الجمهورية الجزائرية الديموقراطية الشعبية
La république algérienne démocratique et populaire
وزارة التعليم العالي و البحث العلمي
Ministère de l'enseignement supérieur et de la recherche scientifique

UNIVERSITE BADJI MOKHTAR-ANNABA
BADJI MOKHTAR UNIVERSITY-ANNABA

جامعة باجي مختار- عنابة -
Année 2019-2020

Faculté des Sciences
Département de Chimie

THÈSE

Pour l'obtention du diplôme de doctorat

Option : Chimie de l'environnement et QSAR

THÈME

Investigations sur la rétention chromatographique des composants d'huiles essentielles par approche QSRR hybride

Présenté par : Mr. DRIOUCHE Youssouf

Devant le jury composé de :

Président :	**Mr. DJELLAL Ahmed**	**Professeur, Université d'Annaba**
Directeur de thèse :	**Mr. MESSADI Djelloul**	**Professeur, Université d'Annaba**
