables explicatives (X) les plus appropriées pour modéliser la variable expliquée (Y), iii) à l'apprentissage du modèle, iv) à la validation du modèle et pour finir v) à la détermination du domaine d'applicabilité. La méthodologie employée pour chacune de ces étapes vous est présentée ci-après.

1.2.1. Prétraitement du jeu de données

1.2.1.1. Valeurs manquantes

Lors du calcul d'un descripteur, il est possible que la plateforme ne parvienne pas à définir une valeur numérique pour certaines structures chimiques, se traduisant par une valeur manquante. Ce phénomène peut être induit par i) une impossibilité du descripteur à capter l'information chimique d'une structure spécifique, ii) un descripteur erroné ne permettant pas de calculer convenablement les caractéristiques de nos composés, ou iii) une structure chimique erronée ne pouvant pas être exploitée convenablement par les outils de calcul de descripteurs. La présence de valeurs manquantes rend inutilisable le jeu de données pour les étapes de modélisation.

Pour pallier ce phénomène, certaines approches consistent à remplacer les valeurs manquantes d'un descripteur par la moyenne ou la médiane de ce dernier. Cependant, ceci ne respecte pas les recommandations (Q)SAR, selon lesquelles un descripteur doit être représentatif de l'information chimique qu'il reflète, et il doit transmettre une information chimique non erronée. Afin de respecter ces recommandations, nous avons traité les valeurs manquantes en deux temps. Dans un premier temps, les descripteurs pour lesquels les valeurs manquantes représentent plus de 1 % des valeurs totales sont considérés comme des descripteurs incomplets. Par conséquent, ils sont supprimés du jeu de données. Dans un second temps, les structures standardisées qui possèdent au moins une valeur manquante de descripteur sont rejetées, afin de ne pas incorporer de valeur erronée dans le jeu de données en leur attribuant une valeur arbitraire.

Dans le cas de la modélisation du LogS, le premier ensemble de descripteurs exploré par la plateforme MetaPredict est celui qui contient les descripteurs 2D provenant de l'outil CDK. Lors de l'étude des valeurs manquantes, 5 descripteurs sur 247 ont été considérés comme incomplets et 6 molécules sur 3 873 ont été supprimées.

Les descripteurs RPCG, RNCG, RPCS, RNCS et Kier3 ont été considérés par la plateforme comme étant des descripteurs incomplets. Nous nous sommes rendu compte que les valeurs manquantes pour les descripteurs RPCG, RNCG, RPCS et RNCS étaient concomitantes et concernaient 80 composés. Les descripteurs RPCS et RNCS sont respectivement calculés à l'aide des descripteurs RPCG et RNCG. Ces derniers traduisent le rapport entre le nombre de charges partielles positives/négatives et le nombre total de charges positives/négatives d'une structure chimique. Nous avons observé que les 80 composés incriminés étaient aliphatiques ou alicycliques (Figure 28.A). Par conséquent, la détermination de ces descripteurs est rendue impossible pour ces composés car ils ne possèdent aucune charge.

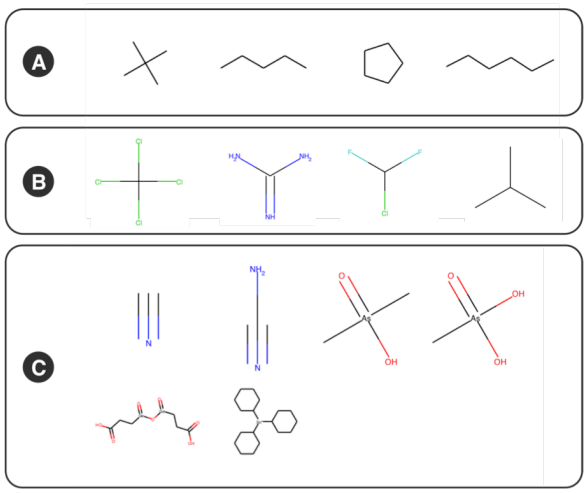


Figure 28 : Structures des molécules qui présentaient des valeurs manquantes selon certains descripteurs.

A) Représentation de 4 molécules pour lesquelles les valeurs des descripteurs RPCG, RNCG, RPCS et RNCS étaient manquantes. B) Représentation de 4 molécules pour lesquelles les valeurs du descripteur Kier3 étaient manquantes. C) Représentation des 6 structures supprimées du jeu de données par la plateforme MetaPredict.

Une analyse similaire pour le descripteur Kier3 nous a montré que les valeurs manquantes étaient observées pour 43 composés qui pouvaient être divisés en deux groupes, à savoir : i) des composés de trois atomes présentant une hybridation sp2, et ii) des composés de quatre atomes d'hybridation sp3 (Figure 28.B). Il existe trois descripteurs de *Kier* topologiques qui ont pour but de saisir différents aspects de la forme moléculaire. Le calcul de ces descripteurs prend en considération le nombre d'atomes lourds qui définissent le graphique moléculaire, le nombre de liaisons observées entre les atomes lourds et une longueur de chemin variable entre 1 et 3. Au final, les descripteurs Kier1, Kier2 et Kier3 sont obtenus pour les longueurs de chemin de 1, 2 et 3 respectivement. Dans notre cas, seul le descripteur Kier3 possède des valeurs

manquantes. Ce descripteur, basé sur des chemins de longueur 3, ne peut tout simplement pas être calculé pour les 43 molécules incriminées, car elles possèdent toutes une longueur de chemin maximale équivalente à 2.

Les molécules présentées dans la Figure 28.C ont été supprimées. Parmi ces molécules, l'isocyanide et la cyanamide ont été rejetées à cause de valeurs manquantes pour le descripteur HybRatio. Ce descripteur calcule le rapport entre le nombre d'atomes de carbone d'hybridation sp3 et le nombre d'atomes de carbone d'hybridation sp3 et sp2. Comme ces deux molécules possèdent une hybridation sp le calcul du descripteur HybRatio est irréalisable. Cependant, comme ces molécules ne représentent que 0,05 % du nombre total de valeurs proposées par ce descripteur, ce dernier n'a pas été jugé comme incomplet par la plateforme. Les quatre molécules restantes sont des composés organométalliques qui disposent de valeurs manquantes pour l'ensemble des descripteurs calculés et ceci quelque soit l'outil. Ce type de composé n'est pas traité par la standardisation des molécules présentée précédemment. Par conséquent, le traitement des composés organométalliques pourrait faire l'objet d'une perspective d'amélioration de notre étape de préparation des structures chimiques.

En résumé, nous venons de voir que la plateforme MetaPredict a été capable d'identifier les 5 descripteurs incomplets et elle a permis de retirer 6 composés qui disposaient de particularités structurales dont les valeurs des descripteurs moléculaires ne permettaient pas l'élaboration d'un modèle de prédiction. Au final, le jeu de données contient 3 867 molécules décrites par 242 descripteurs CDK. Ces descripteurs dépourvus de valeurs manquantes sont ensuite normalisés.

1.2.1.2. Normalisation des descripteurs

Les descripteurs possèdent des échelles de grandeurs variables. Par exemple, un descripteur peut représenter une masse moléculaire tandis qu'un autre peut être sous la forme d'un comptage (descripteurs constitutionnels). L'apprentissage d'un algorithme (Ch1 3.2.3) peut être biaisé si l'ordre de grandeur d'un descripteur moléculaire est nettement supérieur à celui des autres. Ainsi, ce descripteur peut dominer la fonction objective et rendre l'algorithme incapable d'apprendre correctement des autres descripteurs. L'objectif de l'étape de normalisation est de transformer l'ensemble des descripteurs afin de les exprimer selon une échelle de mesure comparable. Deux méthodes sont principalement utilisées pour répondre à cette problématique.

La standardisation Z_{score} d'un descripteur consiste à centrer ses valeurs numériques sur la moyenne puis à les réduire en fonction de l'écart-type (Equation 17). Les descripteurs une fois transformés auront tous une moyenne centrée sur 0 et une déviation standard de 1. L'avantage de la standardisation est qu'elle permet de préserver les tendances de chaque individu vis-à-vis du descripteur transformé.

$$Z_{score} = \frac{x_i - \bar{x}}{\sigma}$$

Equation 17 : Standardisation selon le Z_{score} d'un descripteur.

Avec x_i la valeur du descripteur x pour l'individu i, $\bar{x}$ la moyenne sur tous les x_i et σ l'écart-type du descripteur x.

La normalisation *MinMax* d'un descripteur consiste à borner ses valeurs numériques entre un minima et un maxima. Généralement, les descripteurs transformés possèdent tous une échelle comprise entre 0 et 1 selon l'Equation 18. L'avantage de cette normalisation est qu'elle préserve la gamme de mesure de l'activité prédite. L'inconvénient est qu'elle peut atténuer l'information sur les potentielles valeurs aberrantes présentes dans l'ensemble de données.

$$MinMax = \frac{x_i - x_{min}}{x_{max} - x_{min}}$$

Equation 18 : Normalisation selon le *MinMax* d'un descripteur.

Avec x_i la valeur du descripteur x pour l'individu i, x_{max} la valeur maximale observée dans la gamme des x_i et x_{min} la valeur minimale observée dans la gamme des x_i.

La normalisation *MinMax* est utile lorsque les descripteurs ont des échelles variables et que la méthode d'apprentissage utilisée ne fait pas d'hypothèses sur la distribution des descripteurs, comme par exemple dans le cas des approches de classification. A contrario, la normalisation Z_{score} est utile lorsque les données ont des échelles variables et que la méthode d'apprentissage nécessite que les descripteurs aient une distribution gaussienne, comme par exemple dans le cas des modèles de régression.

Pour la prédiction du LogS nous souhaitons en priorité mettre en œuvre des modèles de régression puis dans un second temps des modèles de classification. Par conséquent, nous avons choisi de normaliser les descripteurs selon la méthode du Z_{score}.

1.2.1.3. Descripteurs de variance quasi-nulle ou de variance nulle

Si la variance d'une variable est nulle, alors cette variable est une constante. Ainsi, des descripteurs de variance quasi-nulle ou de variance nulle apportent une information

peu significative pour expliquer les tendances sous-jacentes qui existent dans le jeu de données. En d'autres termes, ces descripteurs sont peu appropriés pour modéliser la propriété souhaitée. La méthodologie que nous avons mise en place pour identifier ce type de descripteurs repose sur le principe énoncé par *Khun*. Un descripteur de variance quasi-nulle signifie que la fraction des valeurs uniques est inférieure à 10 % par rapport au nombre total de valeurs, mais également que le rapport entre la fréquence de la valeur la plus répandue et la fréquence de la deuxième valeur la plus répandue est supérieur ou égal à 20 [253]. Afin de mieux illustrer ce concept, prenons l'exemple proposé Figure 29.

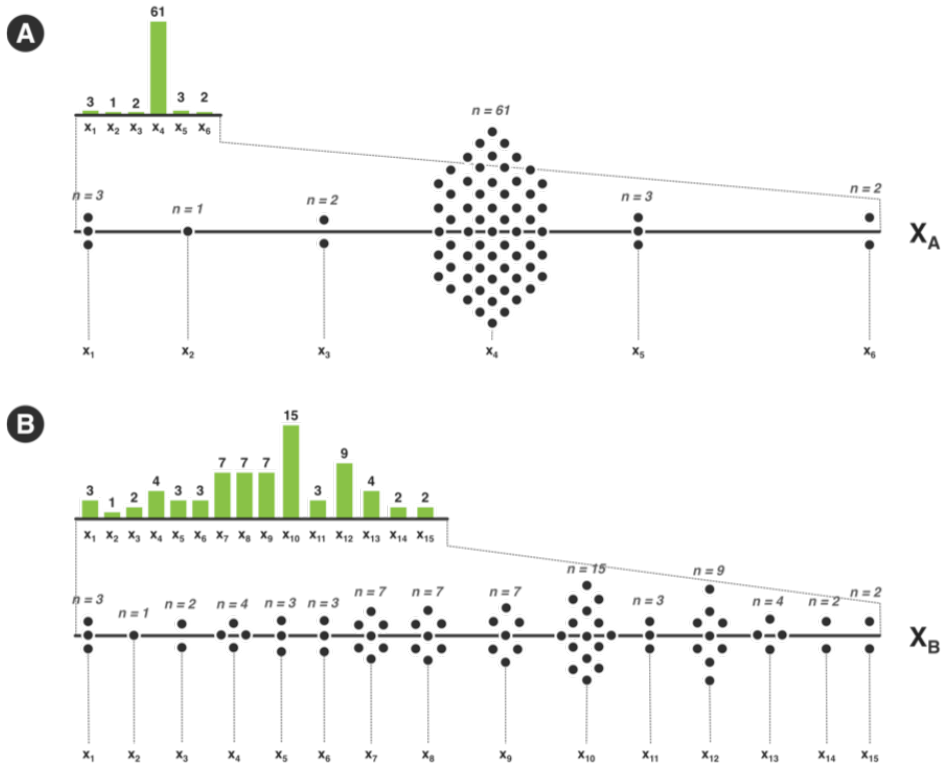


Figure 29 : Représentation schématique de la variance quasi-nulle selon *Khun*.

A) Descripteur X_A de variance quasi-nulle. ; B) Descripteur X_B de variance non nulle. ; Pour chaque descripteur les points noirs représentent les individus qui ont une même valeur numérique x_n. L'histogramme (en vert) représente les fréquences de chaque valeur x_n.

Le descripteur X_A contient au total 72 valeurs numériques, mais seulement 6 valeurs sont uniques. Par conséquent, la fraction des valeurs uniques est égale à 8,33 % (6/72), c'est-à-dire inférieure à 10 %. Ce descripteur est donc de variance quasi-nulle. Le rapport des fréquences des valeurs prépondérantes permet d'affiner l'analyse, afin de déterminer si la distribution de fréquence de la variable est asymétrique. Dans le cas du descripteur X_A, ce rapport vaut 20,33 (61/3) et indique un déséquilibre significatif dans la fréquence des valeurs comme représenté sur la Figure 29.A. Pour l'exemple du descripteur X_B, la fraction des valeurs uniques vaut 21 % (15/72) et le rapport des fréquences

prépondérantes est égal à 1,67 (15/9). Le descripteur X_B est considéré comme étant de variance non nulle et dispose d'un équilibre dans la fréquence des valeurs. Cette approche est très efficace pour identifier les descripteurs n'apportant aucune information chimique pertinente. Cependant, elle ne peut pas être appliquée sur des données catégoriques comme par exemple dans le cas des empreintes moléculaires. Pour ces dernières, seuls les *bits* de variance nulle ont été supprimés.

Cette approche a été utilisée pour identifier et traiter les descripteurs de variance quasi-nulle et de variance nulle du jeu de données LogS. Comme représentés Figure 30, 145 descripteurs disposent d'un pourcentage de valeurs uniques inférieur à 10 % et 29 descripteurs ont été rejetés, car ils avaient un rapport des fréquences prépondérantes supérieur au critère d'acceptabilité. Au total, 68 descripteurs CDK ont été sélectionnés sur les 242 que comptait le jeu de données.

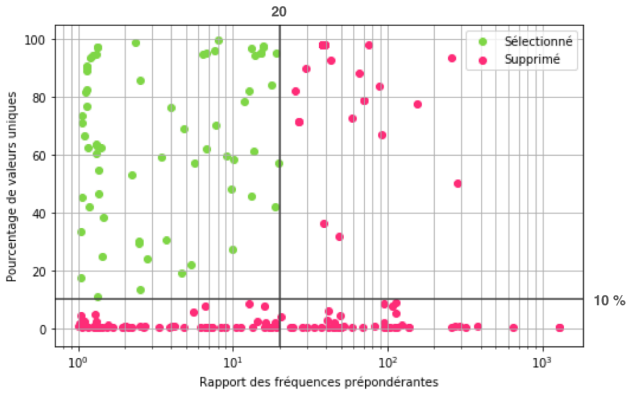


Figure 30 : Identification des descripteurs de variance quasi-nulle.

Ce graphique représente le pourcentage de valeurs uniques en fonction du rapport des fréquences prépondérantes pour l'ensemble des descripteurs CDK. Les seuils grâce auxquels les descripteurs ont été sélectionnés sont représentés en gris (20 et 10).

1.2.1.4. Centralisation de l'espace chimique

Pour un ensemble de descripteurs étudié, la plateforme va focaliser l'espace chimique couvert par le jeu de données. L'objectif de cette étape est de retirer l'ensemble des composés satellites, c'est-à-dire les molécules qui définissent les limites de l'espace chimique connu. Le but est alors de retirer de manière non invasive l'ensemble des individus qui présentent des valeurs de descripteurs extrêmes. Cette étape est indispensable, car ces individus vont rendre certaines analyses complexes, comme par exemple la détermination des descripteurs corrélés. Pour identifier les individus de valeurs extrêmes, le jeu de données est représenté dans un espace ACP, car cette

méthode permet d'appréhender la colinéarité des descripteurs [254]. A l'aide des coordonnées des individus, le centroïde moyen du nuage de point est identifié. La distance euclidienne de tous les individus au centroïde permet ensuite de retirer 5 % des composés les plus éloignés (satellites). Ceci permet de définir un sous-espace chimique de plus haute densité, c'est-à-dire contenant 95% des composés du jeu de données initial.

Cette méthode a été appliquée au jeu de données de LogS et les résultats obtenus sont présentés (Figure 31). On remarque que le rejet de 193 molécules satellites permet de retirer les individus extrêmes observés pour la majorité des descripteurs. Au final, l'espace chimique exploité pour la création des modèles de prédiction comporte 3 674 composés.

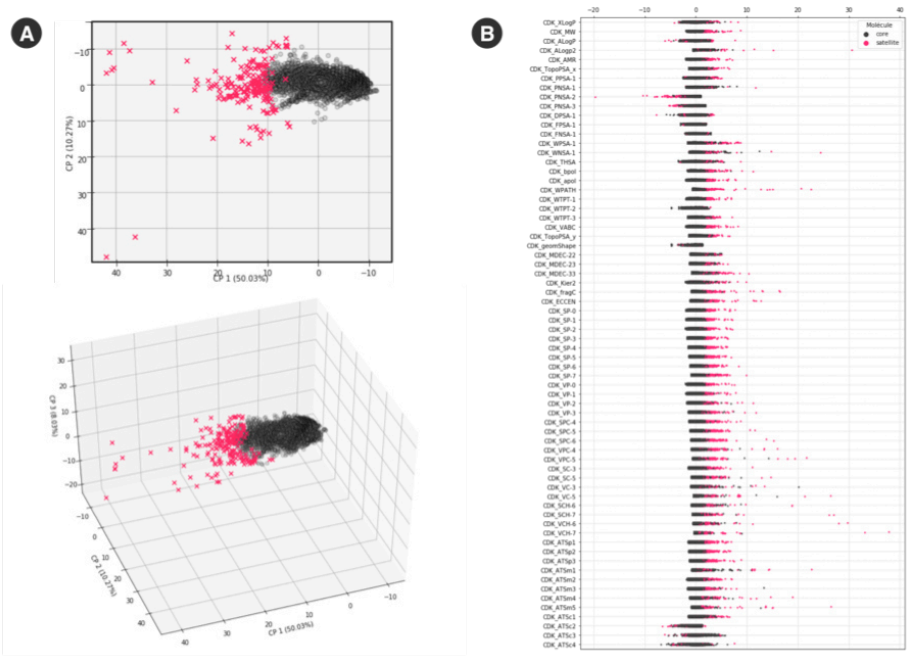


Figure 31 : Centralisation de l'espace chimique pour le jeu de données LogS.

A) Espace ACP 2D (PC1, PC2) et 3D (PC1, PC2, PC3) représentant les molécules satellites (rouge) et les molécules conservées (gris). La variance expliquée des 2 et 3 premières composantes est respectivement de 60,3 % et 68,3 %. B) Répartition des descripteurs en présence des molécules satellites (rouge).

1.2.1.5. Discrétisation des valeurs continues de la propriété modélisée

La discrétisation permet de transformer des valeurs continues de la propriété en valeurs discrètes dans le but de pouvoir générer des modèles de classification. Cette discrétisation est effectuée à l'aide d'un seuil selon lequel les valeurs continues vont être divisées en deux groupes de composés. Un deuxième seuil peut être appliqué pour créer une zone de délétion. La zone de délétion consiste à supprimer les composés qui ont une valeur de la propriété intermédiaire entre les deux classes. Ceci facilite la compréhension

du jeu de données par un modèle de classification, et contribue à la réduction du nombre de faux positifs prédits. Généralement, le seuil attribué pour la création de la zone de délétion est équivalent à l'erreur expérimentale de la propriété modélisée.

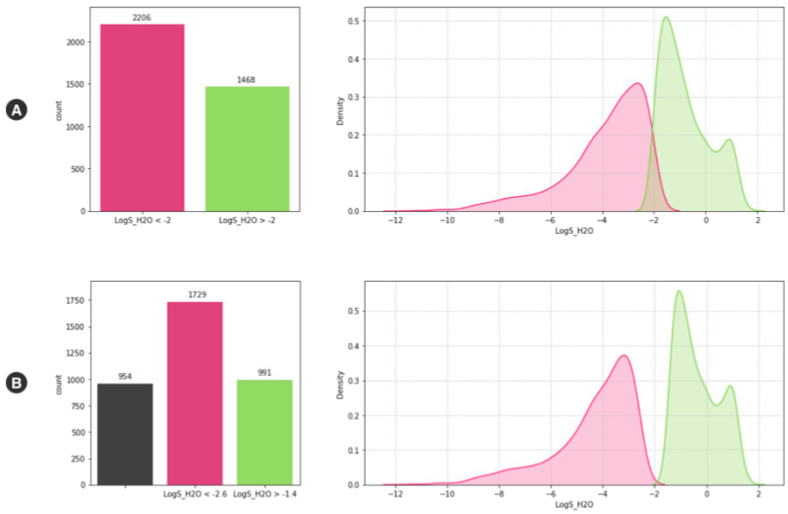


Figure 32 : Discrétisation des valeurs continues du LogS.

La répartition des classes est représentée à gauche. La distribution des classes est représentée à droite. La couleur rouge représente la classe inactive (LogS < seuil), la couleur verte représente la classe active (LogS > seuil) et la couleur grise représente les individus écartés lors de la création de la zone de délétion. A) Représentation des classes issues de l'étape de discrétisation à l'aide du seuil égale à -2. B) Représentation des classes issues de la création d'une zone de délétion entre les seuils -2,60 et -1,40.

Notre objectif lors de la modélisation du LogS est de pouvoir prédire par la suite si une molécule est soluble ou non dans les milieux aqueux. Les mesures de LogS que nous avons à notre disposition sont continues. Par conséquent, il est indispensable d'utiliser un seuil à partir duquel nous sommes capable de séparer les molécules solubles et les molécules insolubles, afin de définir des classes permettant de créer un modèle de classification. En générale un composé est dit insoluble dans les milieux aqueux lorsque sa solubilité est inférieure à 10 mM. Par conséquent, le seuil permettant de discrétiser les mesures de LogS est égale à $\log_{10}(0{,}01)$, soit -2 en unité logarithmique (Figure 32.A). Il est à noter que nous souhaitons limiter la prédiction de faux positifs, c'est-à-dire des molécules prédites comme étant solubles mais qui en réalité ne le sont pas. Pour cela, il est possible de créer la zone de délétion en connaissant l'erreur expérimentale. L'erreur expérimentale pour l'étude de la solubilité a été estimée à 0,6 unité logarithmique par Jorgensen et Duffy [255]. Nous avons donc utilisé deux seuils égaux à -1,40 (-2 + 0,6) et à -2,6 (-2 - 0,6) afin de créer une zone de délétion permettant de limiter l'impact des faux positifs dans le modèle (Figure 32.B).

On remarque que la zone de délétion permet de mieux séparer les deux classes créées. La création de cette zone implique cependant le rejet de 954 composés du jeu de données initial. Au final, le jeu de données utilisé pour générer des modèles de classification contient au total 2720 composés répartis en deux classes, à savoir une classe active nommée « LogS_H2O > -1,4 » qui contient 991 composés, et une classe inactive nommée « LogS_H2O < -2 » qui contient 1729 composés. Il est important de noter que le jeu de données est déséquilibré avec un rapport actif/inactif de 0,57.

Nous venons de voir les étapes de prétraitement d'un jeu de données effectuées par la plateforme MetaPredict. Le jeu de données correspondant à l'ensemble CDK contenait initialement 3873 molécules décrites par 247 descripteurs. Suite aux étapes de suppression des valeurs manquantes, de suppression des descripteurs de variance quasi-nulle et de centralisation de l'espace chimique, le jeu de données est composé de 3674 molécules décrites par 68 descripteurs CDK. Ce dernier sera employé pour la création de modèles de régression. Ce jeu de données a ensuite été discrétisé afin de pouvoir créer des modèles de classification. Le jeu de données utilisé pour les modèles de classification contient 2720 composés.

1.2.1.6. Découpe du jeu de données

L'objectif est d'obtenir un jeu d'apprentissage et un jeu de test qui ont une répartition de la propriété ainsi qu'un sous-espace chimique représentatifs du jeu de données initial. La méthode mise en place dans le cadre de la plateforme MetaPredict vise à vérifier que les sous-ensembles générés respectent les exigences énoncées précédemment. Dans un premier temps, une découpe stratifiée du jeu de données initial est effectuée à l'aide de *Scikit-Learn* [256]. La stratification est une sélection basée sur la propriété (Y). Pour des valeurs discrètes de la propriété, cette stratification permet d'obtenir un rapport Actif/Inactif similaire entre les deux sous-ensembles. Pour des valeurs continues de la propriété, la stratification permet d'obtenir une distribution équivalente entre les deux sous-ensembles. Dans un deuxième temps, nous avons souhaité vérifier que les sous-ensembles créés disposaient d'un espace chimique commun (X). Pour cela, le jeu d'apprentissage et le jeu de test sont représentés dans un espace ACP. Le pourcentage de recouvrement entre les enveloppes convexes de ces deux sous-ensembles est calculé. Si le recouvrement est supérieur à 90 %, le jeu de test est considéré comme représentatif du jeu d'apprentissage, sinon la découpe du jeu de données initial est réitérée jusqu'à ce que ces exigences soient atteintes. Afin d'augmenter nos chances de découvrir des modèles ayant un pouvoir prédictif élevé, la

plateforme effectue 30 découpes du jeu de données, où 75 % des composés sont affectés au jeu d'apprentissage et les 25 % restants au jeu de test. Pour chacune de ces découpes, plusieurs méthodes d'apprentissage vont être testées.

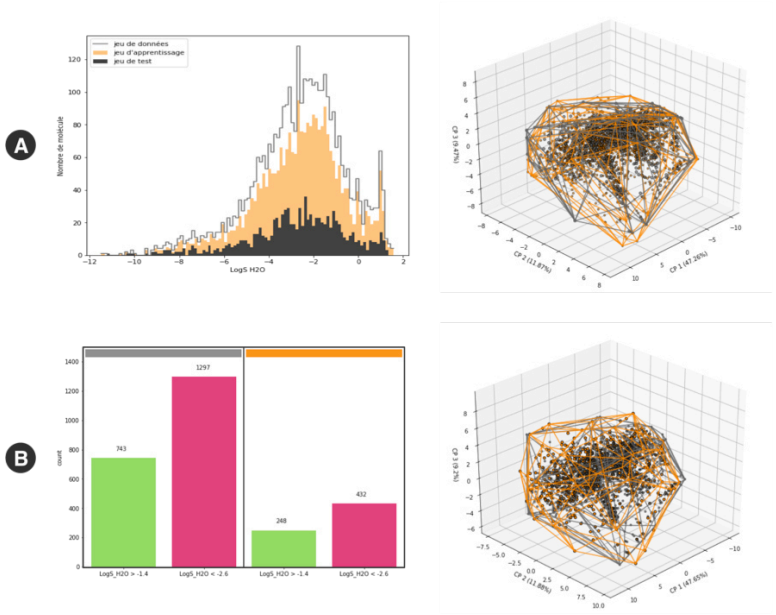


Figure 33 : Répartition du jeu d'apprentissage et du jeu de test.

La couleur orange représente le jeu d'apprentissage et la couleur noire représente le jeu de test. A) Découpe du jeu de données dans le cas d'un modèle de régression. A gauche vous est présentée la distribution de la propriété continue dans les deux sous-ensembles de données produits. A droite, le recouvrement des espaces chimiques couverts par le jeu d'apprentissage et le jeu de test. B) Découpe du jeu de données dans le cas d'un modèle de classification. A gauche vous est présentée la répartition des classes dans les deux sous-ensembles. A droite, le recouvrement des espaces chimiques couverts par le jeu d'apprentissage et le jeu de test.

Les résultats obtenus pour la création d'un modèle de régression et d'un modèle de classification pour le LogS vous sont présentés par la Figure 33. Dans le cas de la découpe du jeu de données pour un modèle de régression (Figure 33.A), on remarque que la distribution du jeu de test couvre bien toute la gamme de LogS explorée par le jeu d'apprentissage. De plus, les distributions des deux sous-ensembles créés sont représentatives du jeu de données initial. Dans le cas d'une découpe du jeu de données destinée à la création d'un modèle de classification, on remarque que l'équilibre des classes est identique entre le jeu d'apprentissage et le jeu de test. Le rapport actif/inactif que nous avions observé dans le jeu de données initial était de 0,57. Ce même rapport entre les deux classes est observé dans le jeu d'apprentissage et le jeu de test. D'autre part, la représentation des espaces chimiques couverts par les jeux d'apprentissage et les jeux de test, nous montre que dans les deux cas les sous-ensembles créés couvrent

un espace chimique commun. Cette observation est extrêmement importante, car cette vérification nous permet de nous assurer que le jeu de test est bien approprié pour valider les modèles créés sur la base du jeu d'apprentissage.

Le jeu d'apprentissage déséquilibré peut biaiser un modèle de classification. Pour cette raison, nous verrons ci-après comment le jeu d'apprentissage est équilibré pour pallier à cette limitation. Il est à noter que le jeu de test présente le même déséquilibre. Cependant, nous avons choisi de ne pas le modifier dans le but de respecter les bonnes pratiques (Q)SAR. Nous avons considéré que l'utilité du jeu de test était de pouvoir tester en conditions réelles le modèle généré, c'est-à-dire de voir s'il est capable d'apporter des prédictions justes pour un ensemble de données représentatif du jeu de données initial.

1.2.1.7. Equilibre des classes

Pour certains jeux de données, pour lesquels la propriété modélisée est discrète, la répartition des classes peut être non équilibrée. Les modèles de classification créés à partir d'ensembles de données non équilibrées peuvent présenter un biais. A titre d'exemple, si un modèle est basé sur un jeu de données contenant 10 actifs pour 100 inactifs, le modèle sera plus enclin à prédire des inactifs que des actifs. L'approche que nous avons mise en place est basée sur la méthode de regroupement *k*-moyennes (Ch1 3.2.3.1.b)). L'objectif est d'équilibrer les classes, c'est-à-dire supprimer des composés de la classe majoritaire, sans modifier le recouvrement de l'espace chimique. La classe minoritaire est identifiée afin de déterminer le nombre de composés à extraire dans la classe majoritaire. Le regroupement *k*-moyennes est ensuite appliqué à la classe majoritaire pour extraire des composés afin d'obtenir le même nombre de composés que la classe minoritaire. Ce regroupement va fournir plusieurs groupes définis par des centroïdes. Ainsi, le composé le plus proche du centroïde est sélectionné comme étant le représentant de chaque groupe.

La Figure 34 présente la répartition des classes avant et après avoir équilibré le jeu d'apprentissage, mais également les sous-espaces chimiques représentés par chacune d'entre elles. Le jeu d'apprentissage a été équilibré en sélectionnant 743 composés parmi les 1297 composés de classe inactive, qui est la classe majoritaire. On remarque que le recouvrement des sous-espaces chimiques couvert par chacune des classes a été préservé.

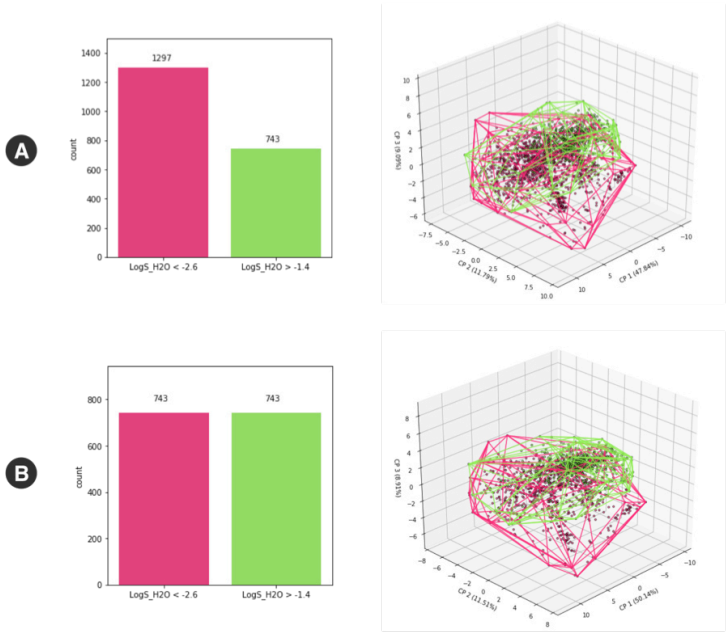


Figure 34 : Equilibre des classes active et inactive.

A gauche est représenté le nombre de composés par classes. A droite est représenté l'espace ACP avec les enveloppes convexes représentants chacune une des deux classes du jeu de d'apprentissage. La couleur verte est attribuée à la classe active (LogS H_2O > -1,4) et la couleur rouge est attribuée à la classe inactive (LogS H_2O < -2,6). A) Représentation du jeu d'apprentissage avant l'équilibre des classes. B) Représentation du jeu d'apprentissage après l'équilibre des classes.

Nous venons de voir l'ensemble des étapes de prétraitement spécifiques à un ensemble de descripteurs et qui sont indispensables pour l'élaboration de modèles (Q)SAR. Le jeu de données qui en résulte va être utilisé pour les étapes de modélisation présentées ci-après. Ainsi, pour un ensemble de descripteurs, plusieurs découpes du jeu de données vont être réalisées pour lesquelles plusieurs algorithmes vont être utilisés.

1.2.2. Choix de l'algorithme

En fonction de l'ensemble de descripteurs étudié et de l'espace chimique exploré, il est possible de générer des modèles à l'aide de différents algorithmes. La plateforme MetaPredict donne l'opportunité de créer des modèles de régression et de classification en fonction des algorithmes présentés dans la Table 11.

ALGORITHME	TYPE	PARAMÈTRE
PLS	R	n_comp : nombre de composantes
MLR	R	
SVM	R/C	kernel : type de noyaux C : courbure des plans supports ε : largeur de la marge
RF	C	mtry : nombre de descripteurs par arbre n_estimators : nombre d'arbres
LGR	C	

Table 11 : Liste des méthodes d'apprentissage utilisées par la plateforme.

Les algorithmes sont présentés avec le type de modèle pour lesquels ils sont utilisés (R : Régression et C : Classification), ainsi que les paramètres optimisés.

Avant toutes étapes de modélisation, il est indispensable de paramétrer le modèle en fonction du jeu d'apprentissage étudié. Pour cela, une recherche par grille a été adoptée pour obtenir un modèle optimisé. Cette recherche consiste à tester toutes les combinaisons possibles des différents paramètres du modèle que nous souhaitons explorer. Une validation interne permet d'estimer la qualité du modèle pour chaque combinaison de paramètres testée. Cette étape indispensable présente néanmoins un inconvénient majeur. Le temps nécessaire à l'identification des conditions optimales augmente exponentiellement lorsque le nombre de paramètres et le nombre de combinaisons à explorer augmentent.

Pour la suite des explications, nous allons prendre l'application des machines à vastes marges (SVM). Cette méthode d'apprentissage nécessite de définir trois paramètres à savoir la courbure (C) de la marge, la largeur (ε) de la marge, ainsi que le noyau utilisé (*kernel*). Dans le cas d'un modèle de régression, l'utilisation des paramètres par défaut (C = 1 ; ε = 0,1 ; kernel linéaire) nous apporte des performances contestables pour le modèle de prédiction du LogS. Ainsi, le coefficient de détermination observé pour le jeu d'apprentissage (R^2) est de -0,04, et le coefficient de validation croisée (Q^2) vaut -0,44. Par conséquent, on en déduit que les paramètres par défaut ne sont pas adaptés pour l'étude du jeu d'apprentissage. L'exploration des paramètres du modèle consiste a tester toutes les combinaisons de 1 à *n*, entre C ($C_{n+1} = C_n * 10$ avec C compris entre 0,001 et 10), ε ($\varepsilon_{n+1} = \varepsilon_n * 10$ avec ε compris entre 0,001 et 1), et les *kernels* (linéaire, sigmoïdal ou polynomial), afin d'identifier les paramètres permettant d'obtenir les meilleurs performances. L'identification du paramètre optimum est obtenue lorsque C vaut 0,01, ε est égal à 0,1 et lorsque le *kernel* est linéaire. Le modèle qui en résulte possède un R^2 de 0,75 et un Q^2 de 0,74. Le pouvoir prédictif du modèle (R^2f), estimé à partir du jeu de test, a été amélioré de 79 points en passant de -0,02 à 0,77 après le paramétrage du modèle

(Figure 35). Il est important de noter que les performances sur le jeu de test ne sont pas utilisées pour la sélection des paramètres optimum, mais est donné à titre informatif.

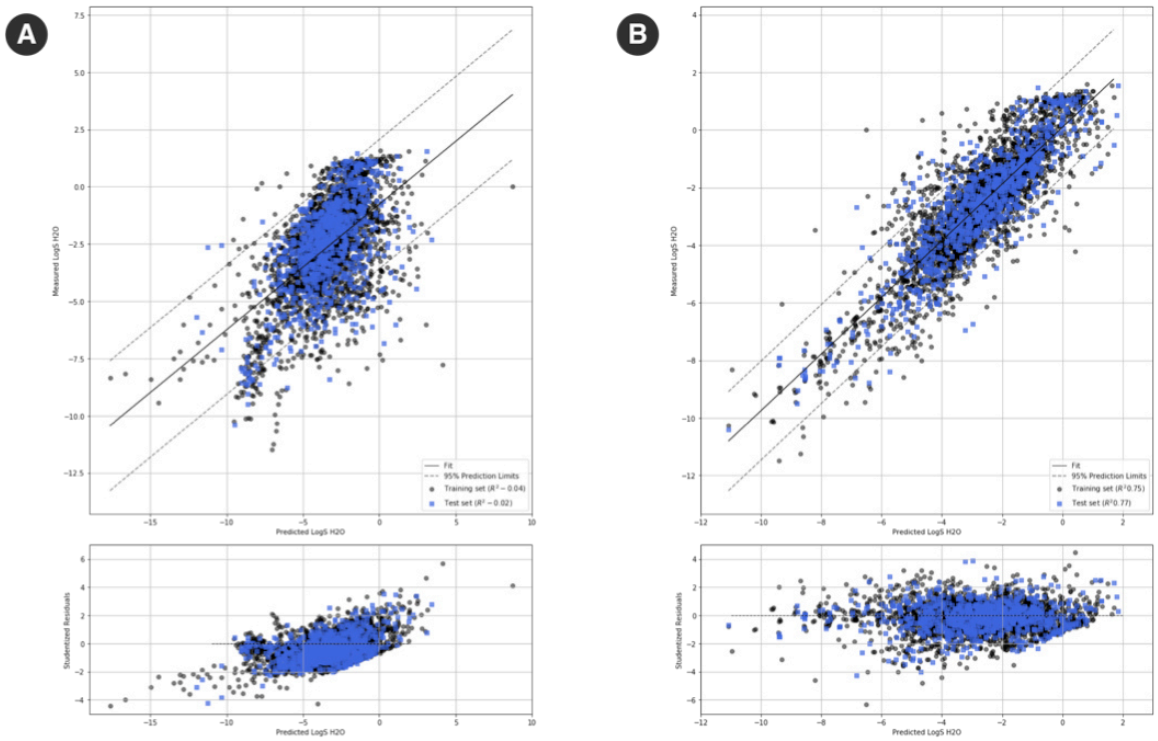


Figure 35 : Ajustement du LogS.

A) Représentation des valeurs expérimentales en fonction des valeurs prédites du LogS avant paramétrage du modèle. B) Représentation des valeurs expérimentales en fonction des valeurs prédites du LogS après paramétrage du modèle.

Dans le cas d'un modèle de classification, seuls les paramètres C ($C_{n+1} = C_n * 10$ avec C compris entre 0,001 et 10) et *kernel* (linéaire, sigmoïdal, polynomial ou fonction de base radiale) peuvent être optimisés. Le modèle qui en résulte possède un C équivalent à 10 et un *kernel* de fonction de base radiale. Les performances du modèle ont été améliorées avec une AUC passant de 0,97 à 0,99 sur le jeu d'apprentissage et une AUC de validation croisée allant de 0,91 à 0,93. Les performances sur le jeu de test restent inchangées avec une AUC égale à 0,97.

Le paramétrage du modèle nous permet d'être dans les meilleures conditions pour élaborer un modèle performant de la propriété visée. Ce modèle est utilisé pour effectuer toutes les étapes de modélisation et de sélection ultérieures.

1.2.3. Détection des points aberrants

Dans le cas des modèles de régression, il est possible que la modélisation d'une propriété soit rendue difficile par la présence de points aberrants (*outliers*). Ainsi, un

modèle peut être prédictif, mais les performances de ce dernier peuvent être impactées par la présence d'individus du jeu d'apprentissage extrêmement mal prédit. Ces points aberrants peuvent survenir à cause de différents phénomènes comme : i) des valeurs de descripteurs extrêmes (points aberrants X), ii) des mesures erronées de la propriété (points aberrants Y), ou iii) les deux phénomènes combinés (points aberrants X et Y). Afin de pouvoir identifier et traiter ces points aberrants, nous avons utilisé l'approche proposée par Xiao [257] qui est inspiré des travaux de Cao *et al* [258]. Nous avons choisi cette approche car elle permet d'identifier les types de points aberrants énoncés précédemment, elle est simple à mettre en œuvre et elle est applicable à tous les modèles de régression.

La méthodologie mise en place repose sur une validation croisée 5-*folds* (jeu d'apprentissage divisé en 5 sous-ensembles stratifiés) répétée 10 fois (Figure 36.A). Durant chaque réplicat, les résidus des composés présents dans les jeux de validation sont déterminés. Suite aux 10 réplications, une matrice (*n* x 10) contenant l'erreur des *n* composés est obtenue. La déviation standard ainsi que l'erreur moyenne absolue (MAE) sont calculées pour chaque composé. La déviation standard traduit l'instabilité du modèle sur les prédictions (problème lié au modèle ou aux descripteurs), et la MAE permet d'identifier les biais systématiques sur la prédiction du LogS (problème lié aux mesures expérimentales de la propriété). Il est alors possible de déterminer quels sont les individus les plus mal prédits, mais également de caractériser l'erreur observée (Figure 36.B). Pour cela, deux seuils sont définis à l'aide des quantiles au risque α de 1 % sur la distribution des MAE et la distribution de la déviation standard (Figure 36.C). La représentation graphique qui en résulte permet ensuite d'identifier et de caractériser les points aberrants.

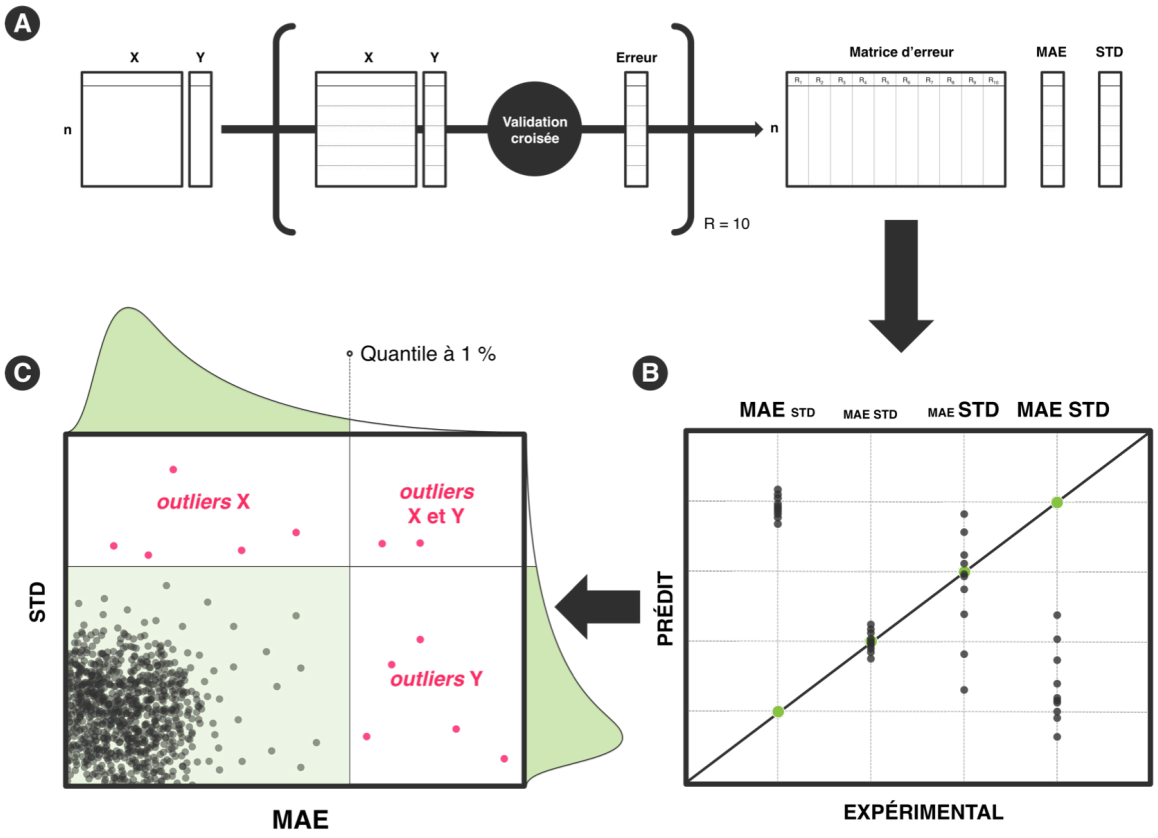


Figure 36 : Représentation schématique de la méthodologie suivie pour la détection des points abérrants.

A) Protocole suivit pour la détermination de l'erreur moyenne absolue (MAE) et de la déviation standard (STD) des résidus. B) Représentation des quatre cas de figures pouvant être rencontrés lors de la prédiction d'une propriété. Les points verts représentent la valeur expérimentale de la propriété pour les composés étudiés ; les points gris représentent les valeurs prédites de la propriété obtenues lors de la réplication de la validation croisée. C) Graphique permettant de différencier les sources d'erreurs présentes dans le jeu de données, ainsi que les points aberrants incriminés (*outliers*). Les trois zones qui spécifient le type d'erreurs sont représentées en rouge.

Parmi les 2755 composés initialement présents dans le jeu d'apprentissage utilisé pour la modélisation du LogS, 52 ont été identifiés comme étant aberrants. Les structures chimiques de ces composés vous sont présentées en Figure 37. La représentation graphique de la déviation standard des résidus en fonction du MAE (Figure 38.A) nous montre que 24 composés sont considérés comme aberrants selon des valeurs inappropriées de descripteurs (X), 24 composés sont aberrants selon des valeurs erronées de LogS (Y) et 4 selon les deux phénomènes combinés (X et Y).

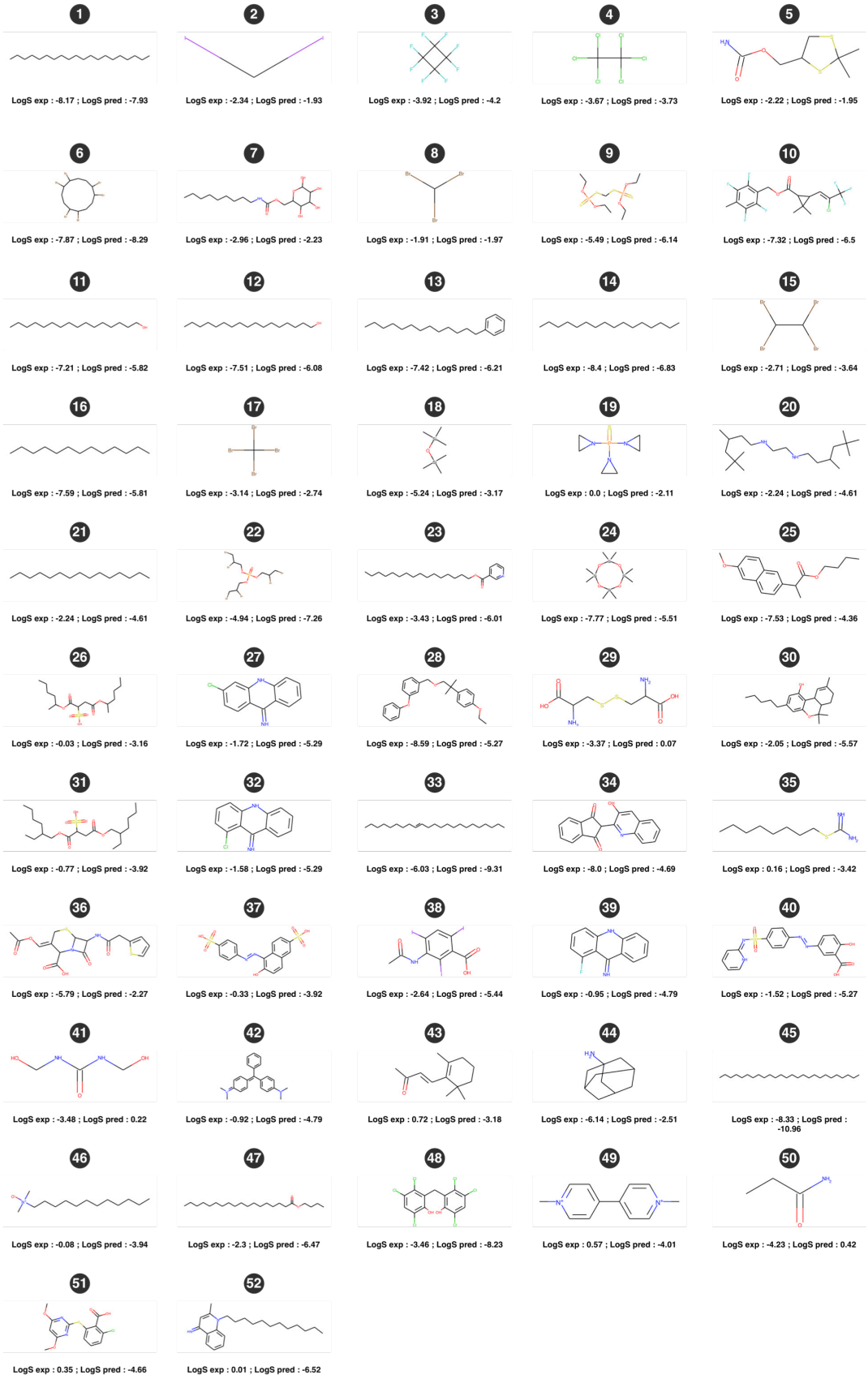


Figure 37 : Structures des 52 individus considérés comme des points aberrants dans le modèle de régression LogS.

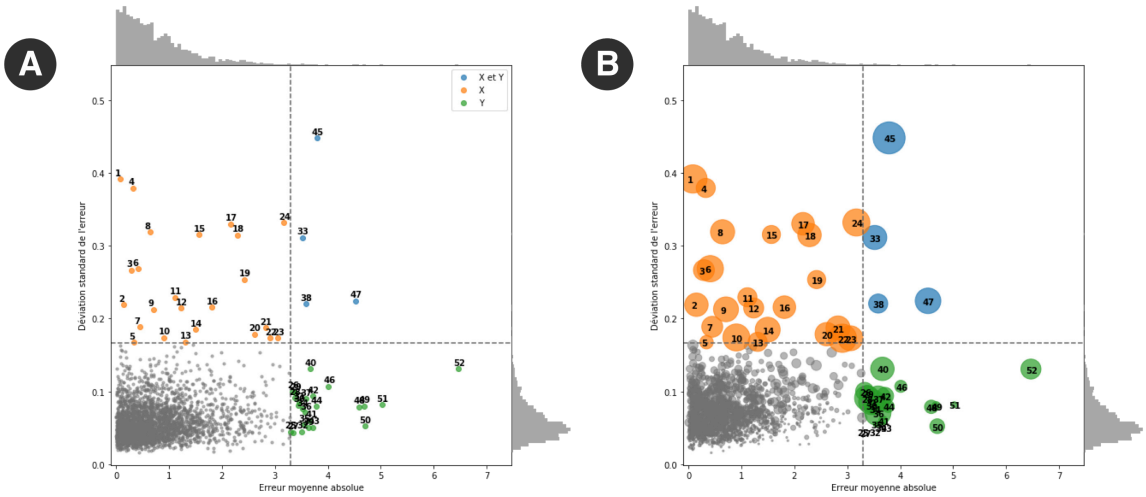


Figure 38 : Identification des individus aberrants dans le modèle de régression.

A) Graphique représentant la déviation standard en fonction de l'erreur moyenne absolue. Ce graphique montre les trois types de points aberrants observés dans le modèle de régression. Les numéros font référence à l'identifiant de chaque individus aberrants. B) Analyse des individus ayant des valeurs de descripteurs extrêmes. La taille d'un individu est proportionnelle au nombre de fois qu'il a été observé en dehors de l'intervalle de confiance à 95% de chaque descripteur.

Les points aberrants selon X sont censés avoir des valeurs de descripteurs extrêmes. Afin de vérifier cette hypothèse, nous avons déterminé quels étaient les individus en dehors de l'intervalle de confiance (95 %) de chaque descripteur (ANNEXE C). L'occurrence des individus a ensuite été utilisée pour proposer l'analyse présentée par la Figure 38.B. Cette occurrence y est représentée par la taille variable des individus. Plus la taille est importante, plus l'individu possède des valeurs extrêmes de descripteurs. On remarque que les individus pour lesquels la déviation standard est importante ont des valeurs de descripteurs extrêmes. Les prédictions de LogS de ces individus sont donc instables. On en déduit que ces derniers peuvent être considérés comme étant les limites de l'espace cartésien exploité par le modèle, pour lesquelles il est plus enclin à se tromper. Les structures chimiques de ces composés ont pour la majorité des chaînes aliphatiques ou des halogènes (composés hydrophobes). Ces observations sont valables pour les individus aberrants selon les critères X et Y.

Les points aberrants Y sont censés présenter un biais systématique induit par une erreur dans les données initiales. Pour vérifier cela, nous avons souhaité comparer les valeurs expérimentales en notre possession et les prédictions du modèle avec de nouvelles mesures expérimentales de la solubilité aqueuse (mesures externes). Parmi les 24 composés incriminés seulement 9 mesures externes ont été trouvées à partir d'autres sources de données (Table 12). On remarque que dans 5 cas sur 9, le modèle donne une

prédiction concordante avec les mesures externes (même ordre de grandeur). A titre d'exemple, pour la sulfasalazine (40), la mesure de la solubilité utilisée par le modèle est de 12 g/L, tandis que la mesure externe nous indique une solubilité inférieure à 1 g/L. Le modèle prédit une solubilité de 2,14 mg/L, ce qui est en accord avec la mesure externe. Un autre exemple concerne le composé 1,3-bis(hydroxyméthyl)urée (41), pour lequel nous avons trouvé deux mesures externes de solubilité, à savoir 40 mg/L et 150 g/L. Aucune explication n'a été trouvée dans la littérature sur la raison de ces deux valeurs très différentes de la solubilité aqueuse. La valeur expérimentale que nous avons à notre disposition correspond à 40 mg/L. La prédiction du modèle représente une solubilité de 198 g/L. Cette prédiction possède un ordre de grandeur en accord avec la deuxième valeur expérimentale de solubilité que nous avons trouvé. Par conséquent, il semblerait que les mesures expérimentales de certains composés soient incorrectes. Ainsi, cette approche nous permet d'identifier les composés pour lesquels la mesure expérimentale est suspecte. Malheureusement, cette vérification n'a pas pu être faite pour tous les individus aberrants (Y) faute de valeurs externes.

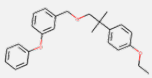
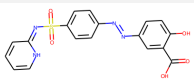
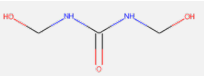
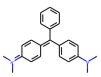
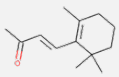
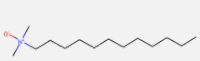
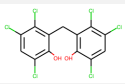
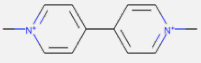
STRUCTURE	ID	LogS EXP	LogS PRED	S EXP	S PRED	S EXTERNE	SOURCE
	28	-8,59	-5,27	95.8 ug/L	2.02 mg/L	100 ug/L	EPA DSSTox
	40	-1,52	-5,27	12 g/L	2.14 mg/L	< 1 g/L	PubChem
	41	-3,48	0,22	40 mg/L	198 g/L	40 mg/L 150 g/L	ParChem
	42	-0,92	-4,79	40 g/L	5.29 mg/L	40 g/L	PubChem
	43	0,72	-3,18	999 g/L	128 mg/L	169 mg/L	EPA DSSTox
	44	-6,14	-2,51	11 mg/L	465 mg/L	6,29 g/L	DrugBank
	46	-0,08	-3,94	190 g/L	26.4 mg/L	190 mg/L	PubChem
	48	-3,46	-8,23	140 mg/L	2.40e-03 mg/L	140 mg/L	DrugBank
	49	0,57	-4,01	700 mg/L	18.3 mg/L	620 mg/L	PubChem

Table 12 : Comparaison des données avec les mesures externes de LogS.

Les valeurs de solubilité (S EXP et S PRED) proviennent de la conversion des valeurs LogS EXP et LogS PRED (MM * 10^{LogS}). La couleur verte est attribuée à la valeur prédite ou expérimentale de la solubilité qui est la plus concordante avec la mesure externe.

Comme mentionné précédemment, des erreurs peuvent également être introduites par des structures chimiques erronées. Nous nous sommes rendu compte que seul laurolinium (44) avait une structure erronée. Dans la structure initiale, l'azote de la quinoline est substitué de telle sorte qu'un ammonium quaternaire est formé. Ce cation n'est plus observé dans la structure standardisée. La neutralisation effectuée lors de l'énumération des tautomères peut expliquer cette anomalie (Figure 39). Ainsi, la perte de cette charge positive réduit considérablement le LogS prédit, ce qui explique la MAE très élevée observée pour ce composé.

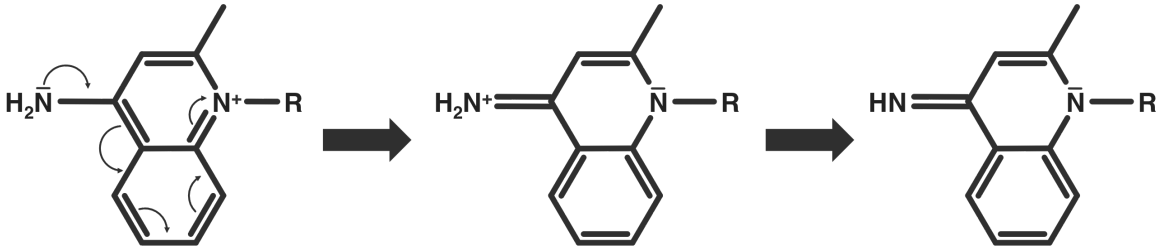


Figure 39 : Perte de la charge pour le laurolinium.

De plus, nous avons remarqué que les structures chimiques des composés 26, 31 et 37 disposaient de groupements sulfates, mais également que les composés 29, 36, 38, 40 et 51 disposaient de groupement carboxylates. Ces groupes fonctionnels sont sous leur forme neutre, alors qu'en milieu aqueux la forme anionique est normalement observée. Ainsi, nous pensons que le biais systématique observé pour ces molécules est induit par la structure chimique qui ne reflète pas l'entité moléculaire testée lors de l'expérimentation. Nous en déduisons que l'ionisation des molécules à pH 7 sera indispensable pour la création des modèles de la solubilité aqueuse. Pour la suite des explications concernant la création de modèles simples, nous resterons dans le cas de molécules neutres.

On en conclut que cette approche nous permet d'identifier les différents phénomènes responsables d'une prédiction inappropriée de certains composés. Nous avons choisi de retirer les individus identifiés comme aberrants, car nous avons considéré ces données comme étant peu fiables et présentant un risque d'incorporer des erreurs dans le modèle. Ainsi, les 52 composés considérés comme aberrants ont été rejetés.

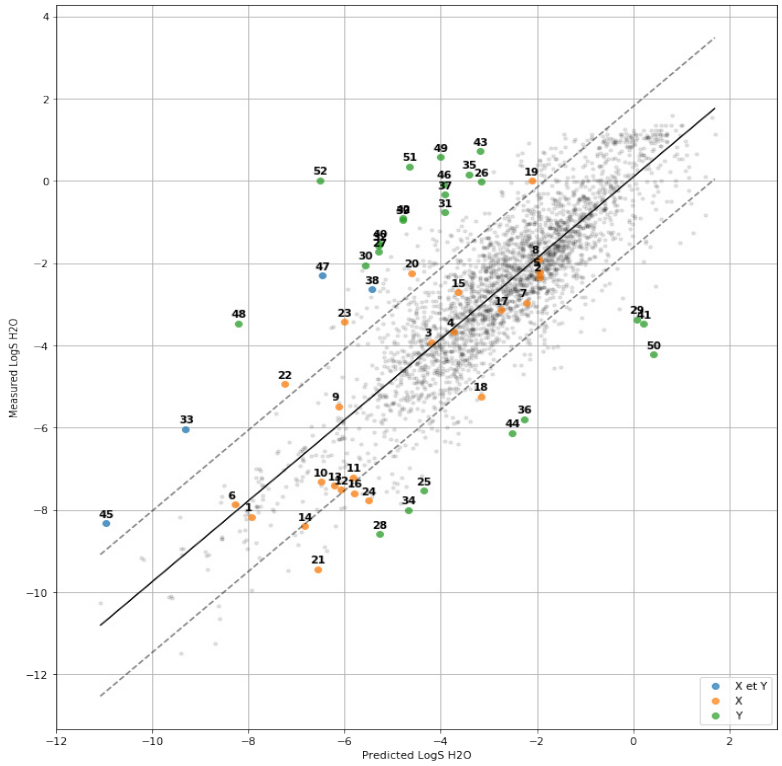


Figure 40 : Représentation des individus aberrants sur l'ajustement du LogS.

Au final, le jeu d'apprentissage contient 2703 composés décrits par 68 descripteurs CDK. La suppression de 2% des individus (Figure 40) permet d'améliorer les performances internes du modèle avec un R^2 et un Q^2 de 0,78. Ceci a également engendré une faible diminution du pouvoir prédictif du modèle avec un R^2_f de 0,76.

1.2.4. Sélection des descripteurs

Pour construire un modèle de prédiction, il est nécessaire de sélectionner des descripteurs appropriés. Cette sélection des descripteurs comporte plusieurs étapes. Elles consistent à supprimer la colinéarité des variables explicatives du jeu de données et à ne sélectionner que les descripteurs les plus pertinents.

1.2.4.1. Suppression des descripteurs corrélés

Certains descripteurs peuvent présenter une colinéarité élevée. Ces descripteurs corrélés apportent une information redondante, et vont statistiquement contribuer de façon équivalente dans un modèle. Ceci peut causer des complications pour i) l'utilisation de certaines méthodes d'apprentissage, ii) l'efficacité du modèle (biais et rapidité), et iii) l'interprétation du modèle. Ainsi, la suppression des descripteurs corrélés consiste à s'assurer que tous les descripteurs sont uniques (non corrélés). Généralement, un seuil de corrélation est appliqué pour supprimer les descripteurs qui possèdent une corrélation

supérieure à ce seuil. Il est admis que deux descripteurs sont corrélés lorsqu'ils possèdent un coefficient de corrélation (r^2) supérieur à 0,8 (ou inférieur à -0,8). La méthode la plus couramment rencontrée dans la littérature consiste ensuite à sélectionner aléatoirement un des deux descripteurs corrélés.

Cependant, nous nous sommes rendu compte que deux descripteurs corrélés pouvaient avoir une influence quelque peu différente vis-à-vis de la propriété modélisée. Comme représenté par la Figure 41.B, les descripteurs CDK_apol et CDK_AMR sont fortement corrélés avec un r^2 égal à 0,93. Néanmoins, la corrélation de chacun d'entre eux avec la propriété LogS est très différente avec un r^2 de 0,62 pour le descripteur CDK_AMR et un r^2 de 0,52 pour le descripteur CDK_apol. On en déduit qu'une sélection aléatoire peut exclure des descripteurs explicatifs vis-à-vis de la propriété. Comme mentionné dans le Ch1 3.2.2, un descripteur doit être corrélé avec l'activité étudiée et montré une corrélation négligeable avec les autres descripteurs.

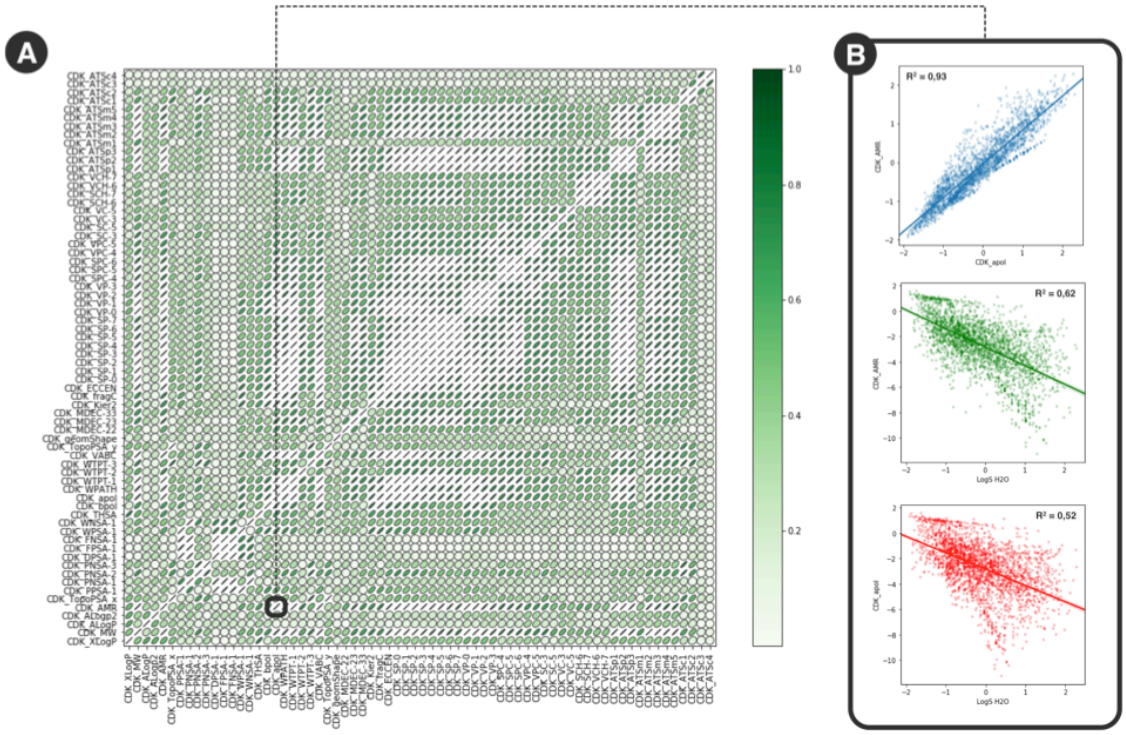


Figure 41 : Matrice de corrélation entre les descripteurs CDK utilisés pour la modélisation du LogS.

A) Matrice de corrélation. B) Corrélation entre les descripteurs CDK_AMR, CDK_apol et la propriété LogS.

Pour cette raison, nous avons souhaité intégrer cette particularité à notre protocole de suppression des descripteurs corrélés. Pour cela, la plateforme MetaPredict identifie dans un premier temps les relations qui existent entre la propriété modélisée et chaque

descripteur. Nous avons défini cette relation de deux manières différentes. Pour un modèle de régression, nous avons considéré le coefficient de corrélation absolue entre chaque descripteur et la propriété modélisée. Pour un modèle de classification, nous avons déterminé pour chaque descripteur une valeur-p (*p-value*) à l'aide d'un test statistique, qui a pour but de vérifier l'égalité des moyennes des deux classes (échantillons) au sein d'un descripteur (population).

Ce test (Figure 42.A) est basé sur le test *t* de Student (paramétrique) et le test de Mann-Whitney-Wilcoxon (non paramétrique). Le test *t* de Student suppose une distribution normale et une variance égale des deux échantillons analysés. Pour vérifier ces critères, deux tests statistiques complémentaires ont été intégrés à savoir : i) le test de Shapiro-Wilk pour vérifier que la distribution des échantillons suit une loi normale, et ii) le test de Fisher-Snedecor pour vérifier l'égalité des variances entre les échantillons. Si les tests de Shapiro-Wilk et de Fisher-Snedecor sont vérifiés, le test *t* de Student est effectué pour comparer l'égalité des moyennes entre les deux échantillons. Si l'une des deux conditions du test *t* de Student n'est pas vérifiée, le test de Mann-Whitney-Wilcoxon est appliqué, afin de comparer l'égalité des médianes observées entre les deux échantillons. Pour plus de détails, une description de chaque test statistique est présentée en ANNEXE D. La valeur-p qui en résulte doit vérifier l'hypothèse H_1 selon laquelle les deux échantillons (classes active et inactive) de la population visée (descripteur) sont différents. Cette hypothèse H_1 est vérifiée lorsque la valeur-p est inférieure au risque α de 5 %. Pour la suite, nous avons défini cette valeur-p comme étant le pouvoir discriminant d'un descripteur, qui traduit sa capacité à séparer les deux classes étudiées. Le pouvoir discriminant de chaque descripteur utilisé pour la modélisation du LogS est illustré par la Figure 42.B. On remarque que la valeur-p permet d'ordonner facilement les descripteurs en fonction de leur capacité à séparer les deux classes. Ainsi, lors de l'étude de la colinéarité des descripteurs, il est à présent possible de sélectionner les plus appropriés, c'est-à-dire les descripteurs ayant une relation étroite avec la propriété modélisée.

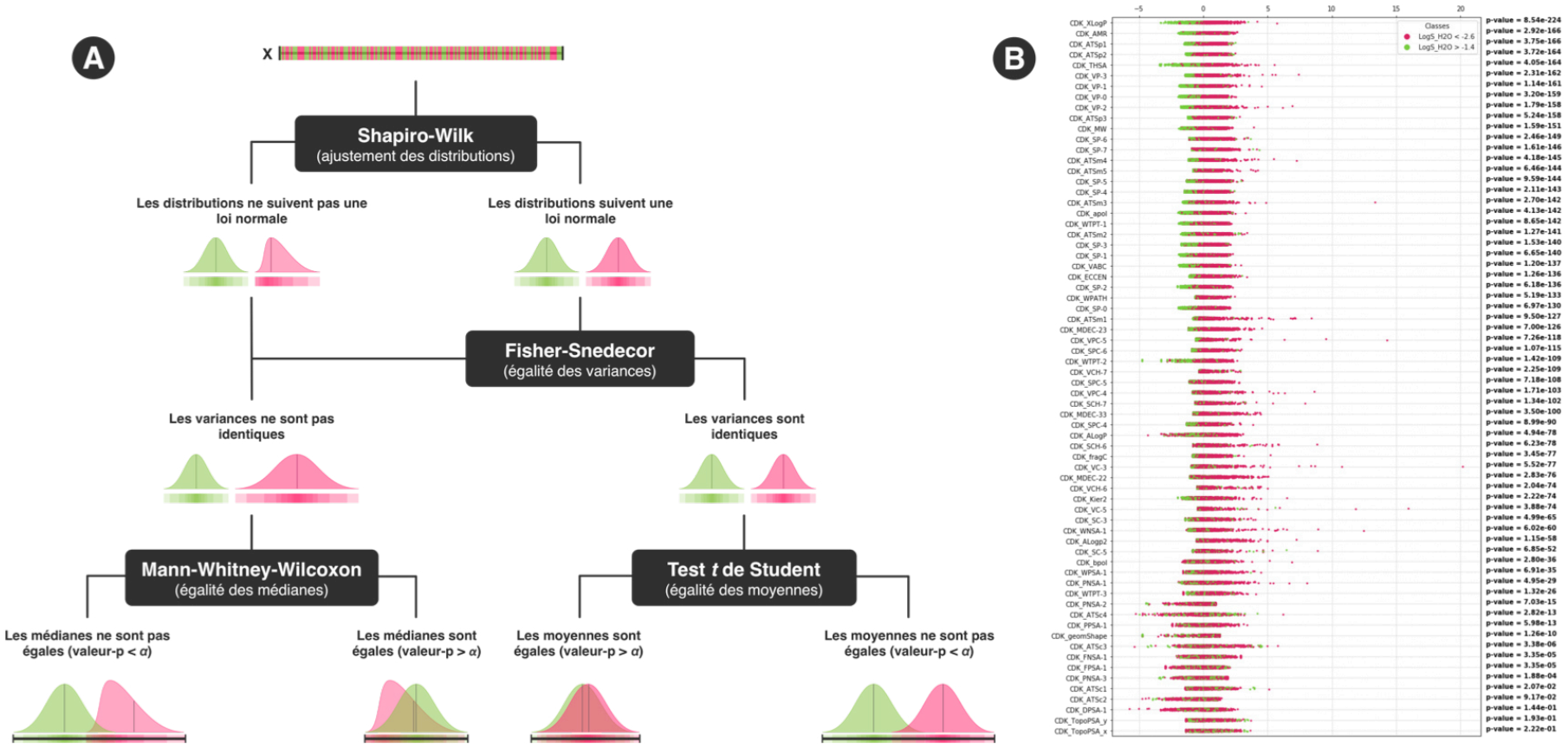


Figure 42 : Pouvoir discriminant des descripteurs.

A) Test statistique permettant de déterminer le pouvoir discriminant d'un descripteur. Le descripteur X est représenté avec la répartition des classes active (vert) et inactive (rouge). Les étapes successives effectuées lors du test statistique sont présentées avec un schéma des conclusions observées. B) Répartition des classes sur les valeurs numériques des descripteurs CDK utilisés pour modéliser le LogS. Les descripteurs ont été ordonnés en fonction de leur pouvoir discriminant, dont la valeur-p est annotée sur la droite de ce graphique.

Dans un deuxième temps, la plateforme génère une matrice de corrélation permettant d'identifier les descripteurs corrélés. Les coefficients absolus sont utilisés afin de traiter les descripteurs corrélés et anti-corrélés. Lorsque deux descripteurs possèdent un coefficient de corrélation absolu supérieur ou égal au seuil choisi, la plateforme sélectionne celui qui dispose du r^2 maximal avec la variable expliquée ou du pouvoir discriminant minimal (valeur-p). Ceci nous assure que les descripteurs sélectionnés permettent d'expliquer la propriété tout en réduisant la colinéarité dans le jeu de données.

Nous avons également remarqué qu'aucune règle empirique n'a été proposée dans la littérature au sujet du choix du seuil de corrélation. Généralement, ce seuil est compris entre 0,8 et 0,95 et il est choisi de façon arbitraire. Cependant, ce paramètre est d'une importance cruciale car il va permettre de retirer un nombre plus ou moins important de descripteurs, et ceci au détriment des performances internes du modèle ou de son pouvoir prédictif. Selon nous, l'optimum correspond au seuil qui permet de retirer un nombre important de descripteurs corrélés, tout en préservant ou en améliorant les performances du modèle. Par conséquent, la plateforme va étudier la réduction de la colinéarité à différents seuils de corrélation, compris entre 0,5 et 0,95 par incrément de 0,05. Nous avons choisi cette gamme de seuils, car l'optimum n'est pas obligatoirement compris entre 0,8 et 0,95. Par conséquent, il est possible que des seuils inférieurs permettent d'obtenir un modèle de performances élevées tout en utilisant un nombre plus restreint de descripteurs. Pour chaque seuil, les descripteurs corrélés vont être identifiés puis supprimés selon la méthode décrite précédemment. Les descripteurs restants sont utilisés pour générer un modèle. Cette opération est réitérée pour tous les seuils testés.

Afin de montrer les avantages de notre méthode par rapport à une sélection aléatoire, nous proposons les résultats obtenus pour la modélisation du LogS (Figure 43). On remarque que notre méthodologie permet de préserver les performances du modèle quelque soit le seuil de corrélation sélectionné. Cette observation ne peut pas être faite dans le cadre d'une sélection aléatoire, pour laquelle les performances chutent lorsque le seuil de corrélation est inférieur à 0,8, et ceci quelque soit le type de modèle. L'utilisation d'une sélection aléatoire fait l'hypothèse que deux descripteurs colinéaires ont une relation équivalente avec la propriété modélisée. Ainsi, plus la corrélation entre les descripteurs diminue, moins cette hypothèse peut être vérifiée. Cela explique pourquoi les seuils de corrélation généralement choisis dans la littérature sont compris entre 0,8 et 0,95. Pour de faibles seuils de corrélation, la sélection aléatoire va donc potentiellement supprimer les descripteurs les plus explicatifs de la propriété modélisée.

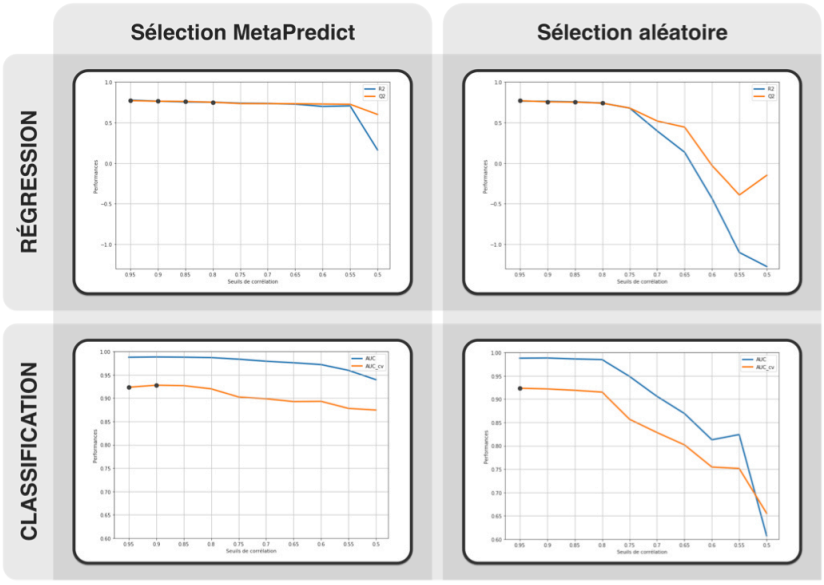


Figure 43 : Etude de l'influence du seuil de corrélation sur les performances d'un modèle de prédiction.

Les résultats sont présentés pour un modèle de régression et un modèle de classification du LogS. Pour chaque type de modèle, nous présentons une sélection des descripteurs non corrélés identifiés selon la méthodologie interne à MetaPredict et une sélection aléatoire. Les courbes bleues représentent les performances du modèle sur le jeu d'apprentissage en fonction des différents seuils de corrélation testés. Les courbes oranges représentent les performances du modèle lors d'une validation croisée *5-folds* en fonction des différents seuils de corrélation testés. Les points noirs représentent les seuils de corrélation identifiés comme étant optimums.

Notre approche nous assure d'être dans les conditions optimales quelque soit le seuil choisi. Les modèles obtenus sont donc plus stables et sont basés sur les descripteurs les plus explicatifs de la propriété à modéliser. Le seuil de corrélation optimal est sélectionné par la plateforme en étudiant conjointement la fluctuation des performances du modèle sur le jeu d'apprentissage et lors de la validation croisée. Seul le seuil qui permet d'apporter les performances internes optimales est choisi pour la suppression des descripteurs colinéaires. Ainsi, pour le modèle de régression du LogS le seuil à 0,80 a été considéré comme optimal (Figure 44.A). Le jeu de données contient au final 24 descripteurs CDK pour lequel, le modèle possède un R^2 (jeu d'apprentissage) et un Q^2 (validation croisée) de 0,78 et pour finir un R^2_f (jeu de test) de 0,74. Pour le modèle de classification le seuil de corrélation de 0,90 a été choisi (Figure 44.B). Le jeu de données qui en résulte contient 35 descripteurs CDK, pour lequel le modèle possède une AUC (jeu d'apprentissage) de 0,99, une AUC_{cv} (validation croisée) de 0,92 et une AUC_f (jeu de test) de 0,97.

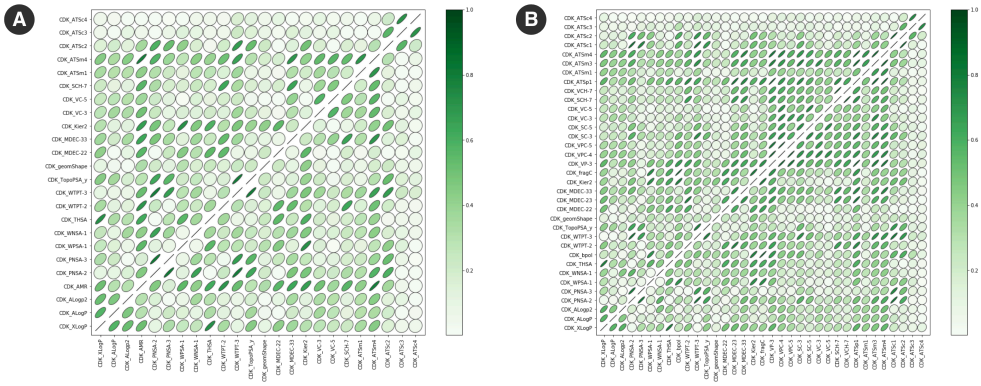


Figure 44 : Matrices de corrélation après suppression des descripteurs corrélés.

A) Matrice de corrélation obtenu suite à la suppression des descripteurs corrélés du jeu d'apprentissage utilisé par le modèle de régression. B) Matrice de corrélation obtenu suite à la suppression des descripteurs corrélés du jeu d'apprentissage utilisé par le modèle de classification.

1.2.4.2. Descripteurs pertinents

Une étude plus approfondie des descripteurs est nécessaire pour justifier leur pertinence. La pertinence peut être définie comme l'influence du descripteur sur les performances d'un modèle. Pour analyser leur pertinence, les descripteurs sont dans un premier temps ordonnés en fonction de leur importance dans le modèle. Une approche graduelle descendante (*stepwise*) est ensuite effectuée en excluant les descripteurs de faible importance. Lors de chaque itération, un modèle est reconstruit sur la base des descripteurs restants et l'ensemble du processus est réitéré. Cette approche permet de ne sélectionner que les descripteurs indispensables pour la modélisation de la propriété, et les descripteurs nécessaires pour assurer les performances du modèle. Les résultats obtenus pour les modèles de LogS sont présentés dans la Figure 45.

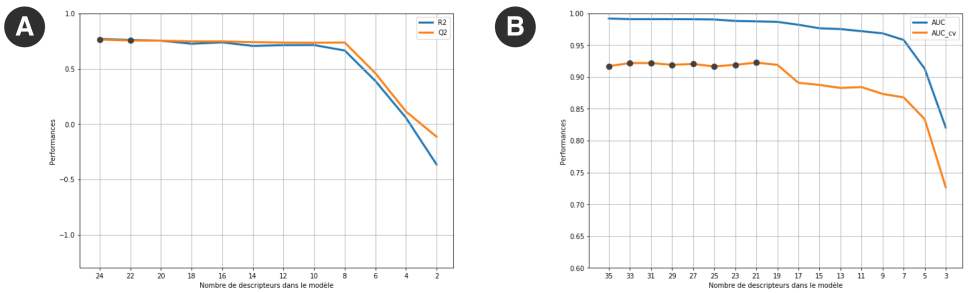


Figure 45 : Sélection des descripteurs pour les modèles de prédiction du LogS.

Les courbes bleue et orange représentent respectivement les performances du modèle sur le jeu d'apprentissage et lors de la validation croisée *5-folds*. Les points noirs représentent les nombres de descripteurs apportant les meilleures performances. A) Sélection des descripteurs pertinents pour le modèle de régression. B) Sélection des descripteurs pertinents pour le modèle de classification.

1.2.5. Validation du modèle

Une fois les descripteurs optimums sélectionnés, le modèle est validé en suivant la procédure standard. La validation interne est effectuée afin de déterminer i) dans un premier temps la qualité du modèle sur le jeu d'apprentissage, puis ii) dans un second temps la robustesse du modèle à l'aide d'une validation croisée 5-*folds*. Afin de vérifier que les performances du modèle ne soient pas dûes au hasard, une étape de randomisation de la propriété est répétée 100 fois en mélangeant aléatoirement les mesures expérimentales (Y). La validation externe sur le jeu de test est ensuite effectuée afin de déterminer le pouvoir prédictif du modèle sur l'ensemble de données qui lui est inconnu.

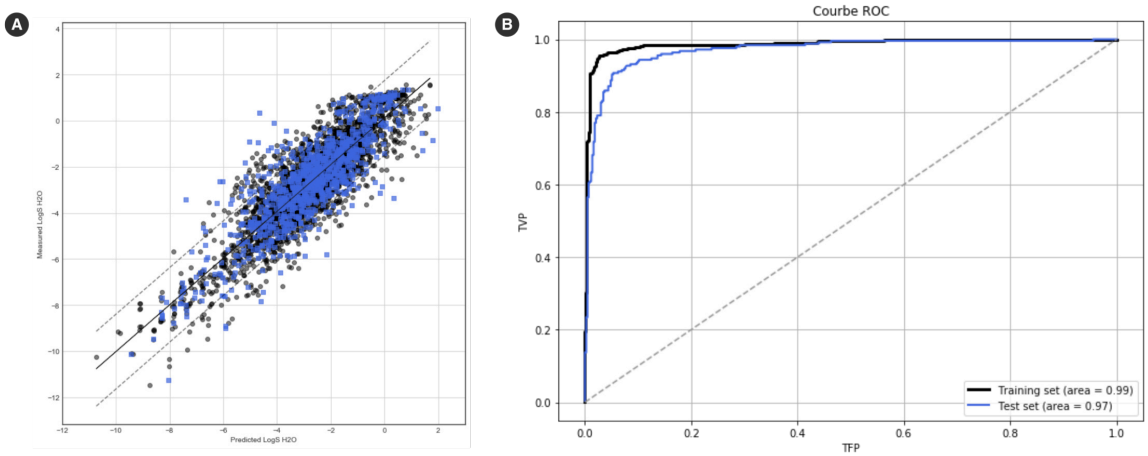


Figure 46 : Représentation des performances des modèles LogS.

A) Ajustement du modèle de régression. B) Courbes ROC du modèle de classification.

Un modèle de régression est considéré comme valide par la plateforme, lorsque le R^2 et le R^2_{adj} déterminés à partir du jeu d'apprentissage sont supérieurs à 0,60 ; lorsque le Q^2 obtenu lors de la validation croisée est supérieur à 0,60 ; lorsque le R^2_S déterminé lors de la randomisation de la propriété est inférieur à 0,10 ; et lorsque le R^2_f calculé durant la validation externe est supérieur à 0,5. Pour la modélisation du LogS (Figure 46.A), le modèle de régression possède un R^2 et un R^2_{adj} de 0,78. Les performances sur le jeu d'apprentissage nous montrent que les 22 descripteurs sélectionnés par la plateforme permettent d'obtenir des performances comparables à celles du modèle construit sur la totalité des descripteurs CDK. D'autre part, le R^2_{adj} nous montre que le nombre de descripteurs sélectionnés n'est pas excessif et que le modèle est donc valide sur le jeu d'apprentissage. Le Q^2 est de 0,78, ce qui nous confirme qu'aucun sur-apprentissage du modèle n'est observé. Le R^2_S de 0,02 ± 0,01 nous permet de confirmer que ce modèle

n'est pas enclin au phénomène de corrélation aléatoire. Pour finir, le pouvoir prédictif R_f^2, est égal à 0,74, ce qui nous prouve que le modèle devrait être capable de prédire correctement de nouvelles séries moléculaires.

Un modèle de classification est considéré comme valide, lorsque le MCC déterminé à partir du jeu d'apprentissage est supérieur à 0,60 ; lorsque le MCC_{cv} obtenu lors de la validation croisée est supérieur à 0,60 ; lorsque le MCC_s déterminé lors de la randomisation de la propriété est inférieur à 0,10 et lorsque le MCC_f calculé durant la validation externe est supérieur à 0,5. Les performances du modèle de classification pour la modélisation du LogS sont présentées par la Table 13.

VALIDATION	AUC	MCC	JUSTESSE	SENSIBILITÉ	SPÉCIFICTÉ	PRÉCISION
Apprentissage	0,99	0,93	0,96	0,96	0,97	0,96
Validation croisée	0,97 ± 0,01	0,82 ± 0,02	0,91 ± 0,02	0,90 ± 0,02	0,92 ± 0,02	0,91 ± 0,01
Randomisation	0,50 ± 0,01	0,00 ± 0,02	0,50 ± 0,01	0,50 ± 0,01	0,49 ± 0,01	0,50 ± 0,01
Test	0,97	0,85	0,93	0,95	0,92	0,92

Table 13 : Performances du modèle de classification pour la prédiction du LogS.

On remarque que le modèle peut être considéré comme valide selon les règles énoncées précédemment. On observe également que la sensibilité est supérieure à la spécificité lorsque le jeu de test dispose de classes déséquilibrées. Le modèle permet donc de limiter la prédiction de faux positifs (Figure 47). On en déduit que l'utilisation d'un jeu d'apprentissage avec des classes équilibrées, ainsi que la création d'une zone de délétion entre les populations active et inactive permet d'améliorer le pouvoir prédictif du modèle.

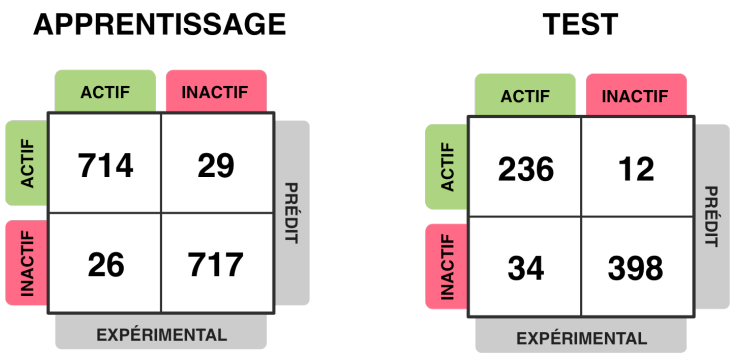


Figure 47 : Tables de contingences des validations internes et externes.

La construction des modèles se finalise par la détermination du domaine d'applicabilité.

1.2.6. Domaine d'applicabilité

La méthodologie que nous avons utilisée est basée sur le principe des *k*-plus proches voisins (*k*-NN) pour définir le DA. Nous avons choisi cette méthode pour les avantages qu'elle présente au sujet de la densité locale, de sa simplicité et de son application possible pour les modèles de régression et de classification. Le *k*-NN permet d'étudier la similarité (distance) entre le jeu d'apprentissage et un jeu de données inconnu. Ainsi, la distance d'un jeu de données inconnu est définie à partir de ses *k* composés les plus proches dans l'espace chimique couvert par le modèle. De faibles valeurs de distance indiquent une grande similitude, tandis que des valeurs de distances élevées signifient une inadéquation importante. Par conséquent, la valeur *k* joue un rôle prépondérant dans la définition des limites du DA. Théoriquement, plus le nombre k augmente, plus le pouvoir prédictif du modèle diminue. L'approche que nous avons mis en place comporte trois étapes principales : i) l'ajustement de l'espace cartésien, ii) la définition de seuils propres à chaque composé du jeu d'apprentissage, iii) l'évaluation du DA pour de nouveaux jeux de données. Pour une compréhension plus claire des étapes décrites ci-après, une représentation de cette méthodologie est présentée (Figure 48).

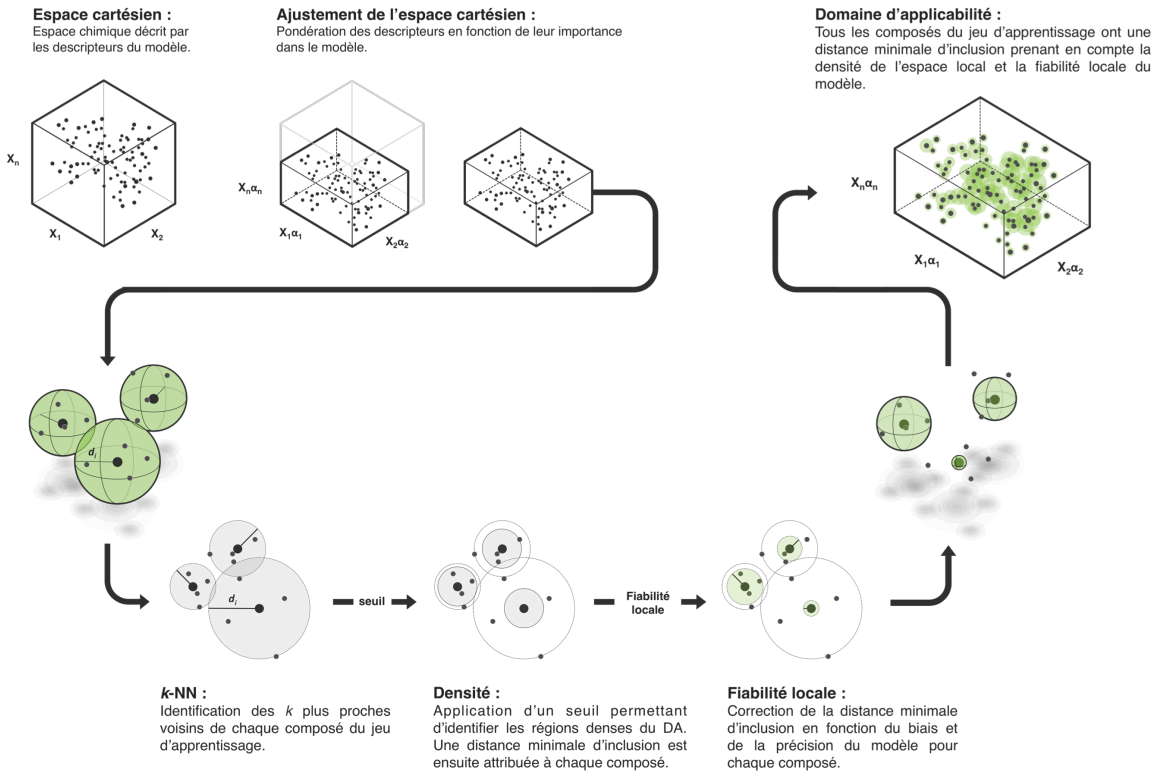


Figure 48 : Représentation schématique de la méthodologie suivie pour définir le domaine d'applicabilité.

1.2.6.1. Ajustement de l'espace cartésien

L'espace chimique exploré par le *k*-NN est défini en fonction des descripteurs utilisés par le modèle de prédiction. Cependant, ces descripteurs n'apportent pas une information chimique ou moléculaire équivalente dans le modèle. En effet, l'équation utilisée par le modèle pour prédire la propriété de nouveaux composés comporte des coefficients différents, pouvant être assimilés à l'importance d'un descripteur. Par conséquent, certains descripteurs sont plus importants que d'autres. L'objectif de l'ajustement de l'espace cartésien consiste à définir un DA sur l'information chimique et moléculaire utile pour le modèle. Les coordonnées de chaque individu sont pondérées en fonction d'un facteur de contribution. Le facteur de contribution est déterminé en normalisant (entre 0 et 1) les coefficients absolus des descripteurs issus de l'équation du modèle. De plus, seuls les descripteurs qui disposent d'un facteur de contribution supérieur à 10 % sont considérés comme pertinents pour définir le DA, et ceci dans le but de pallier à l'impact de la dimensionnalité.

1.2.6.2. Détermination des limites du DA

Les limites du DA ont une grande influence sur la règle d'acceptabilité selon laquelle, un composé inconnu sera considéré dans le domaine d'interpolation ou dans le domaine d'extrapolation. Afin de définir ces limites, nous avons appliqué la méthode proposée par Sahigara *et al.* [207], aussi nommée *density k*-NN (*dk*-NN), car elle permet de définir des seuils uniques pour tous les composés du jeu d'apprentissage en fonction de leur densité locale. Cette méthode débute par la détermination des *k* plus proches voisins de chaque composé i du jeu d'apprentissage à l'aide d'un calcul de distance. Les distances moyennes $\bar{d}_i(k)$ de chaque composé i avec leurs *k* voisins sont ensuite calculées. L'ensemble de ces distances moyennes $\bar{d}\,(k)$ permet de définir un seuil de référence $\tilde{d}(k)$ calculé selon l'Equation 19.

$$\tilde{d}(k) = Q_1\left(\bar{d}\,(k)\right) + 1{,}5\left[Q_3\left(\bar{d}\,(k)\right) - Q_1\left(\bar{d}\,(k)\right)\right]$$

Equation 19 : Calcul du seuil de référence $\tilde{d}(k)$.

Avec les valeurs $Q_1\left(\bar{d}\,(k)\right)$ et $Q_3\left(\bar{d}\,(k)\right)$ correspondent au 25ième et au 75ième centiles dans le vecteur $\bar{d}\,(k)$.

Ce seuil est alors appliqué à chaque composé i pour sélectionner ses *k* voisins représentatifs, c'est-à-dire ceux qui ont une distance inférieure ou égale au seuil de référence, sinon ils sont rejetés. Le nombre *k* de voisins pour lesquels cette condition est

vérifiée va permettre de décrire la densité locale observée au voisinage de chaque composé i du jeu d'apprentissage. Les limites du DA sont déterminées en calculant la distance moyenne des composés i avec leurs *k* voisins représentatifs. Ainsi, à chaque composé du jeu d'apprentissage est associé un seuil individuel (t_i). Le choix du paramètre *k* a une importance pour la détermination de ces seuils individuels. Nous avons choisi d'utiliser une valeur de paramètre *k* égale à $n^{1/3}$ selon les recommandations de Sahigara *et al.* avec *n* égal au nombre d'individus dans le jeu de test.

Nous avons également intégrer la notion de fiabilité locale dans la détermination des seuils individuels t_i. Pour cela, nous avons considéré le biais et la précision du modèle pour chaque composé du jeu d'apprentissage, afin de définir un facteur de correction (f_c). Le biais et la précision du modèle sont déterminés à l'aide d'une validation croisée 5-*folds* répétée 10 fois. Pour un modèle de régression, ce facteur de correction est calculé à partir de l'erreur moyenne absolue et de la déviation standard de l'erreur pour chaque composé. Le facteur de correction est ensuite calculé selon l'Equation 20.

$$f_c = \left(1 - \sqrt{\frac{\sum_1^M \left(\left|\hat{y}_{i,M} - \bar{y}_i\right|\right)^2}{M-1}}\right) * \left(1 - \frac{e_i - e_{min}}{e_{max} - e_{min}}\right) \quad avec \quad e = \sum_{i=1}^{i=n} \left(\frac{\sum_1^M \left|\hat{y}_{i,M} - y_i\right|}{M}\right)$$

Equation 20 : Facteur de correction appliqué lors d'un modèle de régression.

Avec M le nombre de modèles définis lors de la validation croisée; $\hat{y}_{i,M}$ la valeur prédite de la propriété du composé i selon le modèle M ; $\bar{y}_i$ la moyenne des valeurs prédites sur les M modèles ; y_i la valeur expérimentale de la propriété sur composé i ; e l'erreur sur la prédiction pour l'ensemble des individus.

Pour un modèle de classification, le facteur de correction va être déterminé en fonction de la concordance des modèles et de la déviation standard sur la probabilité d'activité (Pa). Le facteur de correction est ensuite calculé selon l'Equation 21.

$$f_c = \left(1 - \sqrt{\frac{\sum_1^M \left(\left|\widehat{Pa}_{i,M} - \overline{Pa_i}\right|\right)^2}{M-1}}\right) * \left(\frac{|y_i \cap \hat{y}_i|}{M}\right)$$

Equation 21 : Facteur de correction appliqué lors d'un modèle de classification.

Avec M le nombre de modèles définis lors de la validation croisée; $\widehat{Pa}_{i,M}$ la probabilité d'activité du composé i selon le modèle M ; $\overline{Pa_i}$ la moyenne des probabilités sur les M modèles ; $\hat{y}_i$ la classe prédite du composé i ; y_i la classe expérimentale du composé i.

Ce facteur de correction est ensuite multiplié aux seuils individuels t_i de chaque composé du jeu d'apprentissage.

1.2.6.3. Evaluation de nouveaux jeux de données

Pour déterminer si un composé testé est assimilable au domaine d'interpolation, sa distance avec tous les individus du jeu d'apprentissage est calculée et comparée aux seuils individuels de ces derniers. Si la distance est inférieure ou égale au seuil individuel d'un composé du jeu d'apprentissage, le composé testé sera considéré dans le domaine d'interpolation.

1.2.6.4. Application au modèle de régression du LogS

Dans le but de montrer les avantages de la méthodologie que nous avons mis en place, nous proposons les résultats de son application sur le modèle linéaire de LogS. Une brève comparaison sera faite avec les résultats obtenus à l'aide du *dk*-NN proposé par Sahigara *et al*. Notre objectif est de démontrer que la méthodologie utilisée est capable de définir un DA représentatif du modèle de régression étudié. Un DA est considéré comme représentatif lorsqu'une dégradation décroissante du pouvoir prédictif est observée à mesure que la distance (*k*) au cœur du modèle augmente. Afin de vérifier cette hypothèse, nous avons étudié la variation du pouvoir prédictif du modèle de régression LogS en fonction du nombre de plus proches voisins *k* utilisé pour définir le DA (Figure 49.A). Les courbes orange et bleue représentent respectivement la variation du R_f^2 lors de la création du DA avec MetaPredict et le *dk*-NN proposé par Sahigara *et al*. Le calcul du R_f^2 est réalisé uniquement à l'aide des composés du jeu de test considérés dans le domaine d'interpolation.

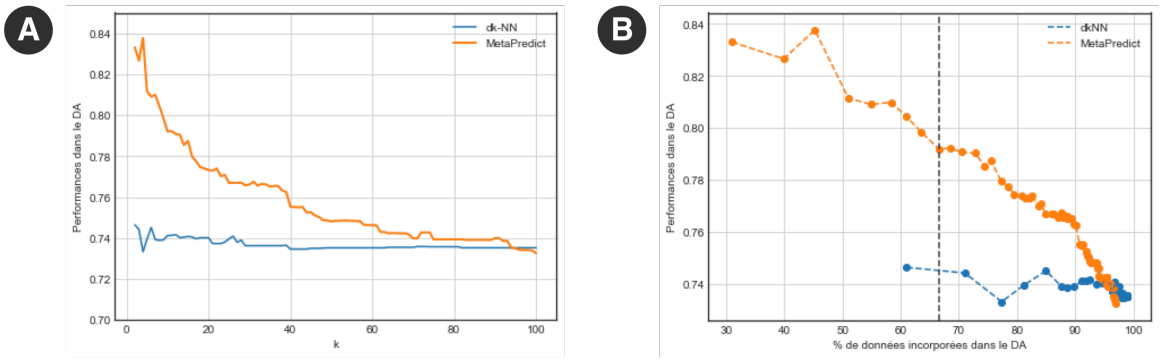


Figure 49 : Comparaison des méthodologies MetaPredict et *dk*-NN.

A) Pouvoir prédictif du modèle en fonction du nombre *k* de plus proches voisins utilisé pour définir les limites du DA. B) Pouvoir prédictif du modèle en fonction du pourcentage du jeu de test incorporé dans le domaine d'interpolation. La ligne verticale grise représente la valeur du paramètre *k* sélectionnée par la plateforme pour définir le DA final.

On remarque que le modèle linéaire possède un pouvoir prédictif supérieur avec MetaPredict qu'avec le *dk*-NN. Cette observation peut être justifiée par le fait que les

seuils t_i corrigés en fonction de la fiabilité locale sont plus restrictifs. Comme représenté par la Figure 50, ces graphiques nous montrent que plus la distance moyenne d'une molécule avec ces plus proches voisins augmente, plus le DA est susceptible d'accepter des molécules ayant une erreur de prédiction élevée. Nous pouvons également voir que les seuils t_i corrigés (Figure 50.B) pénalisent fortement les composés du jeu d'apprentissage pour lesquels l'erreur de prédiction est élevée. Par conséquent, le DA MetaPredict incorpore des groupes de molécules plus petits dans le domaine d'interpolation, et en priorité dans des régions de l'espace chimique correctement prédit par le modèle. Ceci peut être vérifié par la Figure 49.B qui représente la performance externe du modèle en fonction du pourcentage de données incorporées dans le domaine d'interpolation. Ce pourcentage a été calculé à partir du nombre de composés du jeu de test considérés dans le domaine d'interpolation. Il est dépendant du nombre *k* pour lequel il a été défini. Par conséquent, le premier point de chaque profil correspond au même paramètre *k* (*k*=1). On observe que le *dk*-NN incorpore environ 60 % du jeu de test lorsque le nombre de voisins est fixé à 1 (R_f^2 = 0,746), tandis que pour cette même valeur de paramètre MetaPredict incorpore que 30 % des composés testés (R_f^2 = 0,833).

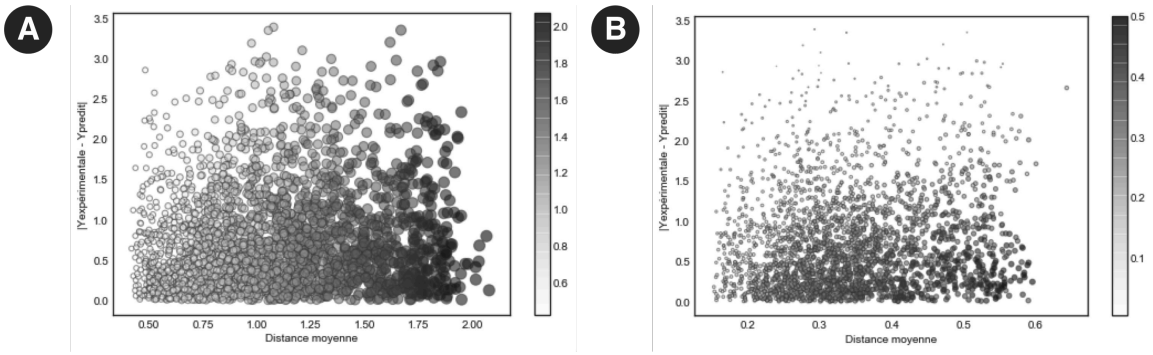


Figure 50 : Représentation des distances moyennes de chaque individu avec leur *k* plus proches voisins en fonction de l'erreur absolue sur la prédiction.

La taille et la coloration des points sont fonction du seuil (t_i) de chaque composé du jeu d'apprentissage. A) Représentation des seuils (t_i) définis par le *dk*-NN. B) Représentation des seuils (t_i) définis par MetaPredict.

Dans le cas de la plateforme MetaPredict, on remarque également une dégradation décroissante des performances. Ainsi, nous pouvons considérer que le DA définit est représentatif du modèle linéaire du LogS. En somme, la plateforme a sélectionné un paramètre *k* égal à 9 pour définir le DA du modèle (Figure 49.B). Le DA pour cette valeur de plus proches voisins incorpore 67 % du jeu de test, tout en assurant un R_f^2 de 0,79.

1.3. Modèles de consensus

Nous venons de voir l'ensemble des étapes assurées par la plateforme afin de produire des modèles (Q)SAR. Nous avons également vu qu'il était possible d'échantillonner plusieurs conditions comme les descripteurs, les découpes successives du jeu de données ou encore l'utilisation de plusieurs algorithmes. Ainsi, l'ensemble des modèles obtenus peuvent être exploités individuellement ou peuvent être combinés afin de générer des modèles de consensus. Comme nous l'avons mentionné précédemment, un modèle de consensus est une combinaison de plusieurs modèles individuels qui disposent tous d'une erreur intrinsèque. L'objectif du consensus est de comparer les prédictions issues des modèles individuels afin d'obtenir des prédictions plus fiables et plus utiles. La création d'un modèle de consensus nécessite une sélection des modèles individuels. Certains outils, comme par exemple AutoQSAR, construisent des consensus en sélectionnant les modèles individuels en fonction de leurs performances. Dans ce cas, les modèles sont ordonnés en fonction de leur score (M_{score}) et seul les *n* premiers sont sélectionnés pour former le consensus (*n* = 10 par défaut).

Comme mentionné précédemment, la création d'un modèle (Q)SAR prête une attention particulière à ce que toutes les données (molécules et descripteurs) apportent une information unique. Le but est de s'assurer qu'aucun biais statistique ne soit introduit dans le modèle. Ces spécificités doivent être également appliquées lors de la création d'un consensus. Ainsi, parmi les *n* modèles de meilleurs scores sélectionnés pour créer un modèle de consensus, il est fort probable que certains d'entre eux soient redondants. Par conséquent, la prédiction obtenue à partir de ce consensus ne reflète que le vote majoritaire du modèle le plus représenté. Dans le cadre de la plateforme MetaPredict, nous avons souhaités caractériser la redondance des modèles individuels afin de ne sélectionner que des modèles uniques pour la création de modèle de consensus. Notre but est de construire un consensus sur la base de modèles utiles et capables de prédire des sous-espaces chimiques propres à chacun.

1.3.1. Identification des modèles uniques

Afin de caractériser cette redondance, nous nous sommes demandés quels étaient les points cruciaux permettant de définir l'identité d'un modèle. Un modèle de prédiction est basé sur un espace chimique qui lui est propre et un ensemble de descripteurs, afin de prédire une propriété ADME-Tox pour de nouveaux composés. Ainsi, nous avons considéré les facteurs relatifs à i) l'espace chimique exploré, ii) les

descripteurs employés, et pour finir iii) les prédictions obtenues, dans le but de caractériser la redondance des modèles à notre disposition. La méthodologie que nous avons mise en place pour la création d'un modèle de consensus comporte trois étapes.

Dans un premier temps, les modèles validés sont ordonnés en fonction de leur score (M_{score}) par ordre décroissant. Les modèles ordonnés sont ensuite comparés deux à deux selon les trois critères définis auparavant : i) le pourcentage de molécules communes entre les jeux d'apprentissages est calculé, ii) le pourcentage de descripteurs communs est défini, et iii) le coefficient de corrélation (R^2 ou MCC) entre les valeurs prédites est déterminé. En résulte trois matrices symétriques pouvant être considérées comme des matrices de similarité entre les modèles.

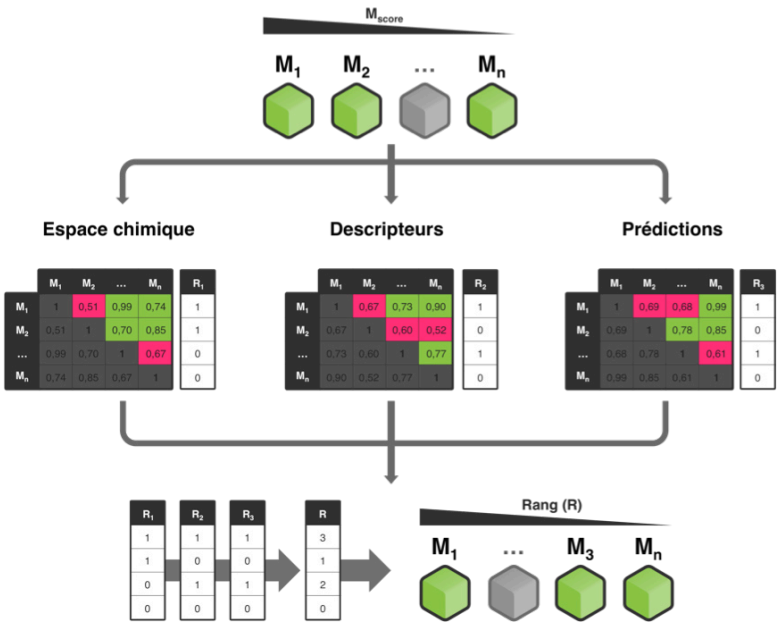


Figure 51 : Identification des modèles uniques pour la création d'un consensus.

Dans un deuxième temps, les matrices d'axes de symétrie diagonale sont traitées de telle sorte que la partie inférieure et la diagonale sont supprimées. Un seuil de similarité (par défaut 0,70) est ensuite appliqué à ces trois matrices, afin d'identifier les modèles prenant en considération l'information traitée par d'autres modèles de score inférieur (Figure 51). Un rang est attribué à chaque modèle en fonction du nombre de critères remplis. Ce rang peut varier entre 3 et 0, avec un rang 3 pour les modèles uniques sur les trois critères étudiés, et un rang 0 pour les modèles non uniques sur les trois critères étudiés.

Dans un troisième temps, les modèles sont réordonnés de manière décroissante en fonction de leur rang et de leur score, avant d'être successivement ajouté au modèle de

consensus. Lors de chaque ajout, l'erreur du consensus est estimée afin de confirmer l'utilisation du modèle ajouté. Seuls les modèles permettant de réduire l'erreur du consensus sont sélectionnés pour continuer l'incorporation des modèles suivants.

Le modèle de consensus peut ensuite être utilisé pour la prédiction de nouvelles molécules. Dans le cas d'une propriété continue, la moyenne arithmétique des valeurs prédites à partir de tous les modèles uniques est calculée. L'information transmise par le modèle de consensus comprend également la déviation standard sur l'ensemble des prédictions issues des modèles uniques. Pour une propriété discrète, le vote majoritaire de tous les modèles uniques est appliqué. La probabilité d'activité moyenne ainsi que sa déviation standard sont également des informations partagées par le modèle de consensus.

1.3.2. Validation des modèles de consensus

La validation nécessite un jeu de données inconnu du modèle de consensus, mais également des modèles uniques qui le composent. Cet ensemble de données va constituer le jeu de données externe. La création du jeu de données externe est faite à partir de l'ensemble des données initiales et ceci dès le début du processus de modélisation (**Figure** 24).

Afin de créer un ensemble de données externe représentatif du jeu de données initial, nous avons choisi la méthode de regroupement Butina (Ch2 3.2.3.1.c)). Cette méthode a été privilégiée car elle est indépendante des ensembles de descripteurs étudiés. Les groupes de composés sont ordonnés en fonction de leur taille et de manière décroissante. Le jeu de données externe est définit en sélectionnant successivement 25 % des composés de chaque groupe, jusqu'à atteindre une taille correspondant à 25 % du jeu de données initial. Une vérification basée sur la propriété (Y) est réalisée pour vérifier que le jeu de données externe est représentatif de l'ensemble des données initiales. Si la distribution de la propriété modélisée ou le rapport actif/inactif ne sont pas identiques entre les jeux de données, la procédure est réitérée.

1.3.3. Application au modèle de régression LogS

Lors de l'étude du jeu de données LogS, nous avons souhaité définir un modèle de régression nous permettant de prédire une valeur continue de la solubilité aqueuse. De plus, nous avons vu précédemment que l'ionisation des molécules était indispensable pour expliquer correctement les valeurs de solubilité observée pour certains composés.

D'une part, le jeu de données a été standardisé et ionisé à pH 7 pour l'obtention des résultats qui sont présentés dans le cadre de cette partie. D'autre part, nous avons choisi de déterminer quel était le meilleur compromis entre les algorithmes de régression linéaire multiple (MLR) et de machine à vastes marges (SVM), mais également entre les ensembles de descripteurs moléculaires RDKit, CDK et MORDRED. Pour chacune des combinaisons explorées entre l'algorithme et les ensembles de descripteurs, l'exploration de l'espace chimique (jeu d'apprentissage et jeu de test) a été effectuée 30 fois. Au total, 180 (2 x 3 x 30) modèles ont été définis pour lesquels 71 ont été validés selon les critères de sélection énoncés précédemment (Ch3 1.2.5). Notre premier objectif a été d'identifier les conditions optimales pour la création d'un modèle linéaire du LogS (Figure 52).

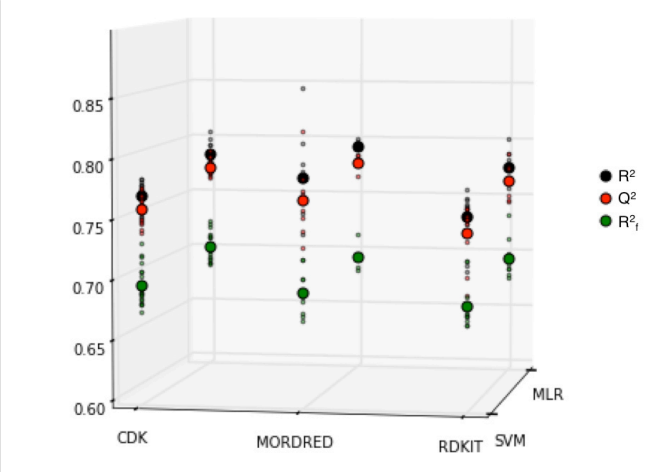


Figure 52 : Performances des 71 modèles linéaires validés pour la prédiction du LogS.

Les points de taille minimale représentent les performances des modèles individuels. Les points de taille maximale représentent les performances moyennes de tous les modèles.

L'analyse des performances moyennes nous montre que les algorithmes testés ont des performances comparables quelque soit l'ensemble de descripteurs étudié. On remarque également que les ensembles de descripteurs MORDRED et CDK semblent être plus appropriés pour modéliser le LogS que l'ensemble de descripteurs RDKit. L'analyse des performances des modèles individuels nous montre que les résultats sont plus variables avec la SVM qu'avec la MLR, et ceci quelque soit l'ensemble de descripteurs sur lequel les modèles sont basés. Nous pouvons expliquer cette observation par le fait que la SVM est paramétrée au début du processus de modélisation, tandis que la MLR ne l'est pas. De ce fait, les paramètres de la SVM s'adaptent à l'espace chimique qu'elle explore. Sachant que nos étapes de traitement des descripteurs sont basées sur une stratégie *Embedded*, les descripteurs vont être sélectionnés en fonction des particularités internes du modèle SVM et ceci pour un espace chimique spécifique. Pour cette raison, des performances variables peuvent être observées dans le cas de la SVM en fonction du jeu

d'apprentissage exploré. Après vérification, nous nous sommes rendu compte que les modèles SVM obtenus dans des conditions équivalentes ne sélectionnaient pas le même nombre et le même type de descripteurs, tandis que dans le cas de la MLR des descripteurs similaires sont observés entre tous les modèles. En somme, ces résultats ne permettent pas de différencier un ensemble de descripteurs ou un algorithme pour optimiser la modélisation du LogS. La seule conclusion que nous pouvons faire est que l'ensemble de descripteurs RDKit semble être légèrement le moins adapté pour modéliser cette propriété.

Les 71 modèles valides ont été exploités afin de générer un modèle de consensus. La Figure 53 présente l'erreur globale du modèle de consensus lors de l'ajout successifs des modèles uniques. Il est important de noter que ces derniers ont été ordonnés en fonction de leur rang et de leur M_{score}. Si l'erreur du modèle de consensus augmente lors de l'ajout d'un modèle unique, ce dernier n'est pas retenu pour les ajouts suivants. Par exemple pour l'ajout du modèle 4, l'erreur globale observée est celle du modèle de consensus construit à l'aide des modèles uniques 1, 2 et 4, le modèle unique 3 ayant été supprimé de l'étude. Cette étape est réitérée jusqu'à ce que l'ensemble des modèles ait été testé.

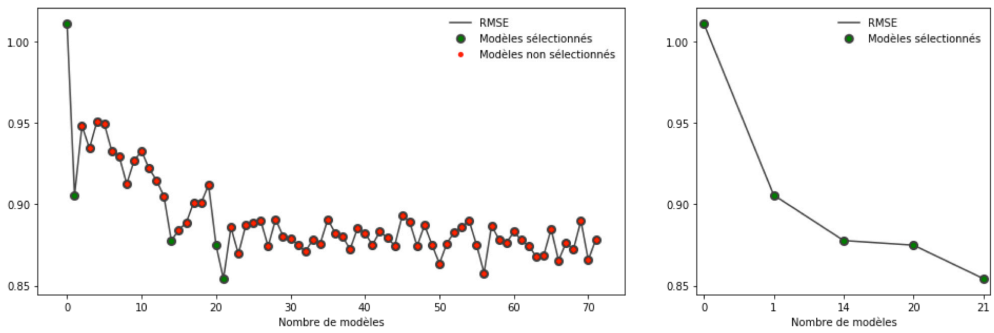


Figure 53 : Erreur du modèle de consensus lors de l'ajout successive des modèles uniques.

Ces deux graphiques représentent le RMSE du consensus en fonction du nombre de modèles testés. Le graphique de gauche présente les résultats de l'ensemble des modèles, tandis que le graphique de droite présente uniquement l'erreur globale du consensus en fonction des 5 modèles uniques sélectionnés.

On remarque que les modèles les mieux classés ne sont pas toujours ceux sélectionnés en priorité. Au final, seulement 5 modèles uniques ont été sélectionnés car ils permettent de diminuer l'erreur globale du modèle de consensus en passant de 1.01 à 0.85. Les paramètres de ces 5 modèles vous sont présentés dans la Table 14. On observe également que le modèle unique 20 réduit faiblement l'erreur global du consensus et qu'il est similaire au modèle 14.

Il arrive parfois qu'aucun modèle unique ne permette de diminuer l'erreur du consensus. Dans ce cas, seul le modèle de plus haut M_{score} est sélectionné pour la prédiction de nouveaux jeux de données.

Modèle	Algorithme	Ensemble	N	Descripteurs
M0	SVM	RDKIT	13	PEOE_VSA10, SMR_VSA5, PEOE_VSA8, PEOE_VSA9, Ipc, SlogP_VSA3, MinPartialCharge, PEOE_VSA1, SlogP_VSA2, Chi1, SMR_VSA10, HeavyAtomMolWt, MolLogP
M1	MLR	MORDRED	26	ATS1dv, ATS5v, Xp-3dv, Xpc-6d, BCUTs-1h, BCUTp-1l, AMID_O, AXp-1d, ATSC0m, AMID, PEOE_VSA7, GGI4, GATS1i, AATS3Z, AATS0d, ATS7Z, BCUTv-1l, SMR_VSA7, AATS1i, piPC10, ATS4Z, AATS4i, AATS1p, PEOE_VSA6, ZMIC2, SLogP
M14	MLR	CDK	21	WNSA-2, geomShape, WTPT-5, SC-5, MDEO-11, FNSA-3, VP-7, VPC-6, VP-5, TPSA, WTPT-2, FPSA-3, ALogp2, SP-6, SP-1, WTPT-1, tpsaEfficiency_x, MW, ALogP, ATSm4, XLogP
M20	MLR	CDK	21	PNSA-3, geomShape, WTPT-4, Kier2, SC-5, MDEO-11, WPATH, VPC-6, FPSA-3, SP-2, SP-6, RPSA, ATSp3, ALogp2, ATSp4, ATSp1, VP-0, THSA, ALogP, ATSm5, XLogP
M21	SVM	MORDRED	133	SM1_Dzare, EState_VSA2, AATSC1se, SddssS, AATSC5s, MATS4are, ETA_epsilon_4, AATSC1d, ATSC6Z, MATS4c, AETA_eta_RL, AATSC5are, AATSC2p, CIC5, AATS0are, IC0, AATS2are, ATSC1v, MATS3Z, BCUTv-1h, AATSC2se, TIC1, GGI10, GATS3c, ATSC2c, MATS3d, JGI10, GATS5p, MATS3dv, MATS1se, ATS4i, ATSC3c, AATS4dv, MATS6p, AATS3dv, EState_VSA6, AATSC0p, AATSC3s, GATS3m, AETA_eta, ATSC7m, ATSC4s, AATS6s, TIC0, GATS3v, ATS7s, ETA_eta_B, GATS5d, MATS2d, BCUTdv-1l, AATSC6d, AATSC4s, SsOH, SsssCH, GATS2s, MATS1p, SlogP_VSA3, GATS2i, AATSC2c, JGI9, MATS2c, AATS5s, ATS6dv, SaaaC, ATSC1s, JGI6, ATS8i, ATSC0s, AATS4s, SpDiam_Dzp, GATS2d, ETA_epsilon_2, ATS5dv, AATS1s, JGI7, IC3, ATSC0p, AATS2s, Xp-7dv, Xc-5dv, CIC0, GATS6v, GGI8, BCUTm-1h, SMR_VSA1, AATSC0dv, AETA_eta_F, SM1_Dzp, AMID_N, AATSC1s, Xpc-6dv, Xc-3dv, ATS2dv, AATS1Z, AETA_beta, ATS3dv, GGI7, EState_VSA3, SpMAD_A, TSRW10, GATS2v, GGI5, TopoPSA(NO), SaasC, AATSC1c, ZMIC5, VR3_A, AXp-3d, Xpc-6d, ETA_eta_FL, AXp-4d, PEOE_VSA7, ATSC0m, BCUTv-1l, AATS0m, AATS6Z, ATS7Z, AATS5Z, AATS3d, AATS3Z, SMR_VSA7, piPC5, AATS4Z, AATS6i, piPC9, AATS3i, AATS4d, AATS2i, AATS5d, AATS5i, AATS3p, ZMIC2, SLogP

Table 14 : Modèles uniques combinés pour définir le modèle de consensus.

Le modèle de consensus est ensuite validé à l'aide du jeu de données externe (Figure 54). Seules les composés identifiés dans le DA de chaque modèle sont utilisés pour calculer les performances externes. La validation interne est réalisée à l'aide du jeu de données initial et selon la même démarche que la validation externe.

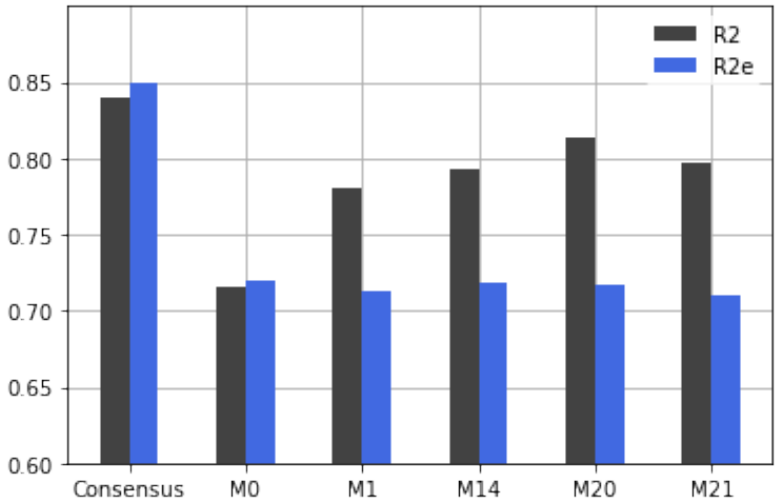


Figure 54 : Validation du modèle de consensus.

Le R^2 a été déterminé à partir du jeu de données initial (jeu d'apprentissage et jeu de test) en ne prenant en considération que les molécules considérées comme étant dans le domaine d'applicabilité. Le R_e^2 a été déterminé à partir du jeu de données externe en ne prenant en considération que les molécules présentes dans le domaine d'interpolation.

Selon les validations interne et externe, le modèle de consensus possède de meilleures performances que les modèles uniques. On remarque que la combinaison de ces 5 modèles uniques apporte une amélioration moyenne de 6 points pour la validation interne et de 12 points pour le coefficient de détermination sur le jeu de données externe. On en déduit que le consensus permet d'obtenir des prédictions plus fiables que celles provenant d'un modèle unique. Concernant l'utilité des prédictions issues d'un modèle de consensus comparé à celles obtenues à partir d'un modèle unique, nous proposons les résultats présentés dans la Table 15 et la Table 16.

La Table 15 présente les prédictions de LogS issues d'un modèle simple (M0). Les résultats proposés comportent la structure des composés testés (SMILES), la valeur expérimentale du LogS (Y_{exp}), la valeur prédite du LogS (Y_{pred}), le statut du DA avec *IN* pour les molécules considérées dans le DA et *OUT* pour les molécules considérées en dehors du DA, et pour finir la structure des composés du jeu d'apprentissage pour lesquels la molécule testée possède une distance inférieure à leurs seuils t_i (AD $_{neighbors}$). Dans le cas des modèles simples, seule l'information sur le DA peut nous permettre de juger la qualité d'une prédiction. L'information sur les plus proches voisins peut permettre de définir un indice de similarité structurale. Cependant, cet indice peut ne pas être représentative de ce que le modèle est capable de prédire, car ce dernier est construit sur la base de descripteurs moléculaires et non structuraux. Par conséquent, le DA prend en considération l'information moléculaire du modèle, ce qui induit que des molécules de

structures très différentes peuvent être considérées comme similaires en fonction de leur comportement moléculaire. Par conséquent, selon ce principe, des sauts d'activité peuvent être observés pour des molécules très similaires.

SMILES	Y_{exp}	Y_{pred}	AD	AD $_{neighbors}$
Cc1c(Cl)ccc(Cc2ccc(Cl)cc2Cl)c1Cl	-7,40	-7,54	IN	[Cc1c(Cl)ccc(Cc2ccc(Cl)cc2Cl)c1Cl]
Cc1cc(Cl)c(Cl)c(Cl)c1Cc1ccccc1Cl	-7,24	-7,67	IN	[Cc1cc(Cl)c(Cl)c(Cl)c1Cc1ccccc1Cl, Cc1cc(Cc2ccc(Cl)cc2)c(Cl)c(Cl)c1Cl]
Clc1cc(Cl)c(-c2ccccc2Cl)cc1Cl	-6,57	-6,70	IN	[Clc1cccc(-c2cccc(Cl)c2Cl)c1Cl, Clc1cc(-c2ccccc2)c(Cl)c(Cl)c1Cl]
Clc1ccc(-c2c(Cl)cc(Cl)cc2Cl)cc1	-6,94	-6,63	IN	[Clc1ccc(-c2ccc(Cl)cc2Cl)c(Cl)c1]
Cn1cnc([N+](=O)[O-])c1Sc1ncnc2[nH]cnc12	-2,62	-2,05	IN	[Cn1cnc([N+](=O)[O-])c1Sc1ncnc2[nH]cnc12]
CC(C(=O)[O-])c1cccc(Oc2ccccc2)c1	-3,60	-3,06	IN	[O=C(N=c1cn[n-]s1)Nc1ccccc1, O=C(c1cccc(F)c1)N(O)c1ccccc1, CC(C(=O)[O-])c1cccc(Oc2ccccc2)c1, [O-]C(=NO)c1ccccc1N=Cc1ccccn1]
NC(=O)c1cnc2n1CCc1ccccc1C21CCC[NH2+]CC1	-2,43	-2,63	IN	[NC(=O)c1cnc2n1CCc1ccccc1C21CCC[NH2+]CC1]
Cc1ccc2c(c1)C=CS2(=O)=O	-2,65	-2,48	IN	[Cc1ccc2c(c1)C=CS2(=O)=O, Cc1ccc(S(=O)(=O)[N-]Cl)cc1]
Cc1cc(=O)oc2c3c4c(cc12)CCCN4CCC3	-3,89	-3,88	IN	[Cc1cc(=O)oc2c3c4c(cc12)CCCN4CCC3, [O-]C(=NO)c1ccccc1N=Cc1ccccn1]
CC1(C)SCC(COC(N)=O)S1	-2,22	-2,18	IN	[Cc1cc(Cl)ccc1OCC(=O)[S-], N=c1[n-]nc(-c2ccc(O)cc2)o1]
COC(=O)c1ccccc1C(=O)c1ccccc1	-3,68	-3,71	IN	[COC(=O)c1ccccc1C(=O)c1ccccc1, CCN(CC)c1ccc2c(C)cc(=O)oc2c1]
c1ccc(C[NH2+]C23OC4C5C6CC(C7C6C4C72)C53)cc1	-3,40	-4,64	OUT	[]
CC1(C)OC(=O)c2cc(C#N)ccc21	-2,91	-2,75	IN	[CC1OC(=O)c2cc([N+](=O)[O-])ccc21]
CCCCCCNC(=O)c1cccnc1	-2,52	-2,38	IN	[CCCCCCCNC(=O)c1cccnc1, CCCCCCNC(=O)c1cccnc1, CCCCCCOC(=O)c1cccnc1]

Table 15 : Prédictions d'un modèle unique.

SMILES	Y_{exp}	Y_{pred}	Y_{std}	AD	AD_{agree}
Cc1c(Cl)ccc(Cc2ccc(Cl)cc2Cl)c1Cl	-7,40	-7,21	0,44	IN	1,0
Cc1cc(Cl)c(Cl)c(Cl)c1Cc1ccccc1Cl	-7,24	-7,34	0,37	IN	1,0
Clc1cc(Cl)c(-c2ccccc2Cl)cc1Cl	-6,57	-6,69	0,26	IN	1,0
Clc1ccc(-c2c(Cl)cc(Cl)cc2Cl)cc1	-6,94	-6,87	0,31	IN	1,0
Cn1cnc([N+](=O)[O-])c1Sc1ncnc2[nH]cnc12	-2,62	-2,77	0,70	IN	0,8
CC(C(=O)O)c1cccc(Oc2ccccc2)c1	-3,60	-3,75	0,30	IN	1,0
NC(=O)c1cnc2n1CCc1ccccc1C21CCNCC1	-2,43	-2,49	0,18	OUT	1,0
Cc1ccc2c(c1)C=CS2(=O)=O	-2,65	-2,50	0,24	IN	0,8
Cc1cc(=O)oc2c3c4c(cc12)CCCN4CCC3	-3,89	-3,80	0,42	IN	1,0
CC1(C)SCC(COC(N)=O)S1	-2,22	-2,14	0,17	IN	0,8
COC(=O)c1ccccc1C(=O)c1ccccc1	-3,68	-3,69	0,38	IN	1,0
c1ccc(CNC23OC4C5C6CC(C7C6C4C72)C53)cc1	-3,40	-3,43	0,79	OUT	0,8
CC1(C)OC(=O)c2cc(C#N)ccc21	-2,91	-2,82	0,26	IN	1,0
CCCCCCNC(=O)c1cccnc1	-2,52	-2,42	0,18	IN	1,0

Table 16 : Prédictions du modèle de consensus.

La Table 16 présente les prédictions de LogS issues du modèle de consensus, avec la structure des composés testés (SMILES), les mesures expérimentales (Y_{exp}) et les valeurs prédites (Y_{pred}) du LogS, mais également la déviation standard de la valeur prédite sur l'ensemble des modèles simples (Y_{std}) et l'information sur le DA. Le DA comporte la concordance entre tous les modèles (AD_{agree}) qui doit être analysée conjointement avec l'information sur le DA. Ainsi, lorsqu'un composé est considéré comme étant dans le DA et que la concordance de tous les modèles est de 1, cela signifie que le composé testé est considéré dans le DA de tous les modèles uniques utilisés par le consensus. Ce type de modèle permet de disposer de plus d'informations utiles qui permettent de juger la qualité d'une prédiction comme par exemple la déviation standard de la valeur prédite.

1.4. Modèles ADME-Tox développés

Cette plateforme nous a permis de créer des modèles de prédiction pour plusieurs propriétés ADME-Tox. Les modèles de classification sont présentés dans la Table 17, et les modèles de régression sont renseignés dans la **Table** 18. Actuellement, les modèles de prédictions pour 21 propriétés ADME-Tox ont été générés par la plateforme MetaPredict.

PROPRIÉTÉ	DESCRIPTION	ALGORITHME	DESCRIPTEURS	VALIDATION	# MOLÉCULE	AUC	MCC	SENSIBILITÉ	SPÉCIFICTÉ	PRÉCISION
ASBT_I	Modèle de classification du transporteur humain ASBT (inibiteur / non-inhibiteur)	RF	IIB24	Apprentissage	32	0,99	0,94	0,94	1	0,97
				Validation croisée	-	0,98	0,97	0,96	0,94	0,95
				Test	21	1	1	1	1	1
ASBT_S	Modèle de classification du transporteur humain ASBT (substrat / non-substrat)	SVM	IIB24	Apprentissage	26	1	1	1	1	1
				Validation croisée	-	0,99	0,98	0,97	0,94	0,96
				Test	17	1	1	1	1	1
BCRP_I	Modèle de classification du transporteur humain BCRP (inibiteur / non-inhibiteur)	SVM	IIB24	Apprentissage	150	0,98	0,83	0,87	0,96	0,92
				Validation croisée	-	0,97	0,86	0,86	0,95	0,91
				Test	69	0,91	0,71	0,85	0,86	0,86
BCRP_S	Modèle de classification du transporteur humain BCRP (substrat / non-substrat)	RF	RDKit	Apprentissage	62	0,99	0,97	0,97	1	0,98
				Validation croisée	-	0,96	0,97	0,93	0,9	0,92
				Test	27	0,89	0,78	0,93	0,84	0,89
MRP2_I	Modèle de classification du transporteur humain MRP2 (inibiteur / non-inhibiteur)	RF	IA25	Apprentissage	30	0,96	0,94	0,93	1	0,97
				Validation croisée	-	0,95	0,91	0,89	0,95	0,92
				Test	15	1	1	1	1	1
MRP2_S	Modèle de classification du transporteur humain MRP2 (substrat / non-substrat)	RF	IA25PFC	Apprentissage	58	0,99	0,97	1	0,97	0,98
				Validation croisée	-	0,97	0,95	0,95	0,89	0,92
				Test	33	0,98	0,88	1	0,90	0,93
OCT1_I	Modèle de classification du transporteur humain OCT1 (inibiteur / non-inhibiteur)	SVM	RDKit	Apprentissage	66	0,96	0,88	0,97	0,91	0,94
				Validation croisée	-	0,95	0,86	0,94	0,87	0,91
				Test	33	0,99	0,94	0,94	1	0,97
OCT1_S	Modèle de classification du transporteur humain OCT1 (substrat / non-substrat)	SVM	CDK	Apprentissage	28	1	1	1	1	1
				Validation croisée	-	0,99	0,98	0,98	0,97	0,97
				Test	14	1	1	1	1	1
PEPT1_I	Modèle de classification du transporteur humain PEPT1 (inibiteur / non-inhibiteur)	SVM	MORDRED	Apprentissage	24	0,92	0,68	0,92	0,75	0,84
				Validation croisée	-	0,88	0,71	0,90	0,77	0,84
				Test	13	0,83	0,69	0,86	0,83	0,85
PEPT1_S	Modèle de classification du transporteur humain PEPT1 (substrat / non-substrat)	RF	IIB24	Apprentissage	46	0,98	0,91	0,96	0,96	0,96
				Validation croisée	-	0,97	0,91	0,91	0,94	0,93
				Test	26	0,96	0,85	1	0,88	0,92
MDR1_I	Modèle de classification du transporteur humain MDR1 (inibiteur / non-inhibiteur)	SVM	MORDRED	Apprentissage	900	0,91	0,70	0,87	0,83	0,85
				Validation croisée	-	0,91	0,71	0,87	0,82	0,85
				Test	575	0,90	0,71	0,86	0,87	0,84
MDR1_S	Modèle de classification du transporteur humain MDR1 (substrat / non-substrat)	SVM	IA25	Apprentissage	155	0,79	0,42	0,81	0,58	0,66
				Validation croisée	-	0,78	0,43	0,76	0,52	0,61
				Test	144	0,69	0,24	0,84	0,50	0,49
Fup	Modèle local de classification pour la fraction libre dans le plasma (Fu < 10 / Fu > 10)	LGR	RDKit	Apprentissage	156	0,88	0,71	0,85	0,86	0,85
				Validation croisée	-	0,86	0,66	0,84	0,80	0,82
				Test	88	0,91	0,72	0,88	0,85	0,78
Fup	Modèle global de classification pour la fraction libre dans le plasma (Fu < 10 / Fu > 10)	**CONSENSUS** LGR SVM	**CONSENSUS** RDKit CDK MORDRED	Apprentissage	1193	0,83	0,66	0,85	0,81	0,83
				Externe	428	0,77	0,55	0,85	0,69	0,78

PROPRIÉTÉ	DESCRIPTION	ALGORITHME	DESCRIPTEURS	VALIDATION	# MOLÉCULE	AUC	MCC	SENSIBILITÉ	SPÉCIFICTÉ	PRÉCISION
CYP2C9_S	Modèle de classification du CYP450 2C9 (substrat / non-substrat)	SVM	IB27	Apprentissage	122	0,72	0,32	0,79	0,52	0,67
				Validation croisée	-	0,71	0,32	0,78	0,52	0,66
				Test	122	0,70	0,33	0,80	0,56	0,65
CYP3A4_S	Modèle de classification du CYP450 3A4 (substrat / non-substrat)	SVM	MORDRED	Apprentissage	256	0,82	0,56	0,78	0,78	0,78
				Validation croisée	-	0,80	0,53	0,74	0,76	0,75
				Test	123	0,69	0,46	0,8	0,66	0,73
CYP2D6_S	Modèle de classification du CYP450 2D6(substrat / non-substrat)	SVM	MORDRED	Apprentissage	168	0,87	0,69	0,86	0,83	0,85
				Validation croisée	-	0,86	0,67	0,82	0,82	0,82
				Test	121	0,82	0,56	0,79	0,81	0,76
BBB	Modèle local de classification du passage de la barrière hémato-encéphalique	RF	IIB24	Apprentissage	298	1,00	0,97	1	0,97	0,99
				Validation croisée	-	0,99	0,98	0,99	0,96	0,97
				Test	295	0,97	0,91	0,99	0,91	0,96

Table 17 : Modèles de classification générés par la plateforme MetaPredict.

PROPRIÉTÉ	DESCRIPTION	ALGORITHME	DESCRIPTEURS	VALIDATION	# MOLÉCULE	R2	Q2	R2f
LogS	Modèle de régression de solubilité aqueuse	**CONSENSUS MLR SVM**	**CONSENSUS RDKit MORDRED CDK**	Initial / Externe	2755 / 1118	0,84	-	0,85
LogP	Modèle de régression du coefficient de partition eau/octanol	MLR	CDK	Apprentissage / Test	2091 / 765	0,88	0,88	0,85
LogD7.4	Modèle de régression du coefficient de partition eau/octanol au pH 7,4	MLR	MORDRED	Apprentissage / Test	2748 / 1294	0,85	0,84	0,83
Fup	Modèle de régression de la fraction libre dans la plasma	MLR	RDKit	Apprentissage / Test	204 / 97	0,74	0,66	0,61

Table 18 : Modèles de régression générés par la plateforme MetaPredict.

1.5. Langage de programmation et aspects pratiques

Dans un premier temps, cette plateforme a été déployée à l'aide du langage de programmation R. Ce logiciel libre a été choisi en priorité pour la diversité des outils statistiques qu'il propose pour la mise en place d'un apprentissage automatique et leur validation par la communauté. Cependant, nous avons dû abandonner ce langage de programmation, car : i) d'une part, la plupart des outils couramment utilisés en chémoinformatique ne sont pas supportés, et ii) d'autre part, la gestion des ressources informatiques n'est pas optimisée. De ce fait, Python a été choisi dans un second temps pour la diversité des outils dédiés à la chémoinformatique et également ses avantages en termes de gestion des ressources informatiques. La plateforme MetaPredict est aujourd'hui disponible sous la forme d'une librairie Python pouvant être installée sur les systèmes d'exploitation Linux et Mac OS. Cette librairie est constituée de 10 modules qui contiennent 13 423 lignes de codes nécessaires pour le fonctionnement de la plateforme.

De plus, cette plateforme est orientée objet, c'est-à-dire qu'elle va fournir des modèles autonomes. Chaque modèle possède l'ensemble des paramètres utilisés pour le créer, ainsi que l'ensemble des procédures employées pour le générer. De ce fait, un modèle va prendre en considération plusieurs caractéristiques lors de la prédiction de nouveaux ensembles de données, comme par exemple la standardisation des molécules si nécessaire, l'ionisation des molécules à pH spécifique si nécessaire, le calcul des descripteurs utilisés par le modèle, la normalisation des descripteurs en fonction des contraintes liées au jeu d'apprentissage, la prédiction de la propriété des composés testés, et pour finir l'étude du domaine d'applicabilité. Ceci est un réel avantage qui nous permet de déployer rapidement des modèles (Q)SAR autonomes, tout en assurant le suivi de ces derniers.

1.6. Limitations et perspectives de l'outil

La plateforme MetaPredict est un outil à la disposition du modélisateur pour l'aider à créer le plus efficacement possible des modèles de prédiction pour diverses applications. Le développement de l'outil a été mené de façon à respecter les bonnes pratiques (Q)SAR. Les limites de MetaPredict exposées dans le cadre de ce paragraphe sont celles rencontrées au cours d'un fonctionnement de la plateforme dans le cas général.

La première limitation concerne les étapes liées à la standardisation des structures chimiques. En effet, comme nous l'avons vu précédemment notre protocole de standardisation ne prend pas en compte l'identification des composés organométalliques pour le moment. Par conséquent, il serait avantageux de pouvoir les identifier dès les étapes précoces du processus de modélisation. Nous avons également remarqué qu'aucun logiciel libre ne permet de protoner les structures à un pH spécifique comme le propose ChemAxon. Il serait probablement intéressant de proposer une alternative publique pouvant répondre à cette demande, ce qui nous permettrait d'être non dépendant d'un logiciel commercial.

La deuxième limitation concerne la diversité des algorithmes à disposition pour la création des modèles de prédiction. Seuls les algorithmes les plus courants ont été initialement introduit dans la plateforme. Il serait néanmoins avantageux d'incorporer d'autres algorithmes qui apportent des modèles de plus hautes performances, comme par exemple l'utilisation des forêts aléatoires pour la prédiction de propriétés continues.

La troisième limitation est la rapidité de la plateforme lors des différentes étapes d'optimisation. La plateforme prend en moyenne 3 jours sur une machine de 8 CPU et 32 G de mémoire vive pour générer les modèles (Q)SAR spécifiques d'une propriété ADME-Tox, pour laquelle nous avons sélectionné 3 algorithmes, 10 ensembles de descripteurs et 30 découpes du jeu de données. La découpe du jeu de données est essentielle pour explorer différents sous-espaces chimiques afin de produire dans la mesure du possible des modèles uniques indispensables pour la création d'un modèle de consensus. A l'heure actuelle, cette découpe est répétée 30 fois sans vérification par plateforme. Elle est donc susceptible de proposer des sous-espaces chimiques déjà explorés. Il serait plus avantageux de créer ces découpes en vérifiant si l'espace chimique nouvellement exploré est redondant et ceci dès les étapes de création des modèles simples. Cette perspective représente une manière efficace de réduire le nombre de découpes nécessaires à l'exploration de sous-espaces chimiques et par conséquent, du temps nécessaire à la création des modèles. Une autre étape qui limite la rapidité de notre outil est la recherche exhaustive par grille pour le paramétrage des modèles. Cette optimisation nécessite la détermination des performances du modèle pour chaque combinaison de paramètres. Ainsi, plus le nombre de paramètres est important, plus la détermination des paramètres optimums est longue. Cependant, cette étape est indispensable pour l'élaboration de modèles performants. L'approche envisagée pour répondre à cette limitation comporte deux étapes. La première étape consiste à effectuer

une recherche par grille contenant moins de paramètres que dans le cas d'une recherche exhaustive. Elle a pour but d'échantillonner grossièrement l'ensemble des paramètres afin de définir la combinaison la proche de la solution idéale. La deuxième étape consiste à rechercher les paramètres optimums du modèle à l'aide d'une optimisation simplexe de Nelder-Mead [259]. Pour chaque combinaison testée le simplexe a pour objectif de rechercher les paramètres qui augmentent le Q^2. L'avantage de cette optimisation est qu'elle est évolutive et capable d'optimiser simultanément plusieurs paramètres, ce qui limite le nombre de combinaison à explorer qui sont nécessaires pour identifier les paramètres optimums du modèle.

La quatrième limitation concerne les aspects pratiques liés aux modèles de consensus. Les prédictions proposées par un modèle de consensus sont actuellement obtenues en moyennant les prédictions de tous modèles simples qui le composent. Cependant, il est possible qu'une molécule d'intérêt soit considérée en dehors du domaine d'interpolation de certains modèles simples. Par conséquent, la valeur prédite par le modèle de consensus prend en considération des valeurs interpolées de la propriété modélisée. Ainsi, nous pensons qu'il serait plus avantageux de n'utiliser que les prédictions interpolables pour fournir une estimation plus juste des propriétés modélisées à l'aide des modèles de consensus.

En somme, nous pensons que le couplage de la plateforme MetaPredict à la base de données ADMET db serait la perspective la plus prometteuse pour la pérennité de ce projet. Ceci peut permettre d'assurer la mise à jour des modèles en fonction des nouvelles données incorporées à la base. L'objectif serait alors d'assurer le suivi des modèles dans le temps et de pouvoir faire évoluer notre collection de modèles ADME-Tox en fonction des projets et des thématiques développées au sein du laboratoire.

2. Validation de la plateforme MetaPredict

Nous avons choisi de valider l'outil sur les modèles de prédiction de la fraction libre dans le plasma (Fu_p). Cette propriété ADME-Tox est la première que nous avons essayé d'exploiter avec la plateforme MetaPredict. La Fu_p a été sélectionnée, car elle est considérée comme une propriété charnière qui influence les phénomènes ADME-Tox liés à la distribution, au métabolisme et à l'élimination d'un médicament.

Comme nous l'avons vu précédemment, les médicaments ingérés sont dans un premier temps absorbés par l'intestin, puis entrent dans la circulation sanguine, où ils interagissent avec les protéines plasmatiques du sang. Si la majorité du médicament est liée aux protéines plasmatiques, seule la fraction non liée, dite fraction libre, peut avoir un effet thérapeutique ou être métabolisée et excrétée. Par conséquent, le degré de liaison aux protéines plasmatiques influence considérablement les propriétés pharmacocinétiques et pharmacodynamiques d'un médicament. L'efficacité de ce dernier sera induite par la quantité de principe actif non lié dans le plasma, c'est-à-dire la proportion de médicament libre de pénétrer dans les tissus environnants. La fraction liée du principe actif peut également servir de réservoir pour le médicament libre soustrait par divers processus d'élimination. Ce phénomène permet alors de prolonger la durée d'action du médicament. On en déduit que la clairance hépatique (CL_H) et la clairance rénale (CL_R) peuvent être réduites lorsque le médicament est fortement fixé aux protéines plasmatiques, ce qui engendre une augmentation du temps de demi-vie ($t_{1/2}$).

Généralement, les médicaments qui se fixent très fortement aux protéines plasmatiques ($Fu_p < 1$ %) et qui disposent d'une dissociation lente peuvent ne pas être en mesure d'exercer leur effet thérapeutique avant d'être éliminés de l'organisme. Une faible fraction libre peut engendrer un faible volume de distribution, ce qui implique un confinement du principe actif dans le compartiment sanguin, ainsi qu'une incapacité à être distribué dans les tissus. Ceci peut alors diminuer la concentration en principe actif au voisinage de la cible nécessaire pour impulser l'effet thérapeutique désiré. Inversement, si le principe actif se fixe moyennement ou faiblement aux protéines plasmatiques ($Fu_p > 20$ %) et dispose d'une dissociation rapide, la distribution, le métabolisme ou encore l'excrétion du médicament pourraient ne pas être modulés. Le médicament est alors rapidement distribué et éliminé de l'organisme, ce qui ne permet pas d'obtenir un effet thérapeutique prolongé. Pour l'étude de cette propriété, notre objectif était d'être en capacité de prédire

la Fu_p des molécules de façon à pouvoir identifier les composés fortement liés aux protéines plasmatiques, c'est-à-dire qui disposent d'une fraction libre inférieure à 10 %.

2.1. Modélisation de la fraction libre dans le plasma

Dans le cadre de notre développement autour de la Fu_p, nous avons élaboré trois modèles de prédiction de cette propriété : i) un modèle local de régression sur le jeu de données proposé par Obach *et al.* ; ii) un modèle local de classification sur le jeu de données Obach *et al.* ; iii) un modèle global de classification sur l'ensemble des données à notre disposition dans la base ADMET db. Ces trois modèles vont ainsi être présentés. Toutes les étapes de préparation de données et de création de modèle présentées ci-après ont été effectuées à l'aide de la plateforme MetaPredict.

2.1.1. Jeu de données Obach

Nous avons initié notre travail sur le jeu de données publié par Obach *et al.* [229] qui à l'intérêt d'être publique, d'avoir été vérifié manuellement par les auteurs, et qui dispose d'un nombre conséquent de molécules. En effet, ce jeu de données propose des valeurs de fraction libre dans le plasma pour 540 molécules. L'étape de standardisation des structures moléculaires a engendré la suppression de 7 composés. Le jeu de données que nous avons utilisé contient donc 533 molécules, qui ont été ionisées au pH sanguin (7,4).

2.1.2. Modèle local de régression

Nous avons dans un premier temps essayé de mettre en œuvre un modèle de régression pour l'estimation de cette propriété. Notre but était de proposer dans la mesure du possible un modèle de consensus. Pour cette raison, un jeu de données externe représentant 25 % du jeu de données initial a été créé. Sur le jeu de données restant, nous avons configuré la plateforme afin d'explorer l'ensemble des combinaisons existantes entre les algorithmes PLS, MLR et SVM ; les descripteurs RDKit, CDK, MORDRED et VolSurf+ ; ainsi que 30 découpes du jeu de données en jeu d'apprentissage (75 %) et jeu de test (25 %). Seuls les outils qui permettent de calculer des descripteurs moléculaires ont été employés pour l'élaboration de ce modèle linéaire. Sur les 360 modèles obtenus, seuls 33 modèles ont été validés. La plateforme n'a retenu qu'un modèle lors de la construction du consensus.

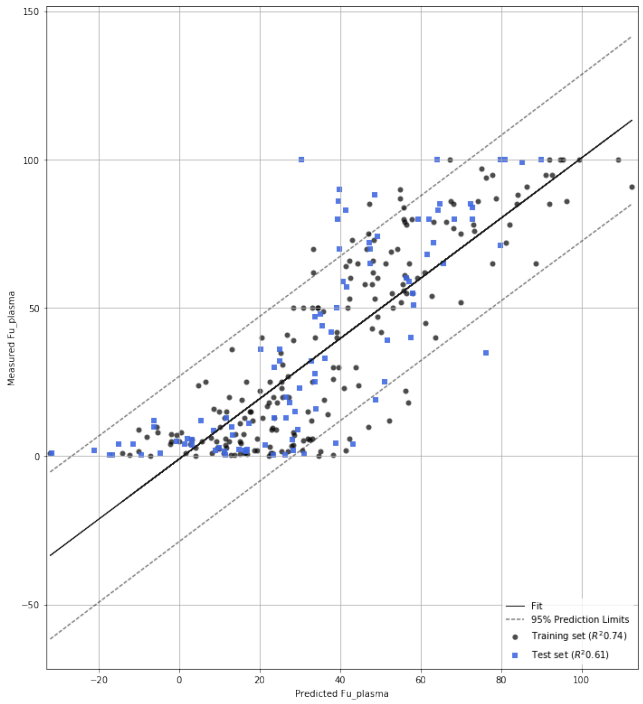


Figure 55 : Représentation de l'ajustement de Fu_p.

Ce modèle de régression linéaire multiple (MLR) est basé sur 24 descripteurs RDKit et il possède un R^2 de 0,74, Q^2 de 0,66, R^2_f de 0,61 et un R^2_e de 0,53. On remarque que ce modèle est peu prédictif sur le jeu de données externe. En visualisant l'ajustement de Fu_p (Figure 55), on observe que le modèle de régression ne parvient pas à prédire correctement les molécules ayant des valeurs de fraction libre extrêmes.

2.1.3. Modèle local de classification

Pour améliorer l'estimation des molécules fortement liées aux protéines plasmatiques, nous avons également mis en place un modèle local de classification. Le même jeu de données que précédemment a été utilisé. Lors de la création de ce modèle, nous avons étudié la discrétisation des données en fonction de différents seuils, à savoir 2 %, 5 %, 10 % et 20 %. Notre objectif était de déterminer quel était le seuil le plus avantageux pour obtenir un modèle performant pour la prédiction de nouvelles molécules. Pour optimiser le modèle nous avons également choisi de créer une zone de délétion avec chacun des seuils énoncés précédemment et une erreur expérimentale de 10 %. Nous avons estimé cette erreur expérimentale sur la base des travaux de Zhang *et al.* [260].

La plateforme a été configurée afin d'explorer l'ensemble des combinaisons existantes entre les algorithmes LGR, RF et SVM ; les descripteurs moléculaires RDKit, CDK, MORDRED, VolSurf+ et les descripteurs structuraux FCFP12, IA25, IA25PFC, IB27, ou

encore IIB24 ; ainsi que 30 découpes du jeu de données en jeu d'apprentissage (75 %) et jeu de test (25 %). Sur les 810 modèles générés pour chaque seuil, seulement 2, 31, 102 et 137 modèles ont été respectivement validés pour les seuils de Fu_p de 2 %, 5 %, 10 % et 20 %. Le coefficient de corrélation de Matthews (MCC) des meilleurs modèles obtenus pour chaque seuil testé est présenté en Figure 56.

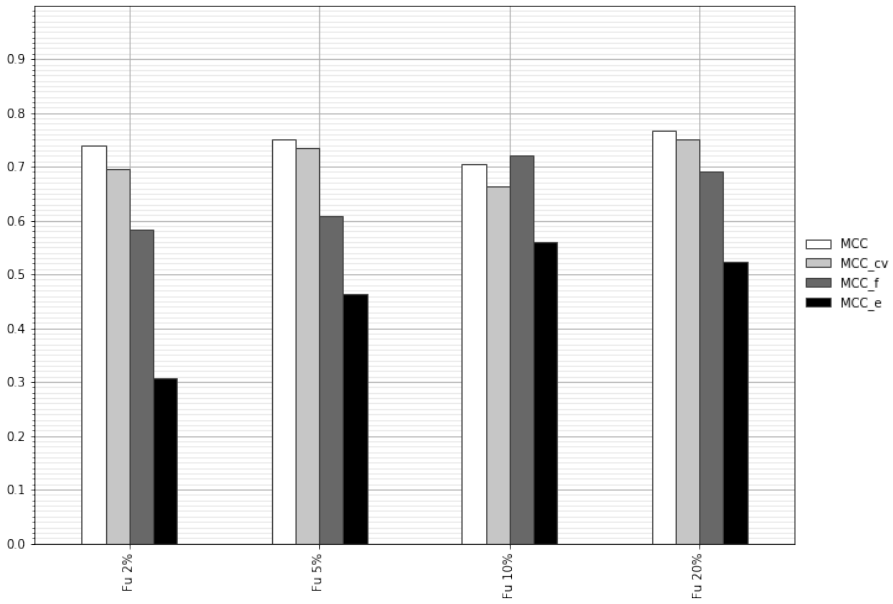


Figure 56 : Recherche du seuil de discrétisation optimum pour l'étude du Fu_p.

Coefficients de Matthews observés pour les seuils de discrétisation testés sur le jeu d'apprentissage (MCC), la validation croisée (MCC_cv), le jeu de test (MCC_f) et le jeu externe (MCC_e).

Comme nous pouvons le voir, les modèles pour les seuils 2 % et 5 % possèdent des performances externes contestables. Ces résultats peuvent être expliqués par le fait que l'utilisation de faibles seuils de Fu_p réduit le nombre de composés observés dans la classe active. Sachant que les classes sont ensuite équilibrées par la plateforme, l'espace chimique couvert par ces modèles se retrouve réduit, avec 58 composés dans le jeu d'apprentissage pour le seuil de 2 % et 96 composés pour le modèle au seuil de 5 %. Par conséquent, il n'est pas surprenant d'observer des bonnes performances internes (MCC et MCC_cv) de ces modèles. Néanmoins, il semble que ces derniers ne permettent pas de prédire les composés du jeu de données externe qui contient 160 molécules, car le MCC_e est inférieur à 0,50 dans les deux cas.

Concernant les modèles aux seuils de 10 % et 20 %, le nombre de composé qui constitue l'espace chimique de ces modèles est de 244 et 296, respectivement. Les jeux d'apprentissage sont donc plus conséquents que ceux observés pour les modèles aux

seuils de 2 % et 5 %, ce qui justifierait les performances externes de ces modèles. Il semblerait également que les performances internes du modèle à 20 % soient meilleures que celles du modèle au seuil de 10 %. On remarque également que le modèle à 10 % dispose de performances externes plus avantageuses que le modèle au seuil de 20 %. D'une part, nous avons remarqué que l'augmentation du seuil au-delà de 10 % desservait les performances externes du modèle. D'autre part, notre objectif initial était de créer un modèle permettant d'identifier les composés qui se fixent fortement aux protéines plasmatiques. Par conséquent, ces résultats nous prouvent que l'utilisation d'un seuil de Fu_p de 10 % permet d'obtenir un modèle de bonnes performances et prédictif pour de nouveaux jeux de données. Le meilleur modèle sélectionné est donc basé sur un seuil de 10 % et a été construit à l'aide d'une régression logistique sur les descripteurs moléculaires RDKit. Le modèle local de classification contient 13 descripteurs dont l'importance est représentée par la Figure 57.

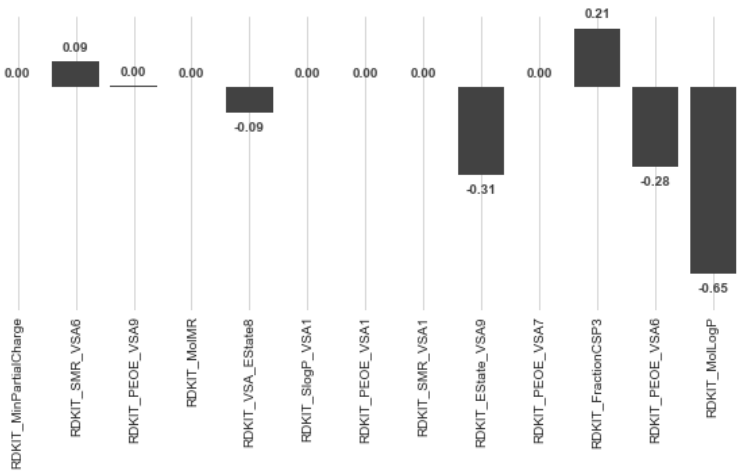


Figure 57 : Coefficient des descripteurs RDKit utilisés par le modèle local de classification.

Suite aux observations faites sur le jeu de données proposé par Obach *et al.*, nous avons souhaité mettre en oeuvre un modèle global de classification à l'aide des données issues de notre base ADMET db.

2.1.4. Modèle global de classification

Le jeu de données sur lequel a été construit le modèle global de classification contenait 2066 molécules uniques. Le jeu de données a été préparé dans les mêmes conditions que précédemment. La plateforme a été configurée selon le même paramétrage. Nous avons essayé de produire des modèles de régression, mais aucun modèle valide n'a été obtenu. Concernant les modèles de classification, 276 modèles ont

été validés sur les 810 combinaisons testées entre les descripteurs, les algorithmes et les découpes du jeu de données. Au final, la plateforme MetaPredict a construit un modèle de consensus basé sur trois modèles uniques dont les caractéristiques sont présentées par la Table 19.

Modèle	Algorithme	Ensemble	N	Descripteurs
M0	LGR	RDKIT	19	MaxAbsPartialCharge, SMR_VSA10, EState_VSA4, SlogP_VSA5, HallKierAlpha, VSA_EState9, EState_VSA8, SlogP_VSA3, MinAbsEStateIndex, EState_VSA10, PEOE_VSA10, SMR_VSA6, SlogP_VSA1, PEOE_VSA9, PEOE_VSA1, FractionCSP3, SlogP_VSA2, PEOE_VSA6, MolLogP
M1	SVM	CDK	14	RNCS, ATSc4, ATSc5, FMF, ATSc3, geomShape, WTPT-2, WTPT-5, MDEO-11, MDEC-12, MDEC-23, FPSA-3, ALogP, XLogP
M2	LGR	MORDRED	98	AATSC8p, ATSC1p, MATS6s, GATS1p, ATSC1i, AATS8m, ATSC2i, SlogP_VSA3, AATSC1p, AATSC4i, AATSC6dv, JGI10, MATS1v, ATSC5i, GATS7v, EState_VSA9, AATSC3d, BCUTdv-1l, Xch-7dv, JGI9, AATS7d, GATS3i, GATS5s, GATS7s, MATS5i, GATS6d, fragCpx, IC2, AATSC5i, GATS2dv, GATS8i, JGI7, SsssN, GATS6s, SsCH3, VR3_A, GATS5dv, AATS6d, SaaaC, MATS3c, ATSC6s, AATS4d, ATSC7i, GATS2d, MID_h, BCUTi-1l, EState_VSA10, SsssCH, RPCG, EState_VSA7, AATSC6s, BalabanJ, JGT10, EState_VSA3, AATS8v, AATS2p, AATS4v, ETA_shape_p, AATS3s, PEOE_VSA11, SMR_VSA1, JGI4, ETA_shape_y, JGI2, EState_VSA8, AATS5s, Kier3, AATS6v, IC0, GATS2i, MID_N, AATS5v, SlogP_VSA1, SdssC, AXp-0d, AATS7p, AATS8p, PEOE_VSA1, AATS3i, AMID_O, AATS4p, AATS4i, ATSC1c, GATS1i, AETA_beta_s, AATS5p, AATSC0v, SlogP_VSA2, SaaCH, MDEC-23, BCUTse-1l, SlogP_VSA6, PEOE_VSA7, MATS1se, ETA_dEpsilon_D, PEOE_VSA6, AATS1i, SLogP

Table 19 : Modèles uniques combinés pour définir le modèle de consensus Fu_p.

Comme dans le cas du modèle local de classification, on remarque que seuls les descripteurs moléculaires permettent d'obtenir des modèles adaptés pour la prédiction de la Fu_p. Selon la Figure 58, nous observons que le consensus possède des performances internes équivalentes aux performances observées pour les modèles uniques.

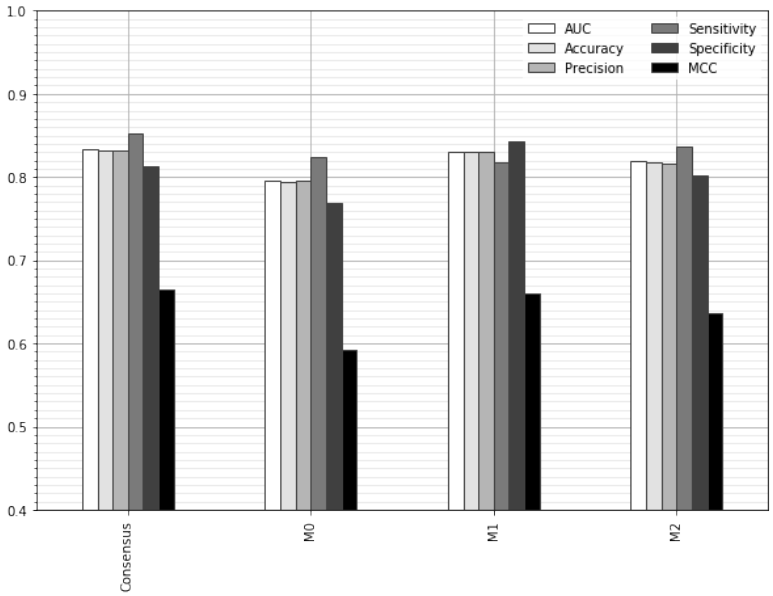


Figure 58 : Performances internes des modèles globaux de Fu_p.

Concernant les performances externes de ces modèles (Figure 59), nous pouvons voir que le consensus fait apparaître les performances moyennes des modèles uniques. Ainsi, contrairement au modèle de consensus créé pour la solubilité aqueuse, le modèle de consensus de la Fu_p dispose d'un pouvoir prédictif supérieur aux modèles uniques M0 et M1 uniquement.

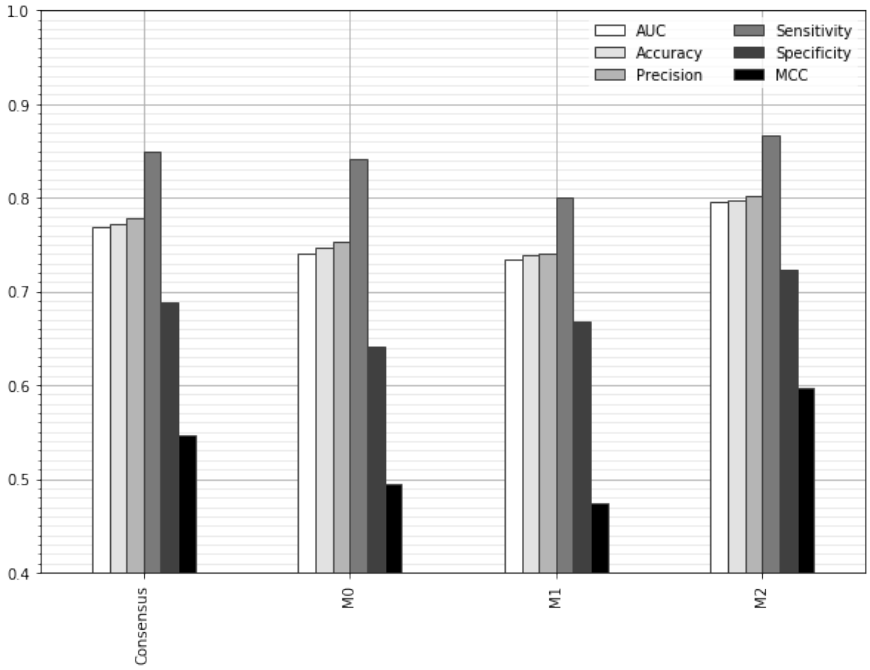


Figure 59 : Performance externes des modèles globaux de Fu_p.

2.1.5. Limitations des modèles développés

Les résultats de ces trois modèles montrent que cette propriété est difficile à modéliser, comme en atteste le nombre de modèles validés par rapport au nombre total de modèles testés. Nous avons également remarqué que seuls les descripteurs moléculaires permettaient d'obtenir des modèles valides. Par conséquent, il semblerait que la fraction libre dans le plasma soit dépendante du comportement moléculaire des composés et non de leurs particularités structurales.

Ceci est en accord avec ce que nous connaissons de la Fu_p. Comme nous l'avons énoncé dans le cadre du chapitre 1, toutes les protéines plasmatiques peuvent simultanément fixer les composés actifs circulant dans le sang. Chaque protéine va être en capacité de fixer des molécules qui possèdent une grande diversité chimique. A titre d'exemple, la forte concentration d'albumine ($C_{Albumine}$ = 30 - 50 mg/mL ≈ 0,5 à 0,7 mM) dans le plasma offre à cette protéine une capacité de liaison très importante. L'albumine est un monomère

de 66,5 kDa qui possède 4 sites de fixation pouvant accueillir des molécules lipophiles, mais également 2 sites de fixation spécifiques pouvant interagir avec des composés hydrophiles ou ioniques [261]. Par conséquent, on en déduit que la fraction libre dans le plasma ne peut pas être facilement expliquée par une relation sous-jacente simple et unique pour tous les composés étudiés. Chaque composé peut être considéré comme un cas particulier pour lequel la fixation aux protéines plasmatiques va dépendre de son comportement moléculaire vis-à-vis d'une protéine plasmatique spécifique et d'un mode d'interaction privilégié.

En somme, les trois modèles que nous venons de décrire ont été choisis pour valider expérimentalement la plateforme MetaPredict.

2.2. Validation expérimentale des modèles

Dans une démarche de validation de nos modèles de prédiction Fu_p, une étude expérimentale a été entreprise afin de constituer une preuve de concept. Cette validation consiste à pouvoir confronter les valeurs théoriques à des valeurs expérimentales obtenues selon un protocole établi pour évaluer la propriété d'intérêt. La démarche de validation que nous avons suivie a tout d'abord comporté une étape de sélection de molécules à tester.

2.2.1. Sélection des molécules testées

La contrainte initiale liée à la sélection des molécules était de constituer un ensemble de validation conséquent, avec une diversité chimique importante, et pour lequel les molécules étaient considérées dans le domaine d'interpolation des trois modèles de prédiction de la fraction libre.

Pour constituer rapidement cet ensemble de validation, nous avons pris le parti de travailler avec des molécules standard et disponibles commercialement via le catalogue AMBINTER, qui contient plus de 7 millions de composés chimiques. Pour identifier les molécules indispensables à notre projet, nous avons effectué une recherche par similarité, à l'aide des empreintes moléculaires FCFP12 de RDKIT, entre les composés du catalogue AMBINTER et les molécules du jeu de données proposé par Obach *et al.*, qui sont communes aux trois modèles. Des groupes de 5 analogues structuraux ont été définis en ne sélectionnant que les composés AMBINTER qui disposaient d'un indice de Tanimoto supérieur à 0,60 avec chaque molécule du jeu de données Obach. Au total, 471 composés ont été identifiés pour constituer l'ensemble du jeu de validation. La Fu_p a été

estimée pour ces composés à l'aide des trois modèles de prédiction présentés précédemment.

Parmi les groupes de molécules identifiées, nous avons supprimé : i) les groupes qui ne contenaient qu'une seule molécule (singletons), ii) les groupes pour lesquels la molécule du jeu de données Obach n'était pas disponible dans le catalogue AMBINTER, iii) toutes les molécules considérées en dehors du domaine d'interpolation des trois modèles. Il est à noter que nous avons voulu sélectionner les composés correspondant aux molécules du jeu de données Obach (ii), dans le but de confronter la valeur publiée à la valeur expérimentale de Fu_p que nous allons déterminer expérimentalement. Ceci nous permettra de juger la pertinence des données publiées pour cette propriété, ce qui n'a jamais été entrepris auparavant. 249 composés ont été choisis pour poursuivre les étapes de sélection. La réduction de l'ensemble de validation a été finalisée par la sélection de tous les composés identifiés dans le domaine d'interpolation des trois modèles. Suite à cela, seuls les groupes qui avaient au minimum 2 analogues structuraux ont été choisis. Au final, l'ensemble de validation contient 68 molécules qui ont été achetées et qui seront testées expérimentalement. La structure de ces molécules est présentée en ANNEXE E.

2.2.2. Protocole expérimental

Les tests expérimentaux sont en cours de réalisation par l'équipe du Pr. Benoit DEPREZ au sein de la plateforme ADME de la faculté de pharmacie de Lille. Le protocole expérimental suivi repose sur une dialyse à équilibrage rapide. Le plasma est tout d'abord dopé avec le composé à tester, qui est incubé à 37°C en triplicat dans un des compartiments de l'insert. L'autre compartiment contient une solution de tampon phosphate à pH 7,2. Après 4 heures d'agitation à 300 rpm, un aliquot de 25 µL de chaque compartiment est prélevé puis dilué. La solution de dilution est adaptée afin d'obtenir une matrice identique pour l'ensemble des compartiments. En parallèle, le retraitement d'un plasma dopé mais non incubé permet d'évaluer le recouvrement de l'étude.

2.2.3. Résultats préliminaires

Les molécules sélectionnées pour constituer l'ensemble de validation sont en cours de test. Par conséquent, nous n'avons pas encore en notre possession les résultats expérimentaux permettant de valider la plateforme. Cependant, l'équipe de Lille utilise 10 molécules de référence afin de suivre le bon déroulement des tests expérimentaux (Figure 60). Nous avons utilisé ces molécules de référence pour effectuer une validation préliminaire de nos modèles.

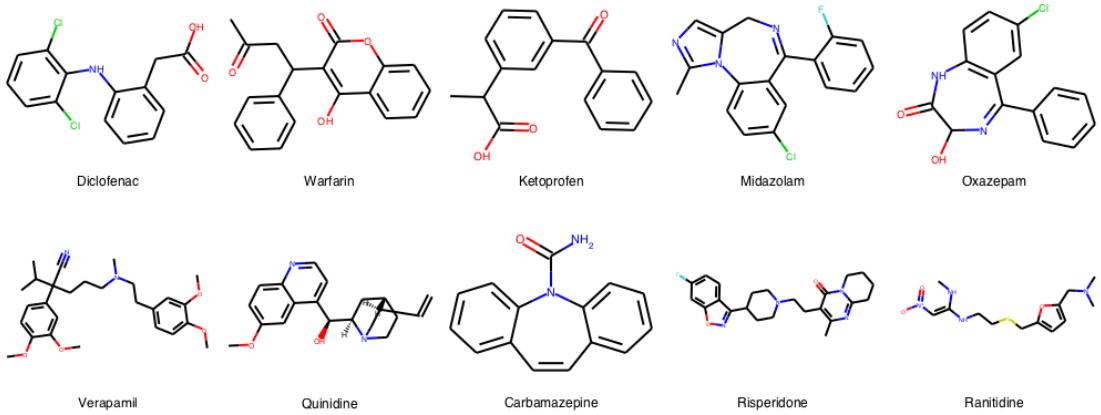


Figure 60 : Molécules de référence pour le test expérimental de la Fu_p.

Les résultats obtenus sur le modèle local de régression (Table 20), nous montrent que 5 molécules sur 10 sont considérées dans le domaine d'interpolation. Une erreur moyenne de 15 % a été estimée sur les prédictions des molécules présentes dans le DA, c'est-à-dire une erreur proche de l'erreur expérimentale observée sur les données Fu_p. On remarque également que le nombre de plus proches voisins ne nous assure pas une prédiction fiable de la propriété modélisée. Cette remarque est justifiée pour la quinidine qui possède 3 voisins dans le DA, mais dont la différence entre les valeurs prédites et expérimentales avoisine les 30 %. D'un point de vue plus général, nous pouvons voir que les prédictions du modèle permettent tout de même de faire une distinction entre les molécules fortement fixées aux protéines plasmatiques (Fu_p < 10 %) et les molécules moyennement ou faiblement fixées (Fu_p > 20 %).

Concernant le modèle local de classification (Table 21), 6 molécules sur 10 sont considérées dans le DA. On remarque que les molécules dans le DA sont toutes bien prédites et que le modèle permet alors d'identifier les molécules fortement fixées. On observe également que pour les molécules identifiées dans le DA, la probabilité d'activité (Pa) est corrélée au nombre de plus proches voisins.

Nom	Fu_p	DA	K	Prédiction
Diclofenac	0,25	IN	1	-19,69
Warfarin	0,83	OUT	0	10,93
Ketoprofen	0,87	OUT	0	7,98
Midazolam	2,3	IN	1	2,63
Oxazepam	5,17	IN	1	-2,33
Verapamil	20,18	OUT	0	19,44
Quinidine	25,66	IN	3	54,17
Carbamazepine	27,5	OUT	0	-9,00
Risperidone	33,35	OUT	0	45,11
Ranitidine	98,73	IN	1	81,04

Table 20 : Prédiction des molécules de référence par le modèle local de régression.

Nom	Fu_p	DA	K	Prédiction	Pa
Diclofenac	0,25	IN	2	Fu < 10	0,85
Warfarin	0,83	OUT	0	Fu < 10	0,76
Ketoprofen	0,87	IN	1	Fu < 10	0,68
Midazolam	2,3	IN	3	Fu < 10	0,8
Oxazepam	5,17	IN	1	Fu < 10	0,76
Verapamil	20,18	OUT	0	Fu < 10	0,68
Quinidine	25,66	IN	1	Fu > 10	0,56
Carbamazepine	27,5	OUT	0	Fu < 10	0,79
Risperidone	33,35	OUT	0	Fu < 10	0,52
Ranitidine	98,73	IN	3	Fu > 10	0,77

Table 21 : Prédiction des molécules de référence par le modèle local de classification.

Nom	Fu_p	DA	Concordance	Prédiction	Pa	Pa_std
Diclofenac	0,25	IN	1	Fu < 10	0,78	0,17
Warfarin	0,83	IN	1	Fu < 10	0,84	0,12
Ketoprofen	0,87	IN	1	Fu < 10	0,76	0,07
Midazolam	2,3	IN	1	Fu < 10	0,77	0,05
Oxazepam	5,17	IN	0,67	Fu < 10	0,7	0,09
Verapamil	20,18	IN	1	Fu < 10	0,61	0,07
Quinidine	25,66	IN	1	Fu > 10	0,55	0,01
Carbamazepine	27,5	IN	1	Fu < 10	0,64	0,2
Risperidone	33,35	IN	1	Fu > 10	0,6	0,07
Ranitidine	98,73	IN	1	Fu > 10	0,9	0,02

Table 22 : Prédiction des molécules de référence par le modèle de consensus global.

Fu_p : valeur expérimentale déterminée par l'équipe de Lille.
DA : domaine d'applicabilité.
K : nombre de plus proches voisins incorporant la molécule étudiée.
Prédiction : estimation de la Fu_p par le modèle.
Pa : probabilité d'activité.
Pa_std : déviation standard de la probabilité d'activité.
Concordance : concordance des modèles uniques dans le consensus.

Pour le modèle global de classification (Table 22), nous pouvons voir que la totalité des molécules testées sont considérées dans le DA. Cependant, contrairement aux deux modèles précédents, on remarque que 2 molécules sont mal prédites (Verapamil et Carbamazepine). Il est important de noter que ces deux molécules ont une faible probabilité d'activité et une déviation standard élevée. Ainsi, on en déduit que les informations du modèle de consensus nous permettent de considérer les prédictions de Fu_p comme étant peu fiables pour ces composés. Par extension, la Quinidine et la Risperidone sont bien prédites, mais elles possèdent également les mêmes caractéristiques. Dans le cas d'une utilisation normale de ce modèle de classification, nous aurions également considéré les estimations de Fu_p suspectes pour ces composés.

En règle générale, ces résultats nous confirment qu'un modèle local est potentiellement plus précis qu'un modèle global. Cependant, les modèles locaux intègrent moins de molécules dans le DA que le modèle global. Ceci peut être expliqué par le fait que les modèles locaux couvrent un sous-espace chimique plus restreint que les modèles globaux. Néanmoins, nous pouvons voir que l'utilisation d'un modèle de consensus peut apporter des informations complémentaires permettant de mieux juger les prédictions et ainsi d'obtenir une fiabilité importante sur les prédictions. On en déduit que l'utilisation du modèle de consensus peut permettre d'apporter des estimations intéressantes de la propriété d'intérêt et ceci pour un plus grand nombre de composés.

2.3. Conclusion de l'étude

La fraction libre est une propriété centrale de notre projet, c'est pourquoi nous avons souhaité la modéliser et valider notre approche à l'aide de tests expérimentaux. Lors de la mise en place des modèles de prédiction, nous avons pu voir que la modélisation de cette propriété ADME-Tox était une tâche difficile. Cependant, comme en atteste la validation préliminaire, les premiers résultats sont encourageants, car le niveau d'adéquation entre les prédictions et les données expérimentales est important.

Néanmoins, la validation sur les 68 molécules actuellement testées par la plateforme ADME de Lille est nécessaire pour établir la preuve de concept que nous avons souhaité mettre en œuvre. Cette validation permettra : i) d'estimer l'adéquation des données publiées par Obach *et al.* avec les données expérimentales, ii) de valider les modèles mis en place pour la prédiction de la Fu_p, iii) de prouver la robustesse de notre approche pour la création du domaine d'applicabilité, iv) de voir si les modèles sont capables de prendre en compte les phénomènes d'*activity cliffs* en fonction des faibles modifications

structurales observées pour les analogues sélectionnés, et v) de voir les différences et confirmer l'intérêt des modèles locaux et des modèles globaux.

3. Valorisation de la plateforme MetaPredict

Aucune interface graphique n'est disponible pour l'utilisation des modèles produits par la plateforme. De ce fait, un modèle est uniquement exploitable en ligne de commande, ce qui n'est pas adapté pour une utilisation en routine au sein de notre laboratoire. De plus, un modèle validé est capable de prédire la propriété modélisée pour plusieurs molécules, mais ces prédictions doivent être analysées avec l'ensemble des informations transmises comme par exemple, les performances du modèle, l'erreur relative sur la prédiction, ou encore l'information sur le domaine d'applicabilité. L'analyse synthétique de ces différentes informations peut représenter un frein pour la compréhension et l'utilisation des prédictions par les chimistes. Dans le but de rendre plus facile l'utilisation des modèles en routine et plus aisée la compréhension des résultats par nos collaborateurs, nous avons mis en place une application en ligne. Suite à de nombreux échanges avec les chimistes de notre laboratoire, nous avons également ajouté à ce service la possibilité de générer des cartes d'activité. Ces cartes d'activité ont pour but de représenter la contribution de sous-structures vis-à-vis des propriétés modélisées. Elles peuvent être modulées en fonction des besoins des chimistes et des caractéristiques d'un projet. L'application en ligne que nous présentons ci-après est en cours de déploiement au sein de notre laboratoire.

3.1. Serveur et application en ligne

Le serveur a pour but de centraliser l'ensemble des modèles produits au sein de notre laboratoire pour les rendre accessibles à nos collaborateurs. Ce serveur a été implémenté à l'aide de PHP version 7.1.8 (http://php.net). Nous voulions que l'application en ligne supportée par le serveur soit facile à utiliser et intuitive. Pour cela, nous avons choisi de développer un site internet basé sur le concept du « one page », c'est-à-dire que le site est intégralement contenu dans une seule page *html*. La navigation sur le site est alors effectuée de manière verticale et par défilement. Cette stratégie présente plusieurs avantages : i) ceci nous permet de réduire le temps nécessaire à la mise en place de l'application grâce à une architecture minimaliste, et ii) ceci est plus ergonomique et apporte à l'utilisateur une vue d'ensemble des fonctionnalités de l'outil sans nécessiter l'ouverture de multiples fenêtres. Nous avons mis en place cette application en ligne en utilisant HTML5 et CSS (https://www.w3.org/html) pour le corps du site internet et l'identité

visuelle. Au total, 4529 lignes de codes ont été nécessaires pour mettre en place cette application en ligne.

Le site internet que nous avons créé comporte plusieurs parties. La première partie correspond à la présentation de l'outil et des fonctionnalités qui peuvent être utilisées à l'aide de l'application (Figure 61). Cette brève présentation permet de donner à l'utilisateur une vue d'ensemble du service proposé.

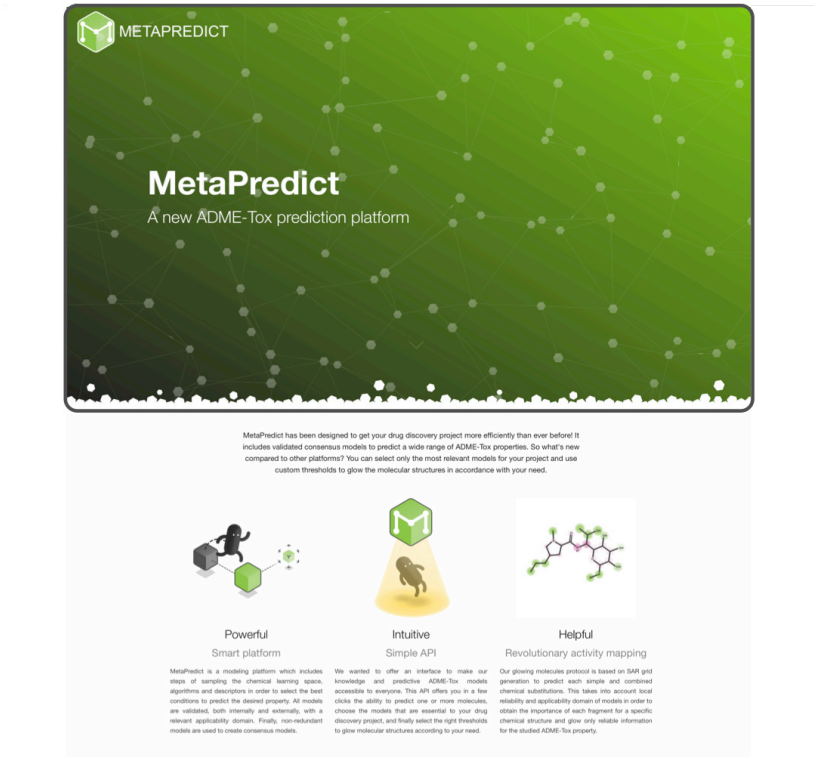


Figure 61 : Page d'accueil du site internet.

Le cadre noir représente la partie visible sur l'écran de l'utilisateur.

La deuxième partie comporte un formulaire permettant à l'utilisateur de paramétrer sa requête pour la prédiction des molécules qu'il souhaite soumettre. Ce formulaire est constitué de trois volets. Le premier volet donne les instructions à suivre dans le cas d'une première utilisation (Figure 62.A). Le deuxième volet est dédié au chargement des structures moléculaires à traiter (Figure 62.B). Le troisième volet comporte la sélection des modèles que l'utilisateur souhaite employer pour l'estimation de propriétés ADME-Tox spécifiques (Figure 62.C).

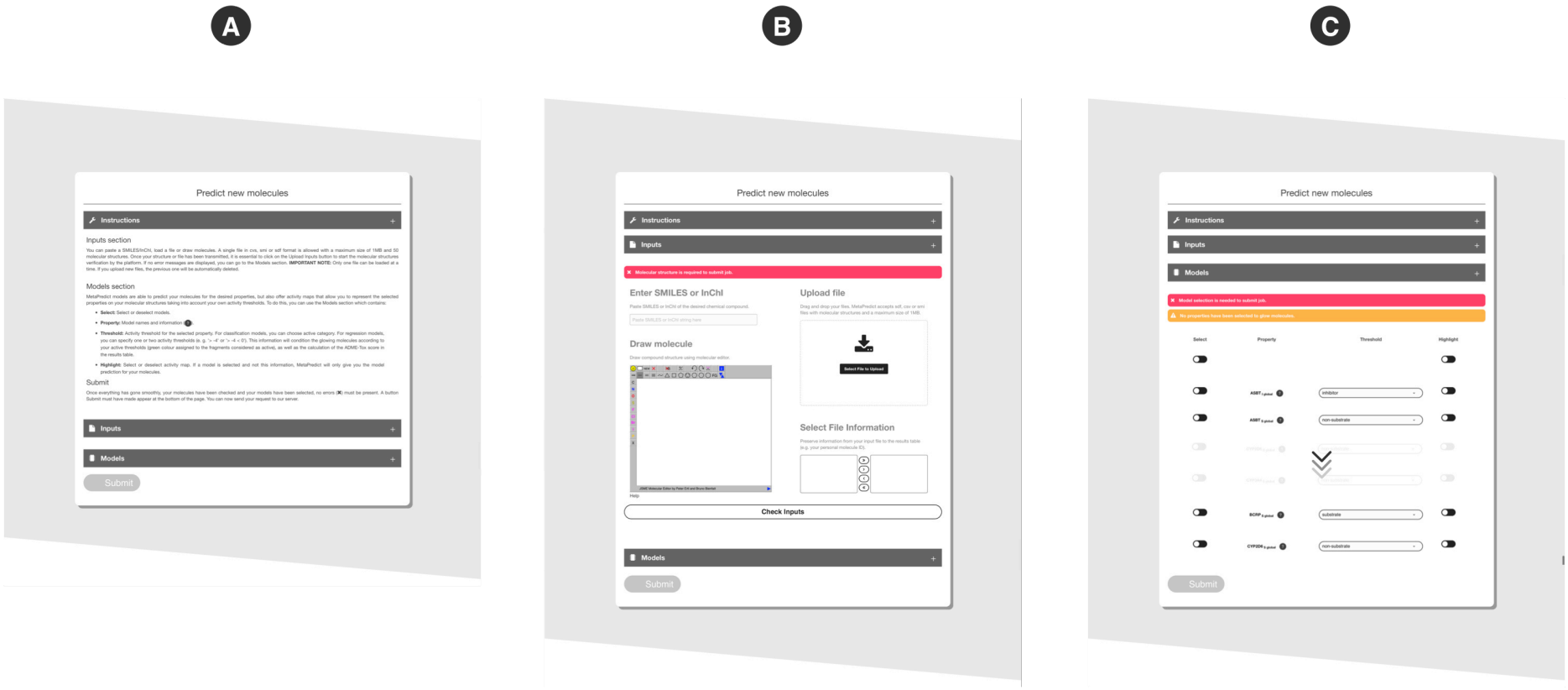


Figure 62 : Formulaire du site internet.

A) Volet qui présente les instructions générales à prendre en considération lors d'une première utilisation du service en ligne. B) Volet qui permet de charger les structures moléculaires à prédire. C) Volet qui répertorie l'ensemble des modèles ADME-Tox à disposition pour la prédiction de propriétés spécifiques.

Le fonctionnement général de ce formulaire est simple et intuitif. La requête ne peut être soumise que lorsque tous les paramètres d'entrées ont été vérifiés par l'application en ligne. Pour cela, un ensemble de procédures a été définit à l'aide de JavaScript (https://developer.mozilla.org/fr/docs/Web/JavaScript) et JQuery (http://jquery.com) pour l'interactivité du site internet et le contrôle des paramètres sélectionnés par l'utilisateur. Ainsi, comme nous pouvons le voir sur la Figure 62, des messages d'erreur (encadrés rouges) et des avertissements (encadrés jaunes) vont informer en temps réel l'utilisateur des changements à apporter au formulaire pour pouvoir soumettre sa requête.

Le volet de chargement des structures moléculaires comporte différentes parties. Il donne l'opportunité de transmettre une structure 1D (SMILES ou InChI), de dessiner une structure chimique à l'aide de l'éditeur de molécule JSME (http://www.peter-ertl.com/jsme), de charger directement un fichier au format *csv*, *sdf* ou *smi*, et pour finir de préserver des informations présentent dans le fichier initial dans le cas d'un fichier téléversé. Concernant ce dernier point, la taille maximale de fichier acceptée par la plateforme est de 1 MB. De plus, l'application va vérifier si des structures moléculaires sont présentent dans le fichier partagé par l'utilisateur. Si l'une de ces conditions n'est pas respectée, le fichier n'est pas téléchargé par le serveur et un message d'erreur correspondant à la nature du problème est affiché. De plus, le nombre de molécules pouvant être exploitées est limité à 50 par requête. Cette restriction assure le bon fonctionnement de l'outil et prévient de la saturation du serveur par un utilisateur. Les informations structurales sont ensuite standardisées afin de vérifier que les structures chimiques à exploiter sont valides. Une fois que toutes les conditions liées aux molécules à traiter sont respectées, il est possible de sélectionner les modèles à utiliser.

Le volet dédié à la sélection des modèles comporte les champs « sélection », « propriété », « seuil » et « mise en lumière ». Le champ « sélection » permet de choisir individuellement les modèles à utiliser. Le champ « propriété » apporte une description du modèle, de la propriété qu'il modélise et de ses performances. Le champ « seuil » donne l'opportunité à l'utilisateur de sélectionner la gamme d'activité qu'il considère active en fonction du projet sur lequel il travaille. Ces seuils seront utilisés afin de colorer les représentations moléculaires en fonction de cette information. Le champ « mise en lumière » permet de choisir les modèles pour lesquelles nous souhaitons représenter la propriété ADME-Tox sur la structure moléculaire selon le principe des cartes d'activité présentées ci-après.

Une fois l'ensemble des paramètres vérifié par l'application en ligne et que plus aucun message d'erreur n'est observé, le bouton de soumission est activé et change de couleur en passant du gris (inactivé) au vert (activé). Il est à noter qu'une requête peut être soumise lorsque des avertissements sont observés.

3.2. Cartes d'activité

Les modèles (Q)SAR sont régulièrement utilisés pour prédire une grande variété de propriétés facilitant les prises de décision pour prioriser un projet de synthèse. Cependant, une critique couramment faite à l'égard des modèles prédictifs est qu'ils offrent peu d'indices sur les raisons pour lesquelles une molécule possède une propriété particulière. Cette particularité est d'autant plus vraie que les méthodes d'apprentissage modernes ou l'utilisation de descripteurs peu explicatifs mais prédictifs favorisent la construction de modèle « boite noire ». Pour faciliter la compréhension de ces modèles, plusieurs groupes de recherches ont proposés des alternatives afin de visualiser de manière intuitive les relations existantes entre des sous-structures chimiques et leurs activités biologiques. Les objectifs de ces approches sont i) d'identifier clairement les régions problématiques d'une molécule qui la défavorise vis-à-vis de la propriété modélisée, ii) de mettre en évidence des groupes fonctionnels qui tendent à améliorer la propriété modélisée, et iii) de comprendre la relation multidimensionnelle structure-activité (RSA) d'une série chimique. Pour une molécule d'intérêt, l'information est généralement représentée sous la forme d'un schéma moléculaire 2D sur lequel des sous-structures vont être colorées en fonction de leur l'impact sur la propriété étudiée.

Pour répondre à ces objectifs, plusieurs approches ont été proposées [262–264]. Elles reposent toutes sur le même principe selon lequel, la coloration du schéma moléculaire est obtenue en attribuant un score à chaque atome. Pour définir ce score, un modèle prédictif est tout d'abord créé sur la base d'empreintes moléculaires. Le score d'un atome est ensuite déterminé en fonction de la contribution des fragments qui le contiennent, et qui encodent la sous-structure correspondante dans la molécule étudiée. La contribution d'un fragment peut être définie comme son importance dans l'équation du modèle. Le vecteur qui contient les scores des atomes est ensuite normalisé entre 0 et 1 [264]. Ce vecteur normalisé est transformé en une couleur sur la représentation moléculaire 2D de la molécule étudiée, allant généralement du rouge pour des scores proches de 0 ; en passant par du jaune pour des scores proches de 0,5 ; jusqu'à une couleur verte pour les scores proches de 1.

Nous avons souhaité mettre en œuvre une approche similaire dans le cadre de l'application en ligne de la plateforme MetaPredict. Cependant, nous avons considéré que les approches basées sur ce principe présentaient plusieurs inconvénients comme : i) la méthodologie suivie ne peut pas être appliquée à des modèles basés sur des descripteurs moléculaires. ; ii) les fragments structuraux sont obtenus en fonction d'un schéma de fragmentation spécifique, qui ne prend pas en compte les particularités structurales de la molécule. Par conséquent, des collisions dans les empreintes moléculaires peuvent être observées, se traduisant par des coefficients erronés dans le modèle, et donc dans le score attribué à chaque atome. ; iii) un atome peut être représenté par plusieurs fragments structuraux, ce qui peut fournir un score erroné lorsqu'il est surreprésenté. ; iv) un score atomique est attribué sur la base de descripteurs structuraux, ce qui implique qu'un atome peut hériter d'une contribution dont il n'est pas responsable. ; v) les fragments étudiés ne sont pas position dépendant et ne permettent pas de prendre en considération l'influence de l'environnement local d'un chémotype. Par conséquent, les scores attribués aux atomes d'un chémotype spécifique seront *a priori* identiques pour toutes les structures qui le contiennent. ; vi) l'information représentée sur la structure est donc globale et ne représente que les tendances sous-jacentes observées à partir des descripteurs du modèle. De ce fait, seule l'importance statistique des fragments est représentée sur la structure, ce qui peut fausser l'interprétation des résultats pour des séries congénériques. ; vii) Ceci pose la question de la pertinence de l'information transmise aux chimistes. Toutes les molécules exploitées ne sont probablement pas contenues dans le domaine d'applicabilité du modèle. Par conséquent, l'information représentée sur leurs structures n'est peut-être pas représentative de ce que le modèle est capable de prédire (extrapolation). ; viii) La coloration ne traduit pas directement la variation de la propriété modélisée.

3.2.1. Stratégie développée pour la mise en place des cartes d'activité

Nous avons choisi de développer une nouvelle approche de visualisation de la propriété sur la structure moléculaire. L'algorithme développé permet d'extraire les relations entre la prédiction du modèle et les particularités structurales du composé testé. Cet algorithme fonctionne en intégrant le modèle dont il extrait les prédictions. Il fournit ensuite une interprétation du modèle afin de représenter les résultats sur les atomes et les liaisons de la structure du composé. De ce fait, cette méthode ne constitue pas une interprétation mécanistique du modèle telle qu'elle peut être obtenue par les approches

présentées précédemment. Cependant, elle permet d'identifier l'importance des sous-structures du composé vis-à-vis de la propriété modélisée.

Notre méthode repose sur un principe simple à partir duquel l'interprétation de la prédiction prend la forme suivante : « Le modèle prédit une propriété de {x} pour l'entité structurale {s} avec une confiance de {y}. Une entité structurale est une sous-structure du composé initial, qui peut contenir un ou plusieurs fragments {z}. Par conséquent, la prédiction {x} est induite par la présence du/des fragment(s) {z} dans l'entité structurale {s} ». Le principe peut être assimilé à une approche RSA couramment utilisée par les chimistes médicinaux. Ainsi, il est nécessaire d'interroger le modèle pour chaque fragment individuel, mais également leurs combinaisons pour élucider les relations existantes entre des sous-structures. Les prédictions sont ensuite employées pour définir des scores. Ces derniers permettent de moduler la coloration des atomes afin de créer notre carte d'activité.

La composante fondamentale de notre approche est l'entité structurale ({s}) qui est soumise au modèle pour qu'une prédiction soit retournée. Plusieurs entités vont être définies afin de former un réseau de sous-ensembles structuraux représentant le schéma de fragmentation du composé. Pour la génération du schéma de fragmentation d'un composé (Figure 63), l'algorithme identifie dans un premier temps le *scaffold* de Bemis-Murcko [265], qui peut être assimilé au squelette moléculaire défini selon les cycles de la structure chimique et les liens qui permettent de les relier entre eux. La comparaison de ce *scaffold* à la structure initiale permet d'identifier l'ensemble des fragments {z}. Les 2^n combinaisons des fragments {z} vont être énumérées afin de générer les entités structurales ({s}). L'ensemble des entités structurales ({s}) forme ce que nous avons appelé une grille RSA. Dans le cadre de la plateforme MetaPredict, cette grille peut être complétée par la fragmentation du *scaffold* de Bemis-Murcko à l'aide de règles de rétrosynthèse BRICS [266] et RECAP [267] disponibles via RDKit. L'ensemble de ces structures vont être soumises au modèle, afin d'en extraire une valeur estimée de la propriété d'intérêt dans le but de définir les scores de coloration.

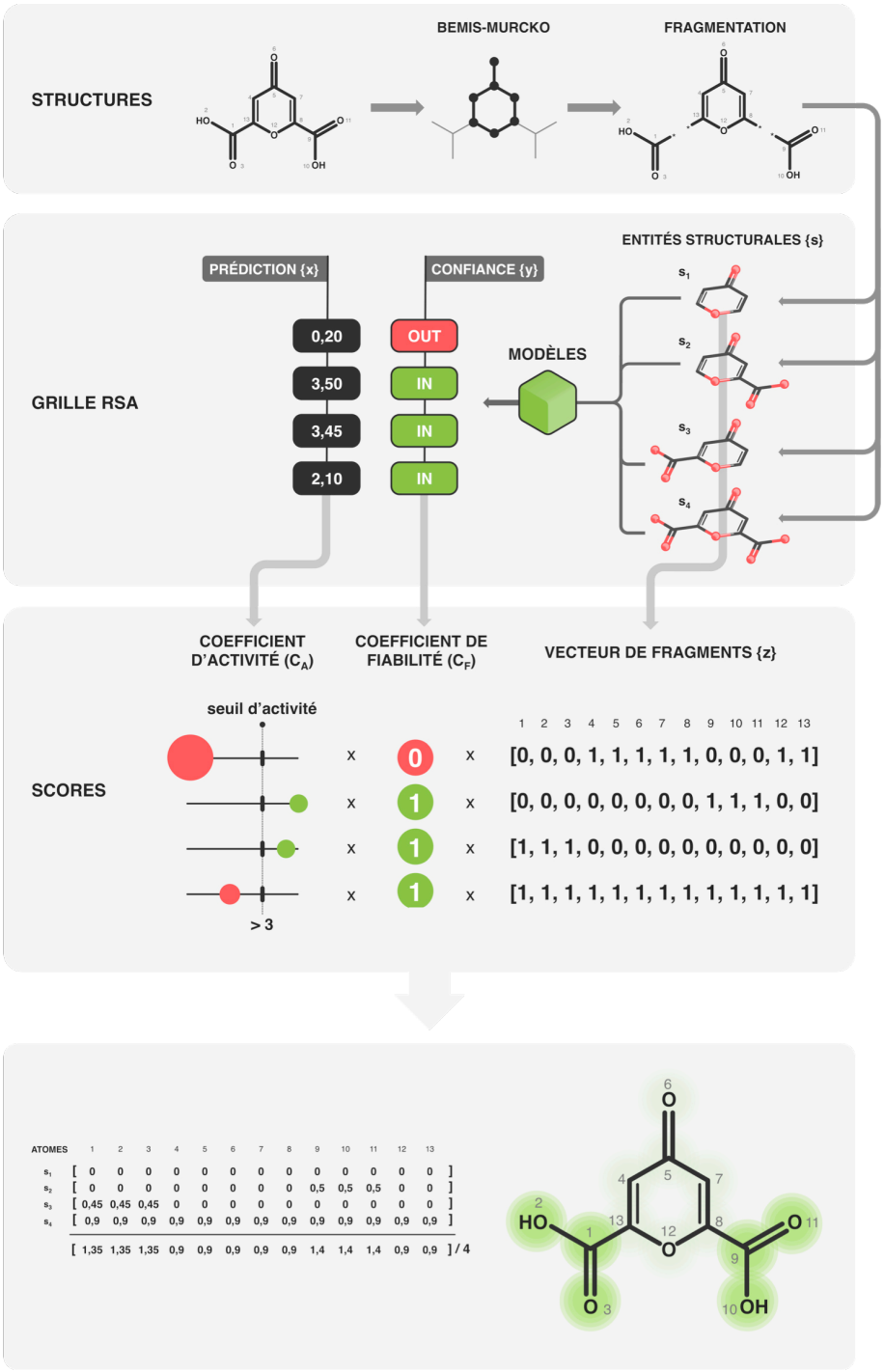


Figure 63 : Processus de génération des cartes d'activité.

3.2.2. Calcul des scores et création des cartes d'activité

L'étape suivante a pour but de définir les scores attribués à chaque atome permettant de colorer la structure du composé. Ces scores sont déterminés en fonction de l'influence de chaque fragment ({z}) sur la propriété modélisée (Figure 63). Pour chaque entité structurale ({s}) on commence par définir un vecteur nul et de longueur

égale au nombre d'atomes présents dans la structure du composé. Pour chaque vecteur des entités structurales ({s}), une valeur de 1 va être attribuée aux positions des atomes observées uniquement dans les fragments ({z}). Les vecteurs sont ensuite pondérés en fonction de coefficients qui reflètent l'activité désirée par l'utilisateur (C_A) et la fiabilité du modèle sur la prédiction (C_F) afin d'établir les scores nécessaires pour créer la carte d'activité.

3.2.2.1. Attribution du coefficient d'activité (C_A)

L'objectif du coefficient d'activité (C_A) est d'adapter la coloration de la structure en fonction de l'activité recherchée par l'utilisateur. Le but est de donner la possibilité de choisir la classe (propriété discrète) ou la gamme de valeurs (propriété continue) à représenter comme active sur la structure du composé. Ainsi, un score négatif traduit l'inactivité de la structure étudiée (couleur rouge), tandis qu'un score positif traduit l'activité de la structure étudiée (couleur verte). Un score égal à zéro signifie qu'aucune information ne peut être représentée sur la structure moléculaire (aucune couleur).

Pour une propriété discrète, l'utilisateur peut définir quelle classe est considérée comme active, afin de la colorer en vert. Si l'entité structurale ({s}) est prédite active, alors son vecteur est multiplié par le coefficient d'activité de valeur +1, sinon un coefficient d'activité de valeur -1 est appliqué. Dans le cas d'une propriété continue, le coefficient d'activité reflète l'écart de la valeur prédite (Y_{pred}) avec un ou deux seuils d'activité (Y_{seuil}). Les seuils qui peuvent être appliqués sont définis à l'aide d'un signe mathématique (supérieur ou inférieur) et d'une valeur numérique, comme par exemple « > -2 » pour un seuil unique ou « > -2 » et « < 0 » lorsque deux seuils sont choisis. Il est possible que des seuils incohérents soient rencontrés, comme par exemple « > -2 » et « > 0 ». Lorsque ce cas de figure est observé, l'algorithme prend en compte le seuil de plus faible valeur pour proposer des scores définis en fonction d'un seuil unique (« > -2 »). Dans le cas d'un seuil unique d'activité, le score de chaque atome est ajusté en transformant la valeur prédite de la propriété à l'aide de l'Equation 22.

$$C_A = k * (Y_{pred} - Y_{seuil})$$

Equation 22 : Coefficient d'activité pour une propriété continue dans le cas d'un seuil d'activité unique.

Avec C_A le coefficient d'activité ; k la constante attribuée en fonction du seuil d'activité ; Y_{pred} la valeur prédite de la propriété pour l'entité structurale ({s}) ; Y_{seuil} la valeur numérique du seuil choisi.

Cette équation traduit la relation linéaire existante entre le coefficient d'activité (C_A) et l'écart ($Y_{pred} - Y_{seuil}$). Lorsque nous souhaitons que la couleur verte indique des valeurs supérieures au seuil d'activité, la constante k est égale à +1. Inversement, lorsque nous souhaitons que la couleur verte indique des valeurs inférieures au seuil d'activité, la constante k est égale à -1. Ceci peut être représenté schématiquement par la Figure 64.

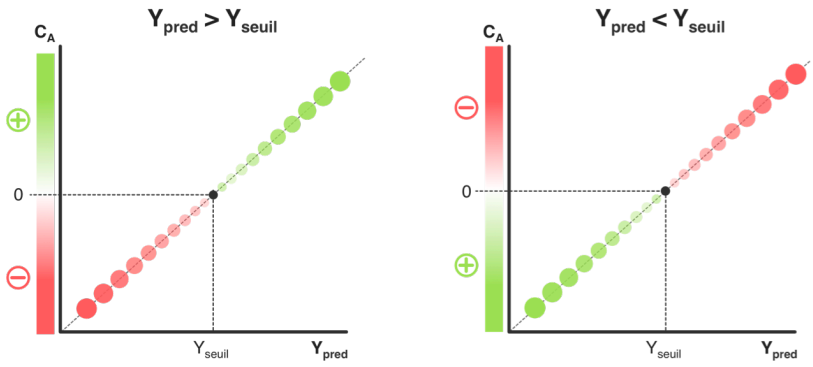


Figure 64 : Représentation schématique du coefficient d'activité dans le cas d'un seuil unique.

La représentation de gauche correspond au coefficient C_A attribué lorsque l'utilisateur considère que les prédictions supérieures au seuil sont actives. La représentation de droite correspond au coefficient C_A attribué lorsque l'utilisateur considère que les prédictions inférieures au seuil sont actives.

L'Equation 22 ne peut pas être appliquée lorsque plusieurs seuils d'activité sont utilisés. Pour répondre à cette problématique, nous avons considéré une fonction parabolique pour définir le coefficient d'activité. Comme représenté par la Figure 65, deux cas peuvent être envisagés, à savoir : i) Y_{pred} appartient à l'intervalle compris entre Y_{seuil_1} et Y_{seuil_2}, et ii) Y_{pred} n'appartient pas à l'intervalle compris entre Y_{seuil_1} et Y_{seuil_2}.

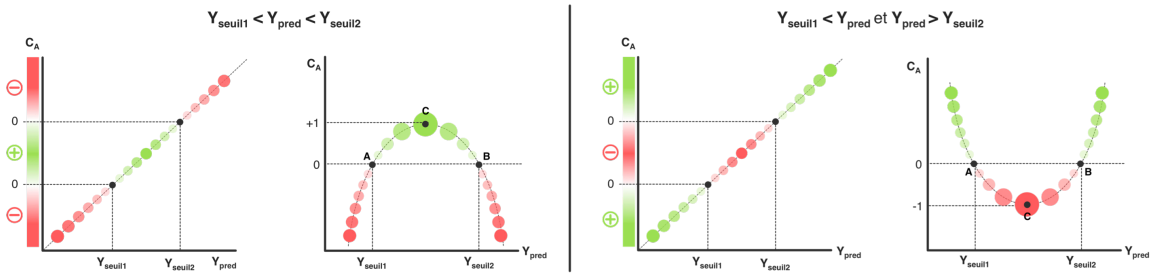


Figure 65 : Représentation schématique du coefficient d'activité dans le cas de deux seuils.

La représentation de gauche est observée lorsque les seuils d'activité impliquent que $Y_{pred} \in [Y_{seuil1} ; Y_{seuil2}]$ et la représentation de droite lorsque les seuils d'activité impliquent que $Y_{pred} \notin [Y_{seuil1} ; Y_{seuil2}]$. Pour chaque cas de figure, la représentation de la fonction parabolique utilisée pour définir les coefficients d'activité est présentée. Les points verts correspondent aux valeurs actives et les points rouges aux valeurs inactives selon les seuils d'activité choisis.

L'équation de cette parabole permet de définir le coefficient d'activité C_A (noté y par la suite) en fonction de la valeur prédite de la propriété Y_{pred} (noté x par la suite). Son équation, du type $ax^2 + bx + c = y$, comporte trois inconnues. Afin de résoudre cette équation à trois inconnues, nous avons considéré trois points A, B et C comme représenté par la Figure 65. Les points A et B sont décris par les seuils d'activité Y_{seuil_1} et Y_{seuil_2}, et ils ont pour coordonnées respectives $(x_1, y_1 = 0)$ et $(x_2, y_2 = 0)$. Le point C permet de définir le sommet de la parabole et possède les coordonnées (x_3, y_3). x_3 est déterminé à l'aide de la moyenne des deux seuils d'activité Y_{seuil_1} et Y_{seuil_2}. Lorsque nous souhaitons que la couleur verte indique des prédictions comprises dans l'intervalle décrit par les seuils d'activité, y_3 est égal à $+1$. Inversement, lorsque la couleur verte doit indiquer des prédictions en dehors de l'intervalle décrit par les seuils d'activité, y_3 est égal à -1. Les coordonnées de ces trois points permettent de résoudre l'équation de la parabole (ANNEXE E) et de déterminer les coefficient a, b et c selon l'Equation 23. L'équation de la parabole permet ensuite d'identifier le coefficient d'activité C_A en fonction des valeurs prédites de la propriété et ceci pour toutes les entités structurales ({s}) de la grille RSA.

$$a = \frac{y_3}{(x_3^2 - x_1^2) - \left(\left(\frac{x_2^2 - x_1^2}{x_2 - x_1}\right) \times (x_3 - x_1)\right)} \qquad b = -a \times \left(\frac{x_2^2 - x_1^2}{x_2 - x_1}\right) \qquad c$$
$$= -ax_1^2 - bx_1$$

Equation 23 : Calcul des coefficients a, b et c de la parabole selon les coordonnées des points A, B et C.

3.2.2.2. Application du coefficient de fiabilité (C_F)

L'objectif du coefficient de fiabilité (C_F) est de pénaliser les entités structurales ({s}) pour lesquelles le modèle est peu fiable. Ce coefficient prend en considération l'information sur le domaine d'applicabilité. Lorsqu'une entité structurale ({s}) est identifiée en dehors du domaine d'interpolation, son vecteur est annulé en multipliant par zéro l'ensemble des scores qu'il contient. Ainsi, seule l'information contenue dans le domaine d'interpolation sera représentée sur la carte d'activité. Le coefficient de fiabilité prend également en compte la notion de confiance ({y}) sur la prédiction. Pour le moment seul les modèles de classification bénéficient de la prise en compte de la confiance ({y}) dans l'énumération du coefficient de fiabilité. Pour cela, le vecteur de chaque entité structurale est multiplié par la probabilité d'activité transmise par les modèles de classification. A

l'heure actuelle, aucun équivalent n'a été trouvé dans le cas des modèles de régression. Ceci représente un des points d'amélioration de la stratégie développée.

3.2.2.3. Détermination des scores

Une fois les coefficients d'activité (C_A) et de fiabilité (C_F) déterminés pour toutes les entités structurales ({s}), les vecteurs sont combinés afin de définir le score de contribution des fragments ({z}) vis-à-vis de la propriété modélisée. Pour cela, les vecteurs qui contiennent les scores propres à chaque entité structurale sont sommés à l'aide de l'Equation 24.

$$S = \frac{1}{n_{\{s\}}} \sum_{1}^{\{s\}} v_s \times C_{A_{\{s\}}} \times C_{F_{\{s\}}}$$

Equation 24 : Détermination des scores d'activité

Avec S le vecteur numérique contenant les scores des atomes ; $n_{\{s\}}$ le nombre d'entité structurales {s} présentes dans la grille RSA ; v_s le vecteur initial contenant la position des atomes dans la structure du composé ; $C_{A_{\{s\}}}$ le coefficient d'activité de l'entité structurale {s} ; $C_{F_{\{s\}}}$ le coefficient de fiabilité de l'entité structurale {s}.

Le vecteur S permet ensuite de générer la carte d'activité à l'aide de la fonction *GetSimilarityMapFromWeights* de RDKit, et ceci sans aucune normalisation des scores [268]. La normalisation est généralement employée pour obtenir des scores qui ont une échelle identique (entre 0 et 1) afin de visualiser une variation de couleur similaire sur l'ensemble des représentations moléculaires d'un jeu de données. Cependant, comme énoncé par Rosenbaum *et al.* [269], cette normalisation peut fournir des scores érronés qui en fonction de la méthode utilisée peut attribuer une couleur à des atomes qui initialement ne disposaient d'aucune information. Dans notre cas, cette normalisation n'est pas nécessaire puisque les scores déterminés pour chaque entité structurale ({s}) sont soumis aux mêmes seuils d'activité. De ce fait, les scores de plusieurs composés sont établis selon les mêmes normes.

Ces cartes d'activité sont intégrées dans les fonctionnalités de l'application en ligne. Elles permettent de visualiser les propriétés ADME-Tox sur les structures moléculaires pour l'ensemble des modèles disponibles sur le site internet. La stratégie mise en place est donc universelle et permet d'interpréter des modèles basés sur des descripteurs structuraux mais également moléculaires.

3.3. Présentation des résultats transmis par l'application en ligne

La visualisation des données est une partie indispensable de l'étude scientifique. Elle permet de représenter graphiquement les données pour aider l'utilisateur à mieux comprendre les processus sous-jacents décrits par les données. Dans le contexte de la découverte moderne de médicaments, les données visualisées peuvent fournir une vue simplifiée des phénomènes complexes étudiés [270]. Notre objectif lors du développement de l'application en ligne était de fournir à l'utilisateur une information simple et utile, lui apportant une visualisation intuitive des résultats générés par les modèles de prédiction. Les résultats transmis par l'application en ligne sont partagés via une page *html* unique pouvant être consultée par défilement. La forme de cette page peut différer en fonction du nombre de molécules traitées par l'application (Figure 66).

Figure 66 : Pages de résultats de l'application en ligne.

A) Page de résultats lorsqu'une seule molécule est traitée par l'application. Les estimations de chaque modèle sont présentées successivement. B) Page de résultats lorsque plusieurs molécules sont traitées par l'application. Les résultats sont présentés dans une table contenant deux colonnes par modèle. La première colonne permet de partager les informations liées à l'estimation du modèle. La deuxième colonne permet de représenter les cartes d'activité rattachées à l'estimation d'un modèle.

Bien que la structure de ces deux pages de résultats est différente, les informations qu'elles transmettent sont identiques. Les résultats contiennent l'ensemble des informations relatives aux prédictions ainsi que les cartes d'activité générées pour les modèles sélectionnés. Les informations sur les prédictions permettent de partager la valeur prédite de la propriété ADME-Tox modélisée, complétée par la probabilité d'activité dans le cas d'un modèle de classification, ainsi que l'information sur le domaine d'applicabilité (dans ou en dehors du domaine d'interpolation). Lorsqu'un modèle de consensus est utilisé pour la prédiction d'une propriété ADME-Tox, les informations précédemment énoncées sont respectivement complétées par l'erreur standard sur la prédiction ou sur la probabilité d'activité, ainsi que la concordance des modèles sur le domaine d'applicabilité et le nombre de modèles qui intègrent le composé testé dans leur domaine d'interpolation.

Les informations sont colorées selon le code couleur gris, vert et rouge (Figure 66). Cette coloration dépend des seuils d'activité appliqués par l'utilisateur pour chaque propriété modélisée. La couleur grise est attribuée à tous les composés prédis en dehors du domaine d'applicabilité du modèle qui a estimé la propriété ADME-Tox renseignée. La couleur verte signifie que le composé est considéré dans le domaine d'interpolation du modèle et qu'il possède une estimation de la propriété en accord avec les attentes (seuils d'activité) de l'utilisateur. Inversement, une couleur rouge indique que le composé est dans le domaine d'interpolation du modèle, mais qu'il dispose d'une estimation de la propriété en désaccord avec les attentes de l'utilisateur. Les cartes d'activité permettent de comprendre les raisons pour lesquelles la molécule ne possède pas une estimation en accord avec les seuils d'activité désirés. Ces cartes peuvent être agrandies comme illustré par la 67 et sont renseignées par les seuils utilisés pour colorer la structure moléculaire.

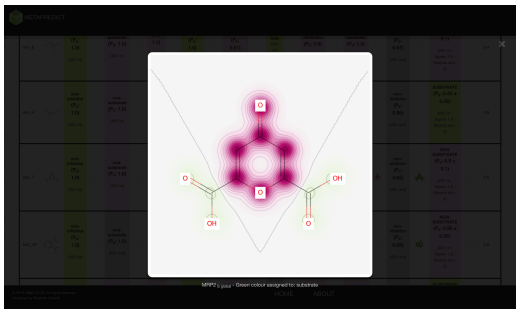


Figure 67 : Agrandissement des cartes d'activité.

L'information transmise au bas de l'image donne explicitement les seuils utilisés pour colorer la structure du composé.

Nous avons également défini un score ADME-Tox qui permet de combiner les prédictions sur plusieurs propriétés pour un composé. Ce score dépend des seuils d'activité appliqués par l'utilisateur. Il traduit le pourcentage de propriété ADME-Tox pour lesquelles un composé possède des estimations en accord avec les seuils d'activité. Ceci permet alors de trier les composés testés et d'identifier rapidement celui qui possède un profile ADME-Tox favorable et en accord avec les contraintes de l'utilisateur.

Comme présenté par la Figure 66, les résultats peuvent être téléchargés aux formats *html* et *xlsx*. Le fichier *html* correspond à la page de résultats proposée par l'application en ligne avec l'ensemble des options qu'elle contient. Ainsi, il est possible d'utiliser hors ligne l'ensemble des fonctionnalités de la page *html* comme par exemple l'agrandissement des cartes d'activité, la possibilité d'ordonner toutes les colonnes de la table ou encore de changer l'ordre des colonnes (Figure 66.B). Le fichier *xlsx* propose à l'utilisateur un rapport détaillé des modèles utilisés et des prédictions obtenues. Ce rapport est généré automatiquement par l'application et il est formé de trois pages (Figure 68). La première page propose une description du fichier et des termes utilisés pour présenter les résultats. La deuxième page référence les modèles utilisés par l'utilisateur pour prédire ses molécules, ainsi que les seuils d'activité qu'il a choisi pour chaque propriété ADME-Tox. La troisième page présente les résultats avec l'ensemble des informations issues des modèles de prédiction, le score ADME-Tox et les scores utilisés pour générer les cartes d'activité de chaque propriété. Ce tableau est également coloré selon les normes présentées précédemment.

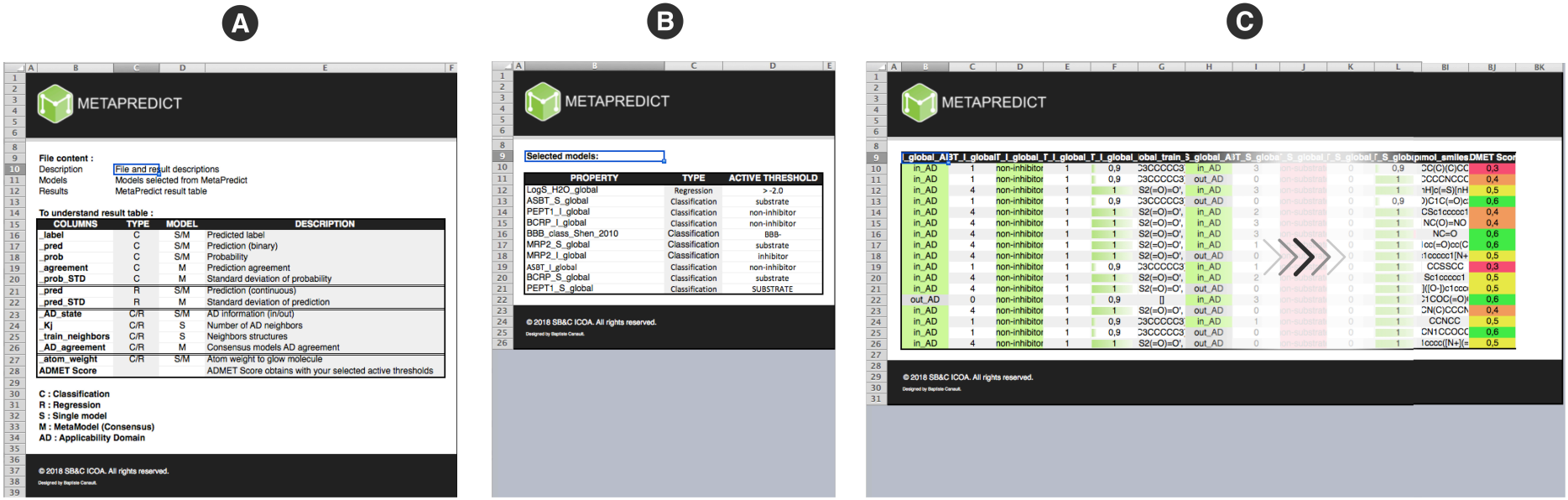


Figure 68 : Aperçu du fichier *xlsx* transmis par l'application en ligne.

A) Page de présentation du rapport et de description des termes utilisés pour exprimer les résultats. B) Page de synthèse des modèles sélectionnés par l'utilisateur avec les seuils d'activité utilisés pour chaque propriété ADME-Tox. C) Page de résultats qui présente la table contenant les prédictions de tous les modèles, le score ADME-Tox et les scores utilisés pour générer les cartes d'activité.

3.4. Pistes d'amélioration et perspectives

L'application en ligne est entièrement développée selon ce qui avait été imaginé en première intention dans le but de mettre en place rapidement un outil fonctionnel. Elle n'est pas encore disponible et est en cours d'installation sur le serveur du laboratoire. Au cours du développement, des pistes d'amélioration ont été discutées et constituent des perspectives de travail pour compléter l'application actuellement élaborée.

La première perspective concerne le format des estimations transmises dans le cas d'une propriété continue. Certaines propriétés comme par exemple le LogS utilisé pour prédire la solubilité aqueuse ne sont pas directement exploitable par les chimistes. Comme nous l'avons présenté précédemment la modélisation cette propriété physico-chimique nécessite une conversion de la concentration (g/L en mol/L) suivie d'une transformation des valeurs selon le logarithme décimal. Par conséquent, la valeur estimée de cette propriété correspond à une concentration convertie et transformée. Le but serait de pouvoir reconvertir les prédictions d'un modèle de ce type, afin de transmettre une estimation pratique et facilement utilisable de tous.

La deuxième perspective concerne les résultats transmis au sujet du domaine d'applicabilité. Actuellement, l'information sur les plus proches voisins observés dans le DA n'est pas exploitée par l'application. Nous pensons qu'il serait intéressant de donner la possibilité aux utilisateurs de visualiser la structure des plus proches voisins dans le domaine d'interpolation. Il serait également avantageux d'ajouter aux résultats une information sur le coefficient de similarité du composé testé avec les composés du modèle. Cela peut permettre d'apporter un niveau d'information supplémentaire permettant de juger la qualité des prédictions transmises.

La troisième perspective concerne la coloration des cellules (gris, vert et rouge) dans la visualisation graphique des résultats. Actuellement, uniquement trois couleurs sont disponibles. Nous pensons qu'à l'avenir il serait plus avantageux de mettre en œuvre un système de coloration progressif, ou par gradient, permettant de faire une distinction parmis les molécules active et inactive selon les seuils d'activité. L'objectif serait alors d'utiliser une approche similaire à celle développée pour la détermination du coefficient d'activité (C_A). Ceci pourrait permettre par exemple d'identifier plus rapidement les composés qui disposent de propriétés très intéressantes dans le cadre d'un projet.

La quatrième perspective concerne le score ADME-Tox. Plusieurs améliorations peuvent être envisagées à ce sujet : i) Le score qui est utilisé à présent est simple et considère que toutes les propriétés possèdent la même importance. Cependant, dans le cadre de certains projets de découverte de nouveaux médicaments, il peut être avantageux de prioriser les résultats de certaines études. Il serait alors possible de pondérer le score ADME-Tox en fonction des propriétés considérées comme plus importantes que d'autres vis-à-vis d'un projet. ii) Ce score ADME-Tox permet de communiquer le potentiel d'un composé vis-à-vis des critères adoptés par l'utilisateur (seuils d'activité). Ce score pourrait être utilisé afin de générer une carte d'activité permettant de visualiser l'importance des sous-structures d'un composé sur son profile ADME-Tox. Le but serait de faciliter l'optimisation multidimensionnelle des composés lors de l'étape d'optimisation des *leads* (Ch1 1.2.4).

4. Conclusions

L'intérêt principal de la plateforme MetaPredict est de pouvoir créer de façon efficace des modèles QSAR pour une propriété d'intérêt. Comparé à un traitement manuel, un gain de temps et une réduction des erreurs sont observés grâce à l'automatisation du processus de modélisation. Comme les étapes de la plateforme sont standardisées, cette automatisation permet d'obtenir une approche reproductible afin d'échantillonner de façon exhaustive l'ensemble des conditions nécessaires à la création d'un modèle QSAR.

Comme présenté dans l'exemple du LogS, cette plateforme permet d'accéder à des modèles de prédiction optimisés. Ainsi, nous avons pu voir que chacune des étapes que nous avons créées permettent de déterminer les conditions optimales pour améliorer et affiner les modèles. Comme mentionné dans ce chapitre, les modèles contiennent l'ensemble des informations pour une utilisation en routine, comme par exemple la création et l'utilisation du domaine d'applicabilité. Concernant ce dernier, nous avons démontré que la prise en compte de la densité locale combinée à la fiabilité locale permettait la mise en œuvre d'un domaine d'applicabilité plus précis. Par conséquent, les modèles générés sont capables de fournir une information plus spécifique sur les molécules qu'ils sont en capacité de prédire avec fiabilité. Nous avons également montré que l'identification des modèles uniques pour la création d'un consensus permettait d'améliorer les prédictions des modèles à notre disposition. Ainsi, l'utilisation des modèles de consensus permet de fournir des prédictions plus utiles.

Nous avons également entrepris une validation expérimentale des modèles développés pour la prédiction de la fraction libre dans le plasma (Fu_p). Les résultats préliminaires nous montrent que les prédictions des modèles sont concordantes avec les mesures expérimentales transmises par l'équipe du Pr. Benoit DEPREZ au sein de la plateforme ADME de la faculté de pharmacie de Lille. Ces premiers résultats sont encourageants et ils devront être confirmés grâce à la validation sur l'ensemble des molécules testées.

Pour finir, nous avons prêté une attention particulière à ce que les modèles générés soient facilement accessibles et utilisables par nos collaborateurs. Pour cela, nous avons proposé une application en ligne. Cette application a pour objectif principal d'être simple et intuitive, afin de faciliter l'accès aux modèles ADME-Tox. Cette application en ligne incorpore de nombreuses options et donne la possibilité de représenter les propriétés d'intérêt sur la structure des composés testés. De surcroît, les cartes d'activité que nous avons proposé peuvent être modulées en fonction des spécificités d'un projet de recherche, afin d'être plus informatives en mettant en lumière l'information utile.

Conclusion générale

Le projet doctoral, mené au sein de l'ICOA et en collaboration avec Technologie Servier, avait pour principale mission de proposer des modèles QSAR pour la prédiction des propriétés ADME-Tox. Ce travail de thèse a été initié pour explorer un nouvel axe de recherche au sein de notre laboratoire et répondre aux exigences actuelles des projets de conception de médicaments. Ainsi, une estimation des propriétés ADME-Tox lors de étapes précoces de ce processus facilite les prises de décision pour prioriser des séries chimiques afin d'accélérer la découverte de nouveaux médicaments.

Pour répondre à cette problématique, nous avons dû dans un premier temps rechercher des données expérimentales pour plusieurs propriétés ADME-Tox. L'obtention de ces données a nécessité un nombre conséquent d'étapes manuelles pour le traitement, pour la vérification des points de mesures et des informations structurales, ainsi que pour l'uniformisation des données. Pour accélérer notre processus, nous avons également proposé les outils *ADMET Xtractor* pour l'extraction d'informations complémentaires contenues dans les descriptions des mesures expérimentales, *WebScraper* pour l'extraction de plusieurs bases de données en ligne et *WebChem* pour la recherche et la vérification des données structurales à partir de plusieurs bases de données chimiques. Au total, 169 496 mesures pour 49 propriétés ADME-Tox ont été extraites et uniformisées pour définir notre base de données ADMET db. Les perspectives envisagées concernent la collecte en quantité plus importante de données publiées, et en parallèle la récupération des conditions expérimentales ayant permis d'effectuer la mesure. Le but serait de centraliser l'ensemble de ces informations dans une base de données. Cela permettrait de mieux trier et sélectionner les mesures expérimentales pour la création de modèles QSAR spécifique à une utilisation, une méthode expérimentale, ou une population d'individus particulière.

Pour l'exploitation centralisée de ces données, nous avons mis au point une plateforme MetaPredict, un outil automatique de création de modèles QSAR. L'intérêt d'un tel outil est la mise en place de procédures standardisées et reproductibles dans les différentes étapes d'élaboration de modèles. La technologie mise en œuvre a pour but d'optimiser toutes ces étapes telles que : i) l'échantillonnage des différentes conditions nécessaires à la création d'un modèle (jeux de données, descripteurs et algorithmes), iii) l'identification du domaine d'applicabilité, iv) la combinaison de plusieurs modèles uniques dans un modèle de consensus, et pour finir v) la mise en forme des résultats. Plusieurs de ces

étapes ont nécessité un temps de développement important. Ce projet s'est voulu exhaustif dans sa capacité à prédire un large panel de propriétés ADME-Tox, tout en assurant l'obtention de modèles fiables et valides quelque soit la propriété modélisée. Sa conception permet alors d'envisager la prédiction de toutes propriétés pour lesquelles des données sont disponibles. A ce jour, la plateforme MetaPredict a permis de créer des modèles locaux et globaux pour 21 propriétés ADME-Tox parmi les 49 représentées dans la base ADMET db.

L'utilisation plus approfondie de la plateforme a été initiée et présentée pour deux propriétés, à savoir la solubilité aqueuse (LogS) et la fraction libre dans le plasma (Fu_p). Les résultats du modèle de solubilité aqueuse indiquent qu'une amélioration des performances est obtenue suite aux étapes d'optimisation effectuées par la plateforme, mais également que le domaine d'applicabilité, basé sur la densité locale et la fiabilité locale, apporte des estimations correctes des limites du modèle. Nous avons mis en lumière l'importance de notre approche de consensus pour améliorer la qualité des modèles produits. Pour la Fu_p, une validation expérimentale est en cours afin de valider la cohérence et la robustesse des prédictions des modèles générés. Les résultats préliminaires nous indiquent que ces modèles permettent de prédire correctement la propriété Fu_p. En règle générale, la plateforme permet de produire des modèles locaux ou globaux utiles et les résultats obtenus à l'aide des modèles de consensus sont très encourageants. La perspective générale de ces travaux est de compléter la collection de modèles ADME-Tox à notre disposition. De plus, comme mentionné dans le chapitre 3, la plateforme MetaPredict peut être améliorée afin d'optimiser la rapidité de l'outil et les informations transmises par les modèles.

Afin de rendre les modèles facilement utilisables par nos collaborateurs, nous avons proposé une application en ligne permettant de regrouper l'ensemble des modèles ADME-Tox générés. Cette application donne la possibilité de représenter l'activité modélisée sur la structure des composés testés. Cette application en ligne a été entièrement développée et est en cours de déploiement au sein du laboratoire. Les perspectives envisagées pour cette application en ligne comportent d'autres approches plus ambitieuses comme les cartes d'activité. L'objectif est toujours d'apporter des outils plus adaptés aux besoins des chimistes. La première approche consiste à proposer des analogues structuraux des molécules testées à l'aide des MMP, afin d'explorer virtuellement l'espace chimique dans le but d'identifier les modifications qui disposent d'une probabilité élevée d'améliorer les propriétés ADME-Tox des composés testés. La

deuxième approche a pour objectif de prédire l'ensemble des métabolites associés à la dégradation d'une molécule par l'organisme. L'énumération et la prédiction de la toxicité des métabolites peuvent permettre d'identifier rapidement les molécules susceptibles d'être dégradées en métabolites toxiques.

La perspective principale est actuellement de déployer l'application en ligne et la plateforme MetaPredict au sein du LMBA, le laboratoire mixte entre l'ICOA et Technologie Servier du campus Orléanais. L'objectif est de donner la possibilité aux chimistes d'avoir accès à cet outil dans le but d'orienter les projets de synthèse et d'accélérer la découverte de nouveaux principes actifs.

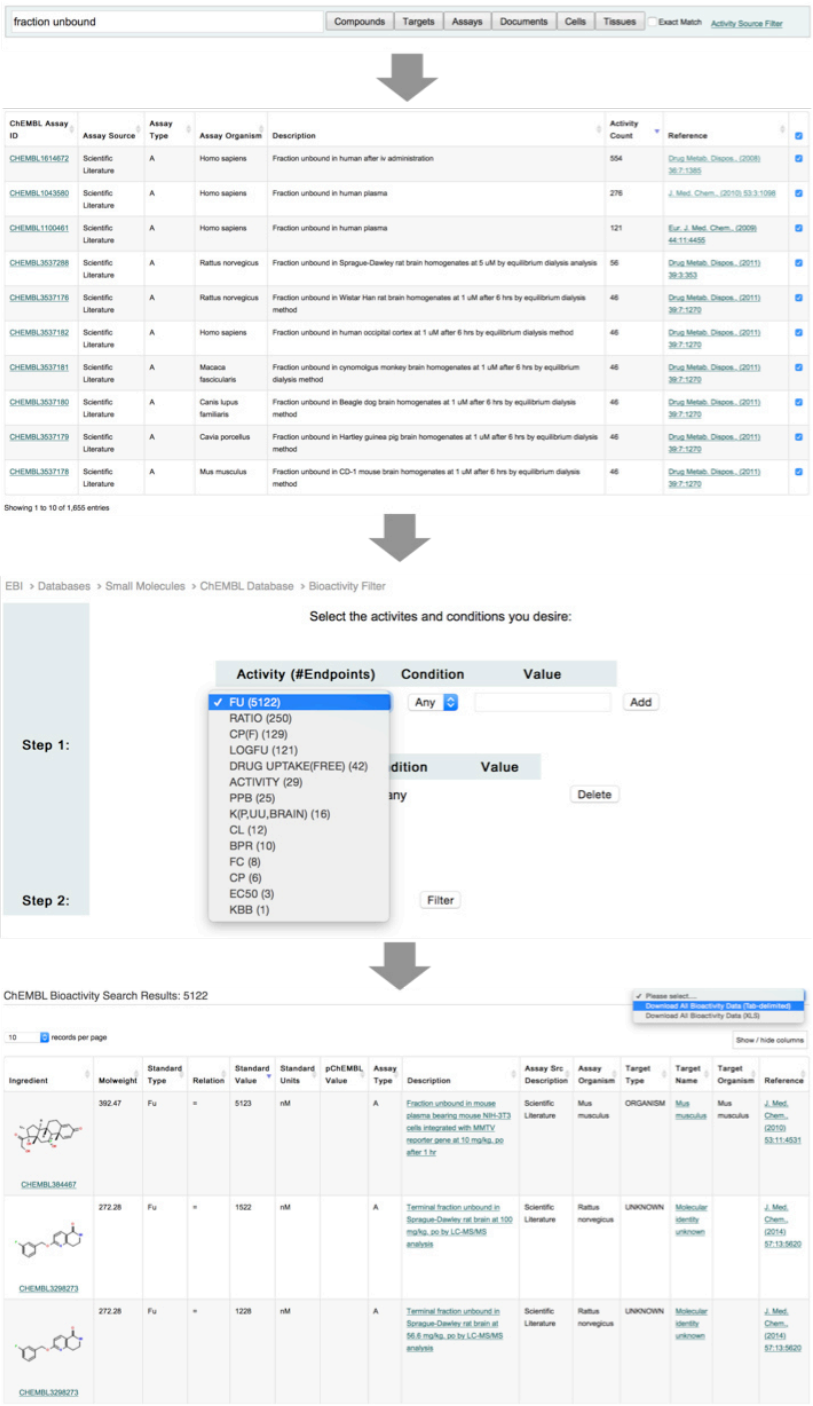
Pour ma part, ce projet m'a permis d'acquérir une vision d'ensemble des contraintes liées au milieu académique et au milieu industriel, qui encouragent la proposition de solutions innovantes tout en assurant un travail de recherche de pointe. Cette thèse a représenté un défi personnel et professionnel. De par ma formation universitaire de chimiste analytique, ce projet m'a donné la possibilité d'enrichir considérablement mes compétences transverses et pluridisciplinaires en chémoinformatique et en programmation, qui me seront utiles à l'avenir. Ainsi, j'ai eu l'occasion de découvrir, d'apprendre et de maitriser R, Python, PostgreSQL, KNIME, HTML, JavaScript, Regex, que j'ai su appliquer avec succès au domaine de la chimie pour développer et proposer plusieurs outils. Le premier est la plateforme MetaPredict actuellement disponible via une librairie Python, le deuxième est l'application en ligne permettant de valoriser les résultats de MetaPredict, et le troisième est un outil de comparaison de bases de données (OverLaper) proposé sous la forme d'un paquet R et développé dans le cadre d'un projet annexe de l'équipe.

ANNEXE A Liste des logiciels de prédiction ADME-Tox

LOGICIELS	LICENCE	PC	A	D	M	E	T
Percepta	Commerciale	✓	✓	✓	✓	✗	✓
QickProp	Commerciale	✓	✓	✓	✗	✗	✓
ADMET Predictor	Commerciale	✓	✓	✓	✓	✓	✓
ADMEWORKS Predictor	Commerciale	✓	✓	✓	✓	✗	✓
IMPACT-F	Commerciale	✗	✓	✗	✗	✗	✗
Metabolizer	Commerciale	✗	✗	✗	✓	✗	✗
BIOVIA	Commerciale	✓	✓	✗	✗	✗	✓
ToxGPS	Commerciale	✗	✗	✗	✗	✗	✓
Derek Nexus	Commerciale	✗	✗	✗	✗	✗	✓
Meteor Nexus	Commerciale	✗	✗	✗	✓	✗	✗
HazardExpert Pro	Commerciale	✗	✗	✗	✗	✗	✓
MetabolExpert	Commerciale	✗	✗	✗	✓	✗	✗
Leadscope	Commerciale	✗	✗	✗	✗	✗	✓
CASE Ultra	Commerciale	✗	✓	✓	✓	✗	✓
SimCYP	Commerciale	✓	✓	✓	✓	✓	✗
Cloe PK	Commerciale	✓	✓	✓	✓	✗	✗
MetaSite	Commerciale	✗	✗	✗	✓	✗	✗
StarDrop	Commerciale	✓	✓	✓	✓	✗	✓
OSIRIS Property Explorer	Libre	✓	✗	✗	✗	✗	✓
SwissADME	Libre	✓	✓	✗	✓	✗	✗
admetSAR	Libre	✓	✓	✓	✓		✓
pkCSM	Libre	✓	✓	✓	✓	✓	✓
Pred-Skin	Libre	✗	✗	✗	✗	✗	✓
Pred-hERG	Libre	✗	✗	✗	✗	✗	✓
Lazar	Libre	✗	✗	✗	✗	✗	✓
MetaPrint2D	Libre	✗	✗	✗	✓	✗	✗
PreADMET	Libre	✓	✓	✓	✓	✗	✓
MolCode Toolbox	Libre	✓	✗	✗	✗	✗	✓
FAME2	Libre	✗	✗	✗	✓	✗	✗
SMARTCyp	Libre	✗	✗	✗	✓	✗	✗
ChemBench	Libre	✗	✗	✗	✗	✗	✓
OCHEM	Libre	✓	✓	✓	✓	✓	✓

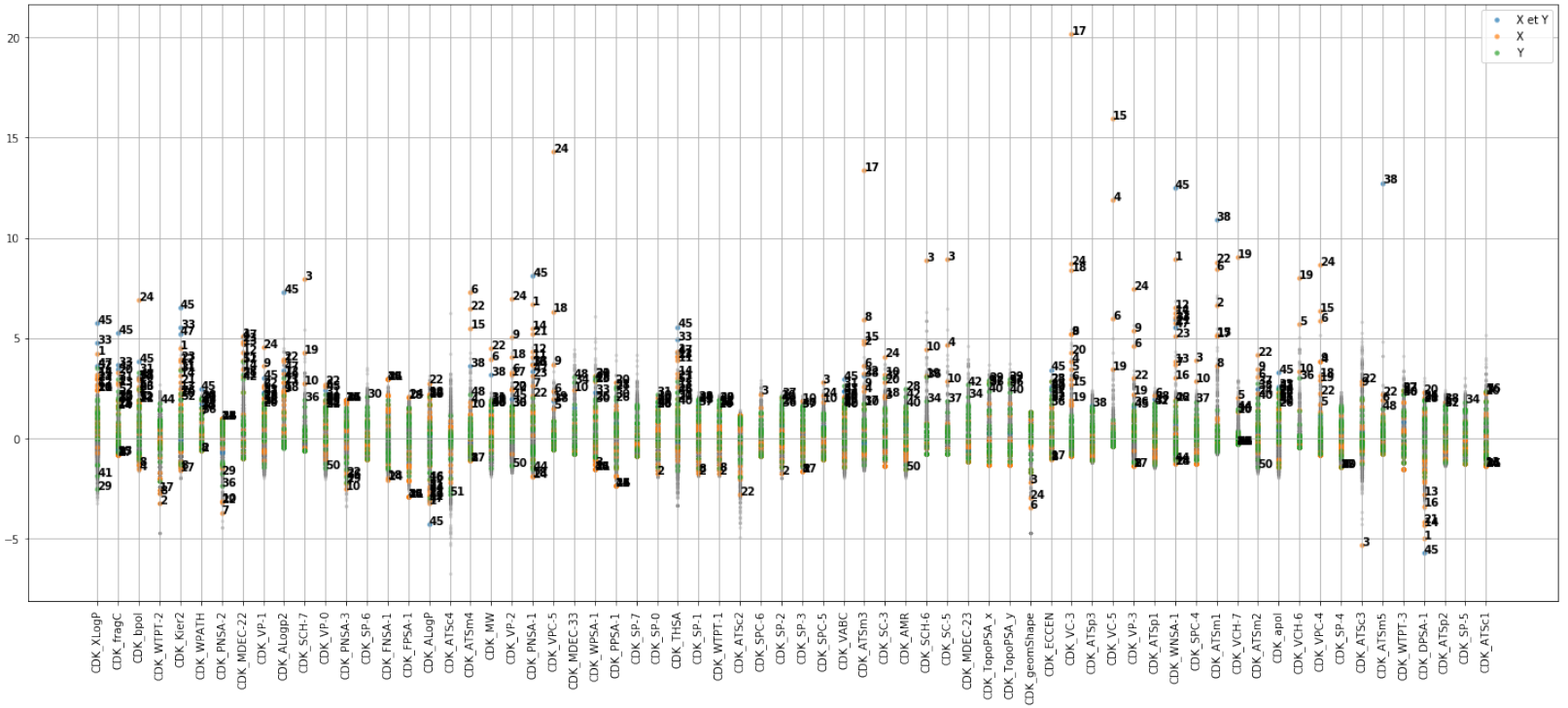
ANNEXE B Extraction manuelle de la base de données ChEMBL.

Exemple des étapes manuelles effectuées sur le site en ligne de la base ChEMBL afin d'extraire les données d'intérêt. Cette approche permet de bénéficier de tous les outils ChEMBL permettant d'appliquer les premiers filtres afin de sélectionner rapidement les données souhaitées.



ANNEXE C Analyse des individus aberrants observés pour le modèle de régression LogS.

Ce graphique représente la répartition des valeurs de tous les descripteurs CDK utilisés pour modéliser le LogS. Les descripteurs ont été ordonnés en fonction de leur importance dans le modèle. Les points aberrants sont colorés en fonction de leurs types d'erreurs (« X », « Y » ou « X et Y »). Seuls les points aberrants en dehors de l'intervalle de confiance de chaque descripteur (95 %) ont été représentés à l'aide de leurs identifiants respectifs.

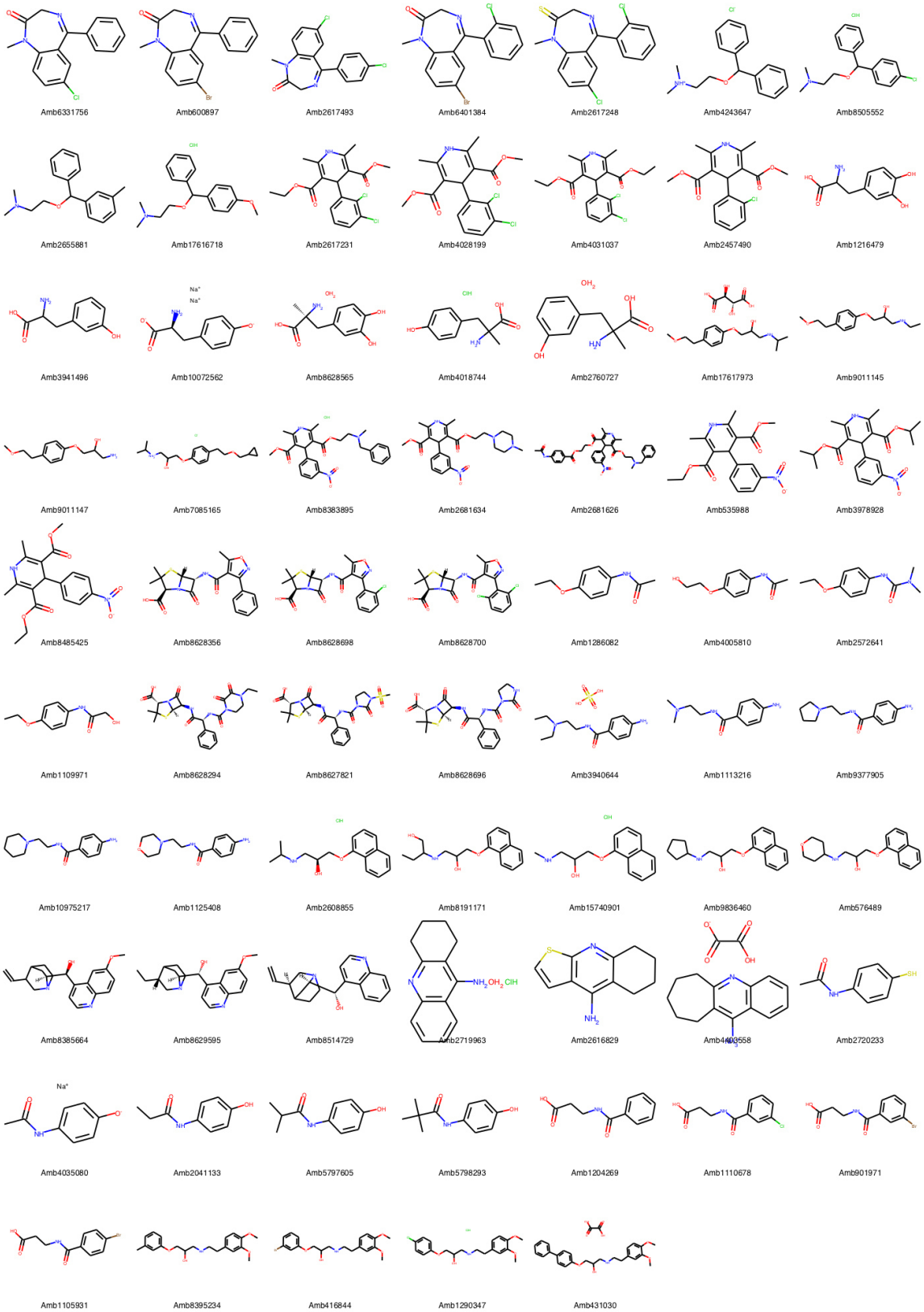


ANNEXE D Tests statistiques employés dans le cadre de la plateforme MetaPredict.

Test	Application	Variable mesurée	Conditions d'application	Hypothèse
Shapiro-Wilk	Ajustement d'une distribution observée à une loi normale	Une variable quantitative	1) La variable quantitative peut être continue ou discrète.	H_0 : La variable suit une loi normale. H_1 : La variable ne suit pas une loi normale.
Fisher-Snedecor	Comparaison de deux variances observées s_1^2 et s_2^2	Une variable quantitative et une variable qualitative à deux classes	1) Les classes de la variable qualitative doivent être exclusives. 2) Dans la population visée, les distributions de la variable quantitative dans chacune des classes de la variable qualitative doivent suivre une loi normale. 3) La variable quantitative doit être de préférence continue. 4) La variable qualitative est nominale. 5) La variable qualitative peut être fixée.	H_0 : Les variances sont identiques. H_1 : Les variances ne sont pas identiques.
***t* de Student**	Comparaison d'une moyenne observée à une moyenne théorique	Une variable quantitative	1) La variable quantitative doit suivre une loi normale dans la population visée. 2) La variable quantitative peut être continue ou discrète.	H_0 : La moyenne théorique est la moyenne réelle. H_1 : La moyenne théorique n'est pas la moyenne réelle.
Mann-Whitney-Wilcoxon	Comparaison de deux médianes observées.	Une variable quantitative et une variable qualitative à deux classes	1) Les classes de la variable qualitative doivent être exclusives 2) Dans la population visée, les distributions de la variable quantitative dans les classes de la variable qualitative doivent avoir la même forme, peu importe celle-ci. 3) La variable quantitative peut être continue ou discrète. Une variable semi-qualitative ou qualitative ordinale convient également. 4) La variable qualitative est nominale. 5) La variable qualitative peut être fixée.	H_0 : Les médianes sont identiques. H1 : Les médianes ne sont pas identiques.

Les informations présentées dans cette annexe sont inspirées du livre « ***Comprendre et réaliser les tests statistiques à l'aide de R : Manuel de biostatistique*** » de Gaël Millot [271].

ANNEXE E Structures des composés testés expérimentallement pour la validation des modèles de fraction libre dans le plasma.



ANNEXE F Résolution de l'équation d'une parabole.

Cette annexe présente la démarche que nous avons mis en place pour résoudre l'équation ($ax^2 + bx + c = y$) d'une parabole utilisée pour définir le coefficient d'activité nécessaire lors de la création des cartes d'activité. Pour cela, nous considérons les points A, B, et C de coordonnées respectives (x_1,y_1), (x_2,y_2) et (x_3,y_3). Les points A et B représentent les seuils d'activité de la propriété modélisée pour lesquels l'ordonnée est nulle. Le point C représente le sommet de la parabole pour lequel l'abscisse (x_3) est égale à la moyenne des seuils d'activité x_1 et x_2. L'équation de la parabole est résolue selon la démarche suivante :

$$\begin{cases} A\ (x_1,0) \\ B\ (x_2,0) \\ C\ (x_3,y_3) \end{cases} \Rightarrow \begin{cases} ax_1^2+bx_1+c=0 \\ ax_2^2+bx_2+c=0 \\ ax_3^2+bx_3+c=y_3 \end{cases} \Leftrightarrow \begin{cases} c=-ax_1^2-bx_1 \\ ax_2^2+bx_2-ax_1^2-bx_1=0 \\ ax_3^2+bx_3-ax_1^2-bx_1=y_3 \end{cases}$$

$$\begin{cases} c=-ax_1^2-bx_1 \\ b=-a\times\left(\dfrac{x_2^2-x_1^2}{x_2-x_1}\right) \\ a\left(x_3^2-x_1^2\right)+b\left(x_3-x_1\right)=y_3 \end{cases} \Leftrightarrow \begin{cases} c=-ax_1^2-bx_1 \\ b=-a\times\left(\dfrac{x_2^2-x_1^2}{x_2-x_1}\right) \\ a=y_3\Big/\left(x_3^2-x_1^2\right)-\left(\left(x_2^2-x_1^2/x_2-x_1\right)\times\left(x_3-x_1\right)\right) \end{cases}$$

Références bilbiographiques

1. Bohacek, R. S., McMartin, C. & Guida, W. C. The art and practice of structure-based drug design: A molecular modeling perspective. *Med. Res. Rev.* **16,** 3–50 (1996).

2. CAS REGISTRY - Chemical Substances. Available at: http://support.cas.org/content/chemical-substances. (Accessed: 15th March 2018)

3. Harvey, A. L., Edrada-Ebel, R. & Quinn, R. J. The re-emergence of natural products for drug discovery in the genomics era. *Nat. Rev. Drug Discov.* **14,** 111–129 (2015).

4. Berg, J. M., Tymoczko, J. L., Jr, G. J. G. & Stryer. *Biochemistry*. (W. H. Freeman, 2015).

5. *La chimie et la nature*. (EDP sciences, 2012).

6. Gashaw, I., Ellinghaus, P., Sommer, A. & Asadullah, K. What makes a good drug target? *Drug Discov. Today* **16,** 1037–1043 (2011).

7. Schmidtke, P. & Barril, X. Understanding and predicting druggability. A high-throughput method for detection of drug binding sites. *J. Med. Chem.* **53,** 5858–5867 (2010).

8. Hajduk, P. J., Huth, J. R. & Fesik, S. W. Druggability indices for protein targets derived from NMR-based screening data. *J. Med. Chem.* **48,** 2518–2525 (2005).

9. Kim, B. *et al.* Clinical validity of the lung cancer biomarkers identified by bioinformatics analysis of public expression data. *Cancer Res.* **67,** 7431–7438 (2007).

10. Cheng, A. C. *et al.* Structure-based maximal affinity model predicts small-molecule druggability. *Nat. Biotechnol.* **25,** 71–75 (2007).

11. Krimm, I. [Fragment-based screening: a promising avenue for drug design]. *Med. Sci. MS* **31,** 197–202 (2015).

12. Spring, D. R. Diversity-oriented synthesis; a challenge for synthetic chemists. *Org. Biomol. Chem.* **1,** 3867–3870 (2003).

13. Schneider, G. Automating drug discovery. *Nat. Rev. Drug Discov.* **17,** 97–113 (2018).

14. Barata, D., van Blitterswijk, C. & Habibovic, P. High-throughput screening approaches and combinatorial development of biomaterials using microfluidics. *Acta Biomater.* **34,** 1–20 (2016).

15. Shun, T. Y., Lazo, J. S., Sharlow, E. R. & Johnston, P. A. Identifying Actives from HTS Data Sets: Practical Approaches for the Selection of an Appropriate HTS Data-Processing Method and Quality Control Review. *J. Biomol. Screen.* **16,** 1–14 (2011).

16. Keserű, G. M. & Makara, G. M. Hit discovery and hit-to-lead approaches. *Drug Discov. Today* **11,** 741–748 (2006).

17. Johnston, P. A. & Johnston, P. A. Cellular platforms for HTS: three case studies. *Drug Discov. Today* **7,** 353–363 (2002).

18. Hughes, J. P., Rees, S., Kalindjian, S. B. & Philpott, K. L. Principles of early drug discovery. *Br. J. Pharmacol.* **162,** 1239–1249 (2011).

19. Lipinski, C. A., Lombardo, F., Dominy, B. W. & Feeney, P. J. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Deliv. Rev.* **46,** 3–26 (2001).

20. Veber, D. F. *et al.* Molecular properties that influence the oral bioavailability of drug candidates. *J. Med. Chem.* **45,** 2615–2623 (2002).

21. Zhang, M.-Q. & Wilkinson, B. Drug discovery beyond the 'rule-of-five'. *Curr. Opin. Biotechnol.* **18,** 478–488 (2007).

22. Doak, B. C., Over, B., Giordanetto, F. & Kihlberg, J. Oral druggable space beyond the rule of 5: insights from drugs and clinical candidates. *Chem. Biol.* **21,** 1115–1142 (2014).

23. Lüllmann, H., Mohr, K. & Hein, L. *Atlas de poche de pharmacologie*. (Médecine Sciences Publ., 2010).

24. Drug Discovery and ADMET process: A Review. *ResearchGate* Available at: https://www.researchgate.net/publication/309425682_Drug_Discovery_and_ADMET_process_A_Review. (Accessed: 16th June 2018)

25. Sliwoski, G., Kothiwale, S., Meiler, J. & Lowe, E. W. Computational Methods in Drug Discovery. *Pharmacol. Rev.* **66,** 334–395 (2013).

26. Nissink, J. W. M., Schmitt, S., Blackburn, S. & Peters, S. Stratified high-throughput screening sets enable flexible screening strategies from a single plated collection. *J. Biomol. Screen.* **19,** 369–378 (2014).

27. Cusack, K. P. *et al.* Design strategies to address kinetics of drug binding and residence time. *Bioorg. Med. Chem. Lett.* **25,** 2019–2027 (2015).

28. De Vivo, M., Masetti, M., Bottegoni, G. & Cavalli, A. Role of Molecular Dynamics and Related Methods in Drug Discovery. *J. Med. Chem.* **59,** 4035–4061 (2016).

29. Braka, A. Prédiction de la cinétique des inhibiteurs de protéines kinases et de leur affinité par docking flexible. (Orléans, 2018).

30. Lavecchia, A. & Di Giovanni, C. Virtual screening strategies in drug discovery: a critical review. *Curr. Med. Chem.* **20,** 2839–2860 (2013).

31. Beuming, T. *et al.* Docking and Virtual Screening Strategies for GPCR Drug Discovery. *Methods Mol. Biol. Clifton NJ* **1335,** 251–276 (2015).

32. McInnes, C. Virtual screening strategies in drug discovery. *Curr. Opin. Chem. Biol.* **11,** 494–502 (2007).

33. Phatak, S. S., Stephan, C. C. & Cavasotto, C. N. High-throughput and in silico screenings in drug discovery. *Expert Opin. Drug Discov.* **4,** 947–959 (2009).

34. Clinical Development Success Rates 2006-2015 - BIO, Biomedtracker, Amplion 2016 | Drug Development | Leukemia. *Scribd* Available at: https://www.scribd.com/document/339203797/Clinical-Development-Success-Rates-2006-2015-BIO-Biomedtracker-Amplion-2016. (Accessed: 15th March 2018)

35. Adams, C. P. & Brantner, V. V. Spending on new drug development1. *Health Econ.* **19,** 130–141 (2010).

36. Cook, D. *et al.* Lessons learned from the fate of AstraZeneca's drug pipeline: a five-dimensional framework. *Nat. Rev. Drug Discov.* **13,** 419–431 (2014).

37. Morgan, P. *et al.* Impact of a five-dimensional framework on R&D productivity at AstraZeneca. *Nat. Rev. Drug Discov.* **17,** 167–181 (2018).

38. van de Waterbeemd, H. & Gifford, E. ADMET in silico modelling: towards prediction paradise? *Nat. Rev. Drug Discov.* **2,** 192–204 (2003).

39. Tetko, I. V., Engkvist, O., Koch, U., Reymond, J.-L. & Chen, H. BIGCHEM: Challenges and Opportunities for Big Data Analysis in Chemistry. *Mol. Inform.* **35,** 615–621 (2016).

40. Novotarskyi, S. *et al.* ToxCast EPA in Vitro to in Vivo Challenge: Insight into the Rank-I Model. *Chem. Res. Toxicol.* **29,** 768–775 (2016).

41. Cases, M. *et al.* The eTOX data-sharing project to advance in silico drug-induced toxicity prediction. *Int. J. Mol. Sci.* **15,** 21136–21154 (2014).

42. Briggs, K. *et al.* Inroads to predict in vivo toxicology-an introduction to the eTOX Project. *Int. J. Mol. Sci.* **13,** 3820–3846 (2012).

43. Kerns, E. H. & Di, L. Pharmaceutical profiling in drug discovery. *Drug Discov. Today* **8,** 316–323 (2003).

44. van De Waterbeemd, H., Smith, D. A., Beaumont, K. & Walker, D. K. Property-based design: optimization of drug absorption and pharmacokinetics. *J. Med. Chem.* **44,** 1313–1333 (2001).

45. Sugano, K. Aqueous boundary layers related to oral absorption of a drug: from dissolution of a drug to carrier mediated transport and intestinal wall metabolism. *Mol. Pharm.* **7,** 1362–1373 (2010).

46. Zakeri-Milani, P. & Valizadeh, H. Intestinal transporters: enhanced absorption through P-glycoprotein-related drug interactions. *Expert Opin. Drug Metab. Toxicol.* **10,** 859–871 (2014).

47. Lundquist, P. & Artursson, P. Oral absorption of peptides and nanoparticles across the human intestine: Opportunities, limitations and studies in human tissues. *Adv. Drug Deliv. Rev.* **106,** 256–276 (2016).

48. N. Brönsted, J. Einige Bemerkungen über den Begriff der Säuren und Basen. *Recl. Trav. Chim. Pays-Bas* **42,** 718–728 (2010).

49. M. Lowry, T. The Uniqueness of Hydrogen. *J. Soc. Chem. Ind.* **42,** 43–47 (2007).

50. Jr Webb, K. E. *Intestinal absorption of protein hydrolysis products: A review*. **68,** (1990).

51. Miyasaki, T., Sato, M., Yoshinaka, R. & Sakaguchi, M. Intestinal Absorption and the Activity of Enzymatic Hydrolysis of Ascorbyl-2-Phosphate in Rainbow Trout. *NIPPON SUISAN GAKKAISHI* **59,** 2059–2064 (1993).

52. Ohura, K. *et al.* Effect of intestinal first-pass hydrolysis on the oral bioavailability of an ester prodrug of fexofenadine. *J. Pharm. Sci.* **101,** 3264–3274 (2012).

53. Nigam, S. K. What do drug transporters really do? *Nat. Rev. Drug Discov.* **14,** 29–44 (2015).

54. Estudante, M., Morais, J. G., Soveral, G. & Benet, L. Z. Intestinal drug transporters: an overview. *Adv. Drug Deliv. Rev.* **65,** 1340–1356 (2013).

55. Kato, M. Intestinal first-pass metabolism of CYP3A4 substrates. *Drug Metab. Pharmacokinet.* **23,** 87–94 (2008).

56. Indiana University, School of Medicine, Department of Medicine. Flockhart Table. (2007). Available at: http://medicine.iupui.edu/clinpharm/ddis/main-table. (Accessed: 18th March 2018)

57. Kivistö, K. T., Niemi, M. & Fromm, M. F. Functional interaction of intestinal CYP3A4 and P-glycoprotein. *Fundam. Clin. Pharmacol.* **18,** 621–626 (2004).

58. Hou, T., Wang, J. & Li, Y. ADME Evaluation in Drug Discovery. 8. The Prediction of Human Intestinal Absorption by a Support Vector Machine. *J. Chem. Inf. Model.* **47,** 2408–2415 (2007).

59. *Recent Advances in QSAR Studies: Methods and Applications*. (Springer Netherlands, 2010).

60. Peters Jr., T. 3 - Ligand Binding by Albumin. in *All About Albumin* 76–132 (Academic Press, 1995). doi:10.1016/B978-012552110-9/50005-2

61. II - BIOCHEMISTRY AND BIOLOGY OF PLASMA ENZYMES. in *Enzymes in Blood Plasma* (ed. Hess, B.) 5–65 (Academic Press, 1963). doi:10.1016/B978-1-4832-3176-1.50008-7

62. Golden, P. L. & Pollack, G. M. Blood-brain barrier efflux transport. *J. Pharm. Sci.* **92,** 1739–1753 (2003).

63. Pardridge, W. Crossing the blood-brain barrier: Are we getting it right? *Drug Discov. Today* **6,** 1–2 (2001).

64. Meyer, U. A. Overview of enzymes of drug metabolism. *J. Pharmacokinet. Biopharm.* **24,** 449–459 (1996).

65. Gillette, J. R. Metabolism of Drugs and Other Foreign Compounds by Enzymatic Mechanisms. in *Progress in Drug Research / Fortschritte der Arzneimittelforschung / Progrès des recherches pharmaceutiques* 11–73 (Birkhäuser Basel, 1963). doi:10.1007/978-3-0348-7050-4_1

66. Hodge, H. C. Detoxication Mechanisms. The Metabolism and Detoxication of Drugs, Toxic Substances and Other Organic Compounds. *J. Am. Chem. Soc.* **83,** 759–759 (1961).

67. Xu, C., Li, C. Y.-T. & Kong, A.-N. T. Induction of phase I, II and III drug metabolism/transport by xenobiotics. *Arch. Pharm. Res.* **28,** 249–268 (2005).

68. Williams, J. A. *et al.* Drug-Drug Interactions for Udp-Glucuronosyltransferase Substrates: A Pharmacokinetic Explanation for Typically Observed Low Exposure (auci/Auc) Ratios. *Drug Metab. Dispos.* **32,** 1201–1208 (2004).

69. Ingelman-Sundberg, M. The human genome project and novel aspects of cytochrome P450 research. *Toxicol. Appl. Pharmacol.* **207,** 52–56 (2005).

70. Lewis, D. F. V. 57 varieties: the human cytochromes P450. *Pharmacogenomics* **5,** 305–318 (2004).

71. Zanger, U. M., Turpeinen, M., Klein, K. & Schwab, M. Functional pharmacogenetics/genomics of human cytochromes P450 involved in drug biotransformation. *Anal. Bioanal. Chem.* **392,** 1093–1108 (2008).

72. Hernandez, J. P., Mota, L. C., Huang, W., Moore, D. D. & Baldwin, W. S. Sexually dimorphic regulation and induction of P450s by the constitutive androstane receptor (CAR). *Toxicology* **256,** 53–64 (2009).

73. Liu, A., Wang, C., Hehir, M., Zhou, T. & Yang, J. In vivo induction of CYP in mice by carbamazepine is independent on PXR. *Pharmacol. Rep. PR* **67,** 299–304 (2015).

74. Briolotti, P. *et al.* Analysis of Glycogen Synthase Kinase Inhibitors That Regulate Cytochrome P450 Expression in Primary Human Hepatocytes by Activation of β-Catenin, Aryl Hydrocarbon Receptor and Pregnane X Receptor Signaling. *Toxicol. Sci. Off. J. Soc. Toxicol.* **148,** 261–275 (2015).

75. Pirmohamed, M. & Park, B. K. Cytochrome P450 enzyme polymorphisms and adverse drug reactions. *Toxicology* **192,** 23–32 (2003).

76. Michalets, E. L. Update: clinically significant cytochrome P-450 drug interactions. *Pharmacotherapy* **18,** 84–112 (1998).

77. Johnson, M. D., Newkirk, G. & White, J. R. Clinically significant drug interactions. *Postgrad. Med.* **105,** 193–195, 200, 205-206 passim (1999).

78. Masimirembwa, C. M., Bredberg, U. & Andersson, T. B. Metabolic stability for drug discovery and development: pharmacokinetic and biochemical challenges. *Clin. Pharmacokinet.* **42,** 515–528 (2003).

79. Nutrition, C. for F. S. and A. Guidance Documents & Regulatory Information by Topic - Guidance for Industry: Preparation of Food Contact Notifications for Food Contact Substances (Toxicology Recommendations). Available at: https://www.fda.gov/food/guidanceregulation/guidancedocumentsregulatoryinformation/ucm081825.htm. (Accessed: 21st April 2018)

80. Liebler, D. C. & Guengerich, F. P. Elucidating mechanisms of drug-induced toxicity. *Nat. Rev. Drug Discov.* **4,** 410–420 (2005).

81. Oyama, T. *et al.* Expression of cytochrome P450 in tumor tissues and its association with cancer development. *Front. Biosci. J. Virtual Libr.* **9,** 1967–1976 (2004).

82. Löhr, M., McFadyen, M. C. E., Murray, G. I. & Melvin, W. T. Cytochrome P450 enzymes and tumor therapy. *Mol. Cancer Ther.* **3,** 1503; author reply 1503-1504 (2004).

83. McFadyen, M. C. E., Melvin, W. T. & Murray, G. I. Cytochrome P450 enzymes: novel options for cancer therapeutics. *Mol. Cancer Ther.* **3,** 363–371 (2004).

84. Pelkonen, O. *et al.* Inhibition and induction of human cytochrome P450 enzymes: current status. *Arch. Toxicol.* **82,** 667–715 (2008).

85. Kongsamut, S., Kang, J., Chen, X.-L., Roehr, J. & Rampe, D. A comparison of the receptor binding and HERG channel affinities for a series of antipsychotic drugs. *Eur. J. Pharmacol.* **450,** 37–41 (2002).

86. Johnson, T. E. *et al.* Statins induce apoptosis in rat and human myotube cultures by inhibiting protein geranylgeranylation but not ubiquinone. *Toxicol. Appl. Pharmacol.* **200,** 237–250 (2004).

87. Zipes, D. P. *et al.* Rosuvastatin: an independent analysis of risks and benefits. *MedGenMed Medscape Gen. Med.* **8,** 73 (2006).

88. Landsteiner, K. & Jacobs, J. STUDIES ON THE SENSITIZATION OF ANIMALS WITH SIMPLE CHEMICAL COMPOUNDS. *J. Exp. Med.* **61,** 643–656 (1935).

89. Chipinda, I., Hettick, J. M. & Siegel, P. D. Haptenation: chemical reactivity and protein binding. *J. Allergy* **2011,** 839682 (2011).

90. Walsh, J. S. & Miwa, G. T. Bioactivation of drugs: risk and drug design. *Annu. Rev. Pharmacol. Toxicol.* **51,** 145–167 (2011).

91. Stepan, A. F. *et al.* Structural alert/reactive metabolite concept as applied in medicinal chemistry to mitigate the risk of idiosyncratic drug toxicity: a perspective based on the critical examination of trends in the top 200 drugs marketed in the United States. *Chem. Res. Toxicol.* **24,** 1345–1410 (2011).

92. Uetrecht, J. & Naisbitt, D. J. Idiosyncratic Adverse Drug Reactions: Current Concepts. *Pharmacol. Rev.* **65,** 779–808 (2013).

93. de Boer, J. & van Bavel, B. European 'REACH' (Registration, Evaluation, Authorisation and Restriction of Chemicals) program. *J. Chromatogr. A* **1216,** 301 (2009).

94. Guengerich, F. P. Mechanisms of Drug Toxicity and Relevance to Pharmaceutical Development. *Drug Metab. Pharmacokinet.* **26,** 3–14 (2011).

95. Kar, S. & Roy, K. Development and validation of a robust QSAR model for prediction of carcinogenicity of drugs. *Indian J. Biochem. Biophys.* **48,** 111–122 (2011).

96. Kar, S. & Roy, K. Predictive toxicology using QSAR: A perspective. *J. Indian Chem. Soc.* **87,** 1455–1515 (2010).

97. Hillebrecht, A. *et al.* Comparative evaluation of in silico systems for ames test mutagenicity prediction: scope and limitations. *Chem. Res. Toxicol.* **24,** 843–854 (2011).

98. Bakhtyari, N. G., Raitano, G., Benfenati, E., Martin, T. & Young, D. Comparison of in silico models for prediction of mutagenicity. *J. Environ. Sci. Health Part C Environ. Carcinog. Ecotoxicol. Rev.* **31,** 45–66 (2013).

99. Milan, C., Schifanella, O., Roncaglioni, A. & Benfenati, E. Comparison and possible use of in silico tools for carcinogenicity within REACH legislation. *J. Environ. Sci. Health Part C Environ. Carcinog. Ecotoxicol. Rev.* **29,** 300–323 (2011).

100. Ames, B. N., Gurney, E. G., Miller, J. A. & Bartsch, H. Carcinogens as frameshift mutagens: metabolites and derivatives of 2-acetylaminofluorene and other aromatic amine carcinogens. *Proc. Natl. Acad. Sci. U. S. A.* **69,** 3128–3132 (1972).

101. Diaz Ochoa, J. G. *et al.* A multi-scale modeling framework for individualized, spatiotemporal prediction of drug effects and toxicological risk. *Front. Pharmacol.* **3,** 204 (2012).

102. Frid, A. A. & Matthews, E. J. Prediction of drug-related cardiac adverse effects in humans--B: use of QSAR programs for early detection of drug-induced cardiac toxicities. *Regul. Toxicol. Pharmacol. RTP* **56,** 276–289 (2010).

103. Huang, L.-C., Wu, X. & Chen, J. Y. Predicting adverse side effects of drugs. *BMC Genomics* **12 Suppl 5,** S11 (2011).

104. Luyendyk, J. P., Roth, R. A. & Ganey, P. E. 9.13 - Inflammation and Hepatotoxicity. in *Comprehensive Toxicology (Second Edition)* (ed. McQueen, C. A.) 295–317 (Elsevier, 2010). doi:10.1016/B978-0-08-046884-6.01031-9

105. U.S. Department of Health and Human Services Food and Drug Administration, Center for Drug Evaluation and Research (CDER) & Center for Biologics Evaluation and Research (CBER). *Guidance for Industry Drug-Induced Liver Injury: Premarketing Clinical Evaluation*. (2009).

106. Union, P. O. of the E. The use of computational methods in the toxicological assessment of chemicals in food : current status and future prospects. (2011). Available at: https://publications.europa.eu/en/publication-detail/-/publication/49b38592-9a5c-4b98-aacf-4b66d71d8377/language-en. (Accessed: 16th April 2018)

107. Cheng, F. *et al.* Adverse drug events: database construction and in silico prediction. *J. Chem. Inf. Model.* **53,** 744–752 (2013).

108. Mitchell, J. B. O. Machine learning methods in chemoinformatics. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **4,** 468–481 (2014).

109. Brown, A. C. & Fraser, T. R. V.—On the Connection between Chemical Constitution and Physiological Action. Part. I. On the Physiological Action of the Salts of the Ammonium Bases, derived from Strychnia, Brucia, Thebaia, Codeia, Morphia, and Nicotia. *Earth Environ. Sci. Trans. R. Soc. Edinb.* **25,** 151–203 (1868).

110. Richet, C. Note sur le rapport entre la toxicité et les propriétés physiques des corps. *Soc Biol Compt Rend* 775–76 (1893).

111. Missner, A. & Pohl, P. 110 Years of the Meyer–Overton Rule: Predicting Membrane Permeability of Gases and Other Small Compounds. *Chemphyschem Eur. J. Chem. Phys. Phys. Chem.* **10,** 1405–1414 (2009).

112. Hammett, L. P. Some Relations between Reaction Rates and Equilibrium Constants. *Chem. Rev.* **17,** 125–136 (1935).

113. Taft, R. W. Linear Steric Energy Relationships. *J. Am. Chem. Soc.* **75,** 4538–4539 (1953).

114. Hansch, C. & Fujita, T. p-σ-π Analysis. A Method for the Correlation of Biological Activity and Chemical Structure. *J. Am. Chem. Soc.* **86,** 1616–1626 (1964).

115. Jaworska, J., Comber, M., Auer, C. & Van Leeuwen, K. V. L. Summary of a Workshop on Regulatory Acceptance of (Q)SARs for Human Health and Environmental Endpoints. *Environ. Health Perspect.* **111,** 1358–60 (2003).

116. H. Unger, S. & Hansch, C. J. On model building in structure-activity relationships. A reexamination of adrenergic blocking activity of β-halo-β-arylalkylamines. *J. Med. Chem.* **16,** 745–9 (1973).

117. OECD. OECD PRINCIPLES FOR THE VALIDATION, FOR REGULATORY PURPOSES, OF (QUANTITATIVE) STRUCTURE-ACTIVITY RELATIONSHIP MODELS. (2004). Available at: http://www.oecd.org/chemicalsafety/risk-assessment/37849783.pdf. (Accessed: 18th April 2018)

118. ORGANISATION FOR ECONOMIC CO-OPERATION AND DEVELOPMENT. GUIDANCE DOCUMENT ON THE VALIDATION OF (QUANTITATIVE)STRUCTURE-ACTIVITY RELATIONSHIPS [(Q)SAR] MODELS. **69,** (2007).

119. Dearden, J. C., Cronin, M. T. D. & Kaiser, K. L. E. How not to develop a quantitative structure-activity or structure-property relationship (QSAR/QSPR). *SAR QSAR Environ. Res.* **20,** 241–266 (2009).

120. Scior, T. *et al.* How to recognize and workaround pitfalls in QSAR studies: a critical review. *Curr. Med. Chem.* **16,** 4297–4313 (2009).

121. Varnek, A. & Baskin, I. Machine learning methods for property prediction in chemoinformatics: Quo Vadis? *J. Chem. Inf. Model.* **52,** 1413–1437 (2012).

122. Cherkasov, A. *et al.* QSAR modeling: where have you been? Where are you going to? *J. Med. Chem.* **57,** 4977–5010 (2014).

123. Tropsha, A., Gramatica, P. & Gombar, V. K. The importance of being earnest: Validation is the absolute essential for successful application and interpretation of QSPR models. *QSAR Comb. Sci.* **22,** 69–77 (2003).

124. T.D. Cronin, M. & Schultz, T. W. Pitfalls in QSAR. *J. Mol. Struct.-Theochem - J MOL STRUC-THEOCHEM* **622,** (2003).

125. Young, D., Martin, T., Venkatapathy, R. & Harten, P. Are the chemical structures in your QSAR correct? QSAR Comb Sci. *QSAR Comb. Sci.* **27,** 1337–1345 (2008).

126. Fourches, D., Muratov, E. & Tropsha, A. Trust, But Verify: On the Importance of Chemical Structure Curation in Cheminformatics and QSAR Modeling Research. *J. Chem. Inf. Model.* **50,** 1189–204 (2010).

127. Roy, P. P., Joseph, T. & Roy, K. Exploring the impact of size of training sets for the development of predictive QSAR models. *Chemom. Intell. Lab. Syst.* **90,** 31–42 (2008).

128. Tropsha, A. Best Practices for QSAR Model Development, Validation, and Exploitation. *Mol. Inform.* **29,** (2010).

129. *Handbook of molecular descriptors*. (Wiley-VCH, 2000).

130. Landrum, G. *RDKit: Open-source cheminformatics*.

131. Steinbeck, C. *et al.* The Chemistry Development Kit (CDK): An Open-Source Java Library for Chemo- and Bioinformatics. *J. Chem. Inf. Comput. Sci.* **43,** 493–500 (2003).

132. Moriwaki, H., Tian, Y.-S., Kawashita, N. & Takagi, T. Mordred: a molecular descriptor calculator. *J. Cheminformatics* **10,** 4 (2018).

133. Ruggiu, F., Marcou, G., Varnek, A. & Horvath, D. ISIDA Property-Labelled Fragment Descriptors. *Mol. Inform.* **29,** 855–868 (2010).

134. *VolSurf+*. (Molecular Discovery, 2018).

135. *Molecular Operating Environment (MOE)*. (Chemical Computing Group ULC, 2018).

136. *Dragon 7*. (Kode chemoinformatics, 2018).

137. *ADRIANA.Code*. (Molecular Networks).

138. JChem 15.2.9, 2015, ChemAxon (http://www;chemaxon.com). *JChem 15.2.9, 2015, ChemAxon (http://www;chemaxon.com)*

139. Cao, D.-S., Xu, Q., Hu, Q. & Liang, Y.-Z. manual for chemopy. (2013).

140. Roy, K., Kar, S. & Das, R. N. Chemical Information and Descriptors. in *Understanding the Basics of QSAR for Applications in Pharmaceutical Sciences and Risk Assessment* 47–80 (Elsevier, 2015). doi:10.1016/B978-0-12-801505-6.00002-8

141. Gasteiger, J. & Engel, T. *Chemoinformatics: a textbook*. (Wiley-VCH, 2003).

142. Johnson, T. W. *et al.* Discovery of (10R)-7-Amino-12-fluoro-2,10,16-trimethyl-15-oxo-10,15,16,17-tetrahydro-2H-8,4-(metheno)pyrazolo[4,3-h][2,5,11]-benzoxadiazacyclotetradecine-3-carbonitrile (PF-06463922), a Macrocyclic Inhibitor of Anaplastic Lymphoma Kinase (ALK) and c-ros Oncogene 1 (ROS1) with Preclinical Brain Exposure and Broad-Spectrum Potency against ALK-Resistant Mutations. *J. Med. Chem.* **57,** 4720–4744 (2014).

143. Baskin, I. & Varnek, A. ChemInform Abstract: Fragment Descriptors in SAR/QSAR/QSPR Studies, Molecular Similarity Analysis and in Virtual Screening. in *ChemInform* **40,** 1–43 (2009).

144. D. Christie, B., A. Leland, B. & G. Nourse, J. Structure searching in chemical databases by direct lookup methods. *J. Chem. Inf. Comput. Sci.* **33,** 545–547 (1993).

145. Chen, X., Rusinko, A. & Young, S. Recursive Partitioning Analysis of a Large Structure-Activity Data Set Using Three-Dimensional Descriptors. *J. Chem. Inf. Comput. Sci.* **38,** 1054–1062 (1998).

146. Durant, J. L., Leland, B. A., Henry, D. R. & Nourse, J. G. Reoptimization of MDL Keys for Use in Drug Discovery. *J. Chem. Inf. Comput. Sci.* **42,** 1273–1280 (2002).

147. Dassault Systèmes. *THE KEYS TO UNDERSTANDING MDL KEYSET TECHNOLOGY*. (2014).

148. Lee, A. C., Shedden, K., Rosania, G. R. & Crippen, G. M. Data mining the NCI60 to predict generalized cytotoxicity. *J. Chem. Inf. Model.* **48,** 1379–1388 (2008).

149. Banerjee, P., Siramshetty, V. B., Drwal, M. N. & Preissner, R. Computational methods for prediction of in vitro effects of new chemical structures. *J. Cheminformatics* **8,** 51 (2016).

150. Bolton, E., Wang, Y., A. Thiessen, P. & H. Bryant, S. Chapter 12 PubChem: Integrated Platform of Small Molecules and Biological Activities. *Annu. Rep. Comput. Chem.* **4,** 217–241 (2008).

151. Barnard, J. M. & Downs, G. M. Chemical Fragment Generation and Clustering Software. *J. Chem. Inf. Comput. Sci.* **37,** 141–142 (1997).

152. Sheridan, R. P., Miller, M. D., Underwood, D. J. & Kearsley, S. K. Chemical Similarity Using Geometric Atom Pair Descriptors. *J. Chem. Inf. Comput. Sci.* **36,** 128–136 (1996).

153. Gütlein, M. & Kramer, S. Filtered circular fingerprints improve either prediction or runtime performance while retaining interpretability. *J. Cheminformatics* **8,** 60 (2016).

154. Carhart, R. E., Smith, D. H. & Venkataraghavan, R. Atom pairs as molecular features in structure-activity studies: definition and applications. *J. Chem. Inf. Comput. Sci.* **25,** 64–73 (1985).

155. Nilakantan, R., Bauman, N., Dixon, J. S. & Venkataraghavan, R. Topological torsion: a new molecular descriptor for SAR applications. Comparison with other descriptors. *J. Chem. Inf. Comput. Sci.* **27,** 82–85 (1987).

156. Varnek, A., Wipff, G., Solov'e, V. P. & Solotnov, A. F. Assessment of the Macrocyclic Effect for the Complexation of Crown-Ethers with Alkali Cations Using the Substructural Molecular Fragments Method. *J. Chem. Inf. Comput. Sci.* **42,** 812–829 (2002).

157. Varnek, A. *et al.* ISIDA - Platform for Virtual Screening Based on Fragment and Pharmacophoric Descriptors. *Current Computer-Aided Drug Design* (2008). Available at: http://www.eurekaselect.com/67604/article. (Accessed: 30th April 2018)

158. *Nomenclature of ISIDA fragments (Fragmentor2015)*.

159. Morgan, H. L. The Generation of a Unique Machine Description for Chemical Structures-A Technique Developed at Chemical Abstracts Service. *J. Chem. Doc.* **5,** 107–113 (1965).

160. Riniker, S. & Landrum, G. A. Open-source platform to benchmark fingerprints for ligand-based virtual screening. *J. Cheminformatics* **5,** 26 (2013).

161. O'Boyle, N. M. & Sayle, R. A. Comparing structural fingerprints using a literature-based similarity benchmark. *J. Cheminformatics* **8,** 36 (2016).

162. Goodarzi, M., Dejaegher, B. & Heyden, Y. *Feature selection methods in QSAR studies*. **95,** (2012).

163. Palczewska, A., Neagu, D. & Ridley, M. Using Pareto points for model identification in predictive toxicology. *J. Cheminformatics* **5,** 16 (2013).

164. Khan, A. *Descriptors and their selection methods in QSAR analysis: Paradigm for drug design*. **21,** (2016).

165. FORGY, E. W. Cluster analysis of multivariate data : efficiency versus interpretability of classifications. *Biometrics* **21,** 768–769 (1965).

166. MacQueen, J. Some methods for classification and analysis of multivariate observations. in (The Regents of the University of California, 1967).

167. Bishop, C. M. *Pattern Recognition and Machine Learning*. (Springer, 2011).

168. Hastie, T., Tibshirani, R. & Friedman, J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition*. (Springer, 2016).

169. Krishna Menon, A., J Jiang, X., Vembu, S., Elkan, C. & Ohno-Machado, L. Predicting accurate probabilities with a ranking loss. *Proc. 29th Int. Conf. Mach. Learn. ICML 2012* **1,** (2012).

170. Maltarollo, V. G., Gertrudes, J. C., Oliveira, P. R. & Honorio, K. M. Applying machine learning techniques for ADME-Tox prediction: a review. *Expert Opin. Drug Metab. Toxicol.* **11,** 259–271 (2015).

171. Eriksson, L. *et al.* Methods for reliability and uncertainty assessment and for applicability evaluations of classification- and regression-based QSARs. *Environ. Health Perspect.* **111,** 1361–1375 (2003).

172. Rosipal, R. & Trejo, L. Kernel Partial Least Squares Regression in Reproducing Kernel Hilbert Space. *J. Mach. Learn. Res.* **2,** 97–123 (2001).

173. VAPNIK, V. A note one class of perceptrons. *Autom. Remote Control* (1964).

174. Boser, B. E., Guyon, I. M. & Vapnik, V. N. A Training Algorithm for Optimal Margin Classifiers. in *Proceedings of the Fifth Annual Workshop on Computational Learning Theory* 144–152 (ACM, 1992). doi:10.1145/130385.130401

175. Cortes, C. & Vapnik, V. Support-vector networks. *Mach. Learn.* **20,** 273–297 (1995).

176. Bennett, K. & Campbell, C. Support Vector Machines: Hype or Hallelujah? *ACM SIGKDD Explor. Newsl.* **2,** 1–13 (2000).

177. Barakat, N. & Bradley, A. Rule extraction from support vector machines: A review. *Neurocomputing* **74,** 178–190 (2010).

178. Smola, A. J. & Schölkopf, B. A tutorial on support vector regression. *Stat. Comput.* **14,** 199–222 (2004).

179. Jaworski, M., Duda, P. & Rutkowski, L. New Splitting Criteria for Decision Trees in Stationary Data Streams. *IEEE Trans. Neural Netw. Learn. Syst.* (2017). doi:10.1109/TNNLS.2017.2698204

180. Breiman, L. Bagging Predictors. *Mach. Learn.* **24,** 123–140 (1996).

181. Leach, A. R. & Gillet, V. J. *An Introduction to Chemoinformatics*. (Springer Netherlands, 2007).

182. Beliakov, G. & Li, G. Improving the speed and stability of the k-nearest neighbors method. *Pattern Recognit. Lett.* **33,** 1296–1301 (2012).

183. Hewitt, M. *et al.* Consensus QSAR Models: Do the Benefits Outweigh the Complexity? *J. Chem. Inf. Model.* **47,** 1460–8 (2007).

184. Sharma, N. & Yap, C. W. Consensus QSAR model for identifying novel H5N1 inhibitors. *Mol. Divers.* **16,** 513–524 (2012).

185. Dixon, S. L. *et al.* AutoQSAR: an automated machine learning tool for best-practice quantitative structure–activity relationship modeling. *Future Med. Chem.* **8,** 1825–1839 (2016).

186. Lever, J., Krzywinski, M. & Altman, N. Points of Significance: Model selection and overfitting. *Nature Methods* (2016). doi:10.1038/nmeth.3968

187. Topliss, J. G. & Costello, R. J. Chance correlations in structure-activity studies using multiple regression analysis. *J. Med. Chem.* **15,** 1066–1068 (1972).

188. Topliss, J. G. & Edwards, R. P. Chance factors in studies of quantitative structure-activity relationships. *J. Med. Chem.* **22,** 1238–1244 (1979).

189. Gramatica, P. & Sangion, A. A Historical Excursus on the Statistical Validation Parameters for QSAR Models: A Clarification Concerning Metrics and Terminology. *J. Chem. Inf. Model.* **56,** 1127–1131 (2016).

190. Roy, K., Kar, S. & Das, R. N. Chapter 7 - Validation of QSAR Models. in *Understanding the Basics of QSAR for Applications in Pharmaceutical Sciences and Risk Assessment* 231–289 (Academic Press, 2015). doi:10.1016/B978-0-12-801505-6.00007-7

191. Gramatica, P. Principles of QSAR models validation: internal and external. *QSAR Comb. Sci.* **26,** 694–701 (2007).

192. Kohavi, R. A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. **14,** (2001).

193. Kim, J.-H. Estimating Classification Error Rate: Repeated Cross-Validation, Repeated Hold-Out and Bootstrap." Computational Statistics & Data Analysis, 53(11), 3735-3745. *Comput. Stat. Data Anal.* **53,** 3735–3745 (2009).

194. Roy, K., Roy, P. P. & Joseph, T. On some aspects of validation of predictive QSAR models. *Chem. Cent. J.* **2,** (2008).

195. Golbraikh, A. & Tropsha, A. Predictive QSAR modeling based on diversity sampling of experimental datasets for the training and test set selection. *J. Comput. Aided Mol. Des.* **16,** 357–369 (2002).

196. Guha, R. & Jurs, P. C. Determining the Validity of a QSAR Model – A Classification Approach. *J. Chem. Inf. Model.* **45,** 65–73 (2005).

197. Leonard, J. T. & Roy, K. On Selection of Training and Test Sets for the Development of Predictive QSAR models. *QSAR Comb. Sci.* **25,** 235–251 (2006).

198. Martin, T. M. *et al.* Does Rational Selection of Training and Test Sets Improve the Outcome of QSAR Modeling? *J. Chem. Inf. Model.* **52,** 2570–2578 (2012).

199. Netzeva, T. *et al.* Current Status of Methods for Defining the Applicability Domain of (Quantitative) Structure-Activity Relationships. *ATLA* **33,** 155–173 (2005).

200. Ap, W. *et al. The Characterisation of (Quantitative) Structure-Activity Relationships: Preliminary Guidance*.

201. Jaworska, J., Nikolova-Jeliazkova, N. & Aldenberg, T. QSAR applicabilty domain estimation by projection of the training set descriptor space: a review. *Altern. Lab. Anim. ATLA* **33,** 445–459 (2005).

202. Preparata, F. P. & Shamos, M. I. Convex Hulls: Basic Algorithms. in *Computational Geometry* 95–149 (Springer, New York, NY, 1985). doi:10.1007/978-1-4612-1098-6_3

203. Houle, M. E., Kriegel, H.-P., Kröger, P., Schubert, E. & Zimek, A. Can Shared-Neighbor Distances Defeat the Curse of Dimensionality? in *Scientific and Statistical Database Management* 482–500 (Springer, Berlin, Heidelberg, 2010). doi:10.1007/978-3-642-13818-8_34

204. Beyer, K., Goldstein, J., Ramakrishnan, R. & Shaft, U. When Is "Nearest Neighbor" Meaningful? in *Database Theory — ICDT'99* 217–235 (Springer, Berlin, Heidelberg, 1999). doi:10.1007/3-540-49257-7_15

205. ORGANISATION FOR ECONOMIC CO-OPERATION AND DEVELOPMENT. OECD SERIES ON TESTING AND ASSESSMENT - THE REPORT FROM THE EXPERT GROUP ON (QUANTITATIVE) STRUCTURE-ACTIVITY RELATIONSHIPS [(Q)SARs] ON THE PRINCIPLES FOR THE VALIDATION OF (Q)SARs. **49,** (2004).

206. Weaver, S. & Gleeson, M. P. The importance of the domain of applicability in QSAR modeling. *J. Mol. Graph. Model.* **26,** 1315–1326 (2008).

207. Sahigara, F., Ballabio, D., Todeschini, R. & Consonni, V. Defining a novel k-nearest neighbours approach to assess the applicability domain of a QSAR model for reliable predictions. *J. Cheminformatics* **5,** 27 (2013).

208. Aniceto, N., Freitas, A. A., Bender, A. & Ghafourian, T. A novel applicability domain technique for mapping predictive reliability across the chemical space of a QSAR: reliability-density neighbourhood. *J. Cheminformatics* **8,** 69 (2016).

209. Sahigara, F. *et al.* Comparison of Different Approaches to Define the Applicability Domain of QSAR Models. *Mol. Basel Switz.* **17,** 4791–810 (2012).

210. Todeschini, R., Ballabio, D., Cassotti, M. & Consonni, V. N3 and BNN: Two New Similarity Based Classification Methods in Comparison with Other Classifiers. *J. Chem. Inf. Model.* **55,** 2365–2374 (2015).

211. Todeschini, R., Ballabio, D., Consonni, V. & Grisoni, F. A new concept of higher-order similarity and the role of distance/similarity measures in local classification methods. *Chemom. Intell. Lab. Syst.* **157,** (2016).

212. Kireeva, N. *et al.* Generative Topographic Mapping (GTM): Universal Tool for Data Visualization, Structure-Activity Modeling and Dataset Comparison. *Mol. Inform.* **31,** 301–312 (2012).

213. Gaspar, H. A., Baskin, I. I., Marcou, G., Horvath, D. & Varnek, A. GTM-Based QSAR Models and Their Applicability Domains. *Mol. Inform.* **34,** 348–356 (2015).

214. Gaspar, H. A., Baskin, I. I., Marcou, G., Horvath, D. & Varnek, A. Stargate GTM: Bridging Descriptor and Activity Spaces. *J. Chem. Inf. Model.* **55,** 2403–2410 (2015).

215. Lombardo, F. *et al.* In Silico Absorption, Distribution, Metabolism, Excretion, and Pharmacokinetics (ADME-PK): Utility and Best Practices. An Industry Perspective from the International Consortium for Innovation through Quality in Pharmaceutical Development. *J. Med. Chem.* **60,** 9097–9113 (2017).

216. Cronin, M. T. D. & Madden, J. C. Chapter 1:In Silico Toxicology—An Introduction. in *In Silico Toxicology* 1–10 (2010). doi:10.1039/9781849732093-00001

217. NIH. ChemIDplus. Available at: https://chem.nlm.nih.gov/chemidplus. (Accessed: 5th June 2018)

218. Wishart, D. *et al.* T3DB: the toxic exposome database. *Nucleic Acids Res.* **43,** D928-934 (2015).

219. Papadatos, G., Gaulton, A., Hersey, A. & Overington, J. P. Activity, assay and target data curation and quality in the ChEMBL database. *J. Comput. Aided Mol. Des.* **29,** 885–896 (2015).

220. Gaulton, A. *et al.* ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Res.* **40,** D1100-1107 (2012).

221. Bento, A. P. *et al.* The ChEMBL bioactivity database: an update. *Nucleic Acids Res.* **42,** D1083-1090 (2014).

222. Mutowo, P. The ChEMBL-og: ChEMBL tissues: Increasing depth, breadth and accuracy of annotations. *The ChEMBL-og* (2018).

223. spaCy · Industrial-strength Natural Language Processing in Python. Available at: https://spacy.io/. (Accessed: 6th June 2018)

224. Research, C. for D. E. and. Data Standards Manual (monographs) - Drug Nomenclature Monographs. Available at: https://www.fda.gov/Drugs/DevelopmentApprovalProcess/FormsSubmissionRequirements/ElectronicSubmissions/DataStandardsManualmonographs/ucm071650.htm. (Accessed: 6th June 2018)

225. Wishart, D. S. *et al.* DrugBank: a comprehensive resource for in silico drug discovery and exploration. *Nucleic Acids Res.* **34,** D668-672 (2006).

226. Fonger, G. C., Stroup, D., Thomas, P. L. & Wexler, P. TOXNET: A computerized collection of toxicological and environmental health information. *Toxicol. Ind. Health* **16,** 4–6 (2000).

227. PhysProp. (2018). Available at: http://esc.syrres.com/fatepointer/search.asp.

228. Kim, S. *et al.* PubChem Substance and Compound databases. *Nucleic Acids Res.* **44,** D1202-1213 (2016).

229. Obach, R. S., Lombardo, F. & Waters, N. J. Trend Analysis of a Database of Intravenous Pharmacokinetic Parameters in Humans for 670 Drug Compounds. *Drug Metab. Dispos.* **36,** 1385–1405 (2008).

230. ChemSpider. Available at: http://www.chemspider.com/. (Accessed: 10th June 2018)

231. Wikipédia. Available at: https://www.wikipedia.org. (Accessed: 10th June 2018)

232. NIH. Chemical Identifier Resolver. Available at: https://cactus.nci.nih.gov/chemical/structure. (Accessed: 10th June 2018)

233. Legehar, A., Xhaard, H. & Ghemtio, L. IDAAPM: integrated database of ADMET and adverse effects of predictive modeling based on FDA approved drug data. *J. Cheminformatics* **8,** 33 (2016).

234. Kornai, A. *Mathematical Linguistics*. (Springer, 2010).

235. Goldberg, Y. Neural Network Methods for Natural Language Processing. *Synth. Lect. Hum. Lang. Technol.* **10,** 1–309 (2017).

236. Kanji, S. *et al.* Reporting Guidelines for Clinical Pharmacokinetic Studies: The ClinPK Statement. *Clin. Pharmacokinet.* **54,** 783–795 (2015).

237. Lima, A. N. *et al.* Use of machine learning approaches for novel drug discovery. *Expert Opin. Drug Discov.* **11,** 225–239 (2016).

238. Wang, T., Yuan, X.-S., Wu, M.-B., Lin, J.-P. & Yang, L.-R. The advancement of multidimensional QSAR for novel drug discovery - where are we headed? *Expert Opin. Drug Discov.* **12,** 769–784 (2017).

239. Cumming, J. G., Davis, A. M., Muresan, S., Haeberlein, M. & Chen, H. Chemical predictive modelling to improve compound quality. *Nat. Rev. Drug Discov.* **12,** 948–962 (2013).

240. Sushko, I. *et al.* Online chemical modeling environment (OCHEM): web platform for data storage, model development and publishing of chemical information. *J. Comput. Aided Mol. Des.* **25,** 533–554 (2011).

241. Cox, R., Green, D. V. S., Luscombe, C. N., Malcolm, N. & Pickett, S. D. QSAR workbench: automating QSAR modeling to drive compound design. *J. Comput. Aided Mol. Des.* **27,** 321–336 (2013).

242. Stevenson, J. M. & Mulready, P. D. Pipeline Pilot 2.1 By Scitegic, 9665 Chesapeake Drive, Suite 401, San Diego, CA 92123-1365. www.scitegic.com. See Web Site for Pricing Information. *J. Am. Chem. Soc.* **125,** 1437–1438 (2003).

243. Wang, J. & Hou, T. Recent advances on aqueous solubility prediction. *Comb. Chem. High Throughput Screen.* **14,** 328–338 (2011).

244. Delaney, J. S. ESOL: estimating aqueous solubility directly from molecular structure. *J. Chem. Inf. Comput. Sci.* **44,** 1000–1005 (2004).

245. Wang, J., Hou, T. & Xu, X. Aqueous solubility prediction based on weighted atom type counts and solvent accessible surface areas. *J. Chem. Inf. Model.* **49,** 571–581 (2009).

246. Huuskonen, J. Estimation of Aqueous Solubility for a Diverse Set of Organic Compounds Based on Molecular Topology. *J. Chem. Inf. Comput. Sci.* **40,** 773–777 (2000).

247. Wang, J. *et al.* Development of reliable aqueous solubility models and their application in druglike analysis. *J. Chem. Inf. Model.* **47,** 1395–1404 (2007).

248. Tetko, I. V., Maran, U. & Tropsha, A. Public (Q)SAR Services, Integrated Modeling Environments, and Model Repositories on the Web: State of the Art and Perspectives for Future Development. *Mol. Inform.* **36,** (2017).

249. Swain, M. *MolVS: Molecule Validation and Standardization*.

250. Halgren, T. A. MMFF VI. MMFF94s option for energy minimization studies. *J. Comput. Chem.* **20,** 720–729 (1999).

251. ChemAxon – Software for Chemistry and Biology. *ChemAxon - Software for Chemistry and Biology R&D* Available at: https://www.chemaxon.com/. (Accessed: 18th May 2017)

252. Gally, J.-M., Bourg, S., Do, Q.-T., Aci ‐ Sèche, S. & Bonnet, P. VSPrep: A General KNIME Workflow for the Preparation of Molecules for Virtual Screening. *Mol. Inform.* **36,** 1700023 (2017).

253. Kuhn, M. & Johnson, K. *Applied Predictive Modeling*. (Springer-Verlag, 2013).

254. Lafi, S. Q. & Kaneene, J. B. An explanation of the use of principal-components analysis to detect and correct for multicollinearity. *Prev. Vet. Med.* **13,** 261–275 (1992).

255. Jorgensen, W. L. & Duffy, E. M. Prediction of drug solubility from structure. *Adv. Drug Deliv. Rev.* **54,** 355–366 (2002).

256. Pedregosa, F. *et al.* Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **12,** 2825–2830 (2011).

257. Ensemble Partial Least Squares Regression. Available at: https://nanx.me/enpls/index.html. (Accessed: 2nd July 2018)

258. Cao, D.-S., Liang, Y.-Z., Xu, Q.-S., Li, H.-D. & Chen, X. A new strategy of outlier detection for QSAR/QSPR. *J. Comput. Chem.* **31,** 592–602 (2010).

259. Nelder, J. A. & Mead, R. A Simplex Method for Function Minimization. *Comput. J.* **7,** 308–313 (1965).

260. Zhang, F., Xue, J., Shao, J. & Jia, L. Compilation of 222 drugs' plasma protein binding data and guidance for study designs. *Drug Discov. Today* **17,** 475–485 (2012).

261. Ghuman, J. *et al.* Structural basis of the drug-binding specificity of human serum albumin. *J. Mol. Biol.* **353,** 38–52 (2005).

262. Kuz'min, V. E., Polishchuk, P. G., Artemenko, A. G. & Andronati, S. A. Interpretation of QSAR Models Based on Random Forest Methods. *Mol. Inform.* **30,** 593–603 (2011).

263. Balfer, J. & Bajorath, J. Visualization and Interpretation of Support Vector Machine Activity Predictions. *J. Chem. Inf. Model.* **55,** 1136–1147 (2015).

264. Marchese Robinson, R. L., Palczewska, A., Palczewski, J. & Kidley, N. Comparison of the Predictive Performance and Interpretability of Random Forest and Linear Models on Benchmark Data Sets. *J. Chem. Inf. Model.* **57,** 1773–1792 (2017).

265. Bemis, G. W. & Murcko, M. A. The properties of known drugs. 1. Molecular frameworks. *J. Med. Chem.* **39,** 2887–2893 (1996).

266. Degen, J., Wegscheid-Gerlach, C., Zaliani, A. & Rarey, M. On the art of compiling and using 'drug-like' chemical fragment spaces. *ChemMedChem* **3,** 1503–1507 (2008).

267. Lewell, X. Q., Judd, D. B., Watson, S. P. & Hann, M. M. RECAP--retrosynthetic combinatorial analysis procedure: a powerful new technique for identifying privileged molecular fragments with useful applications in combinatorial chemistry. *J. Chem. Inf. Comput. Sci.* **38,** 511–522 (1998).

268. Riniker, S. & Landrum, G. A. Similarity maps - a visualization strategy for molecular fingerprints and machine-learning methods. *J. Cheminformatics* **5,** 43 (2013).

269. Rosenbaum, L., Hinselmann, G., Jahn, A. & Zell, A. Interpreting linear support vector machine models with heat map molecule coloring. *J. Cheminformatics* **3,** 11 (2011).

270. Ritchie, T. J., Ertl, P. & Lewis, R. The graphical representation of ADME-related molecule properties for medicinal chemists. *Drug Discov. Today* **16,** 65–72 (2011).

271. Millot, G. *Comprendre et réaliser les tests statistiques à l'aide de R: manuel de biostatistique*. (De Boeck Supérieur, 2014).

Communications scientifiques

PUBLICATION

Canault, Bourg, Vayer et Bonnet
Comprehensive Network Map of ADME-tox Databases
Mol. Inf. (2017), 36, 1700029

Khater, Canault, Azzimani, Bonnet et West
Thermodynamic enantioseparation behavior of phenylthiohydantoin-amino acid derivatives in supercritical fluid chromatography on polysaccharide chiral stationary phases
J. Sep. Sci. (2018), 41 (6), 1450-1459

COMMUNICATIONS ORALES

Canault, Bourg, Vayer, Bonnet
Metapredict : une plate-forme in silico de prédiction ADME-tox
8es journées de la Société Française de Chémoinformatique - SFCi
Octobre 2017 – Orléans.

Canault, Bourg, Aucagne, Loth et Bonnet
Conception rationnelle d'analogues de Kisspeptine
Journée de la Fédération de Recherche PCV FR2708
Janvier 2016 – Orléans.

COMMUNICATIONS PAR POSTER

Canault, Bourg, Vayer, Bonnet
MetaPredict: an in silico platform for ADME-Tox prediction
21st European Symposium on Quantitative Structure-Activity Relationship - EuroQSAR
sept. 2016 - Verona (Italy).

Canault, Gally, Braka, Bourg, Aci-Sèche, Bonnet
Evaluation of rDock docking tool for kinase drug discovery
Journée Jeunes Chercheurs SCT
févr. 2016 - Lille.

Canault, Bourg, Vayer, Bonnet
Network map of integrated ADME-Tox databases
Journées nationales de la Société Française de Chémoinformatique (SFCI)
oct. 2015 - Nice.

Pihan, Canault, Arora, Golib, Bosc, Bourg, Bonnet
Large scale kinase virtual screening plateform
50th International Conference on Medicinal Chemistry (RICT 2014)
juil. 2014 - Rouen.

Gally, Bosc, Pihan, Arora, Bourg, Canault, Bonnet
Development of a universal workflow for the preparation of molecular databases for virtual screening
Chemoinformatics Strasbourg Summer School (CS3)
juin 2014 - Strasbourg.

PRIX

Canault, Bourg, Vayer, Bonnet
Metapredict : une plate-forme in silico de prédiction ADME-tox
8es journées de la Société Française de Chémoinformatique - SFCi
Octobre 2017 – Orléans.

Canault, Gally, Braka, Bourg, Aci-Sèche, Bonnet
Evaluation of rDock docking tool for kinase drug discovery
Journée Jeunes Chercheurs SCT
févr. 2016 - Lille.

Baptiste CANAULT

Développement d'une plateforme de prédiction *in silico* des propriétés ADME-Tox

Résumé :

Dans le cadre de la recherche pharmaceutique, les propriétés relatives à l'Absorption, la Distribution, le Métabolisme, l'Elimination (ADME) et la Toxicité (Tox) sont cruciales pour le succès des phases cliniques lors de la conception de nouveaux médicaments. Durant ce processus, la chémoinformatique est régulièrement utilisée afin de prédire le profil ADME-Tox des molécules bioactives et d'améliorer leurs propriétés pharmacocinétiques. Ces modèles de prédiction, basés sur la quantification des relations structure-activité (QSAR), ne sont pas toujours efficaces à cause du faible nombre de données ADME-Tox disponibles et de leur hétérogénéité induite par des différences dans les protocoles expérimentaux, ou encore de certaines erreurs expérimentales. Au cours de cette thèse, nous avons d'abord constitué une base de données contenant 150 000 mesures pour une cinquantaine de propriétés ADME-Tox. Afin de valoriser l'ensemble de ces données, nous avons dans un deuxième temps proposé une plateforme automatique de création de modèles de prédiction QSAR. Cette plateforme, nommée MetaPredict, a été conçue afin d'optimiser chacune des étapes de création d'un modèle statistique, dans le but d'améliorer leur qualité et leur robustesse. Nous avons dans un troisième temps valorisé les modèles obtenus grâce à la plateforme MetaPredict en proposant une application en ligne. Cette application a été développée pour faciliter l'utilisation des modèles, apporter une interprétation simplifiée des résultats et moduler les observations obtenues en fonction des spécificités d'un projet de recherche. Finalement, MetaPredict permet de rendre les modèles ADME-Tox accessibles à l'ensemble des chercheurs.

Mots clés : chémoinformatique, ADME-Tox, QSAR, molécules bioactives

Development of an *in silico* platform for ADME-Tox prediction

Summary :

Absorption, Distribution, Metabolism, Elimination (ADME) and Toxicity (Tox) properties are crucial for the success of clinical trials of a drug candidate. During this process, chemoinformatics is regularly used to predict the ADME-Tox profile of bioactive compounds and to improve their pharmacokinetic properties. *In silico* approaches have already been developed to improve poor pharmacokinetics and toxicity of lead compounds. These predictive models, based on the quantification of structure-activity relationships (QSAR), were not always efficient enough due to the low number of accessible biological data and their heterogeneity induced by the differences in experimental assays or the significant experimental error. In this thesis, we first built a database containing 150,000 data points for about 50 ADME-Tox properties. In order to valorize all this data, we then proposed an automatic platform for creating predictive models. This platform, called MetaPredict, has been designed to optimize each step of model development, in order to improve their quality and robustness. Third,, we promoted the statistical models using the online application of MetaPredict platform. This application has been developed to facilitate the use of newly built models, to provide a simplified interpretation of the results and to modulate the obtained observations according to the needs of the researchers. Finally, this platform provides an easy access to the ADME-Tox models for the scientific community.

Keywords : chemoinformatics, ADME-Tox, QSAR, bioactive molecules

Institut de Chimie Organique et Analytique UMR CNRS-Université d'Orléans 7311 Université d'Orléans
Rue de Chartres
45067 Orléans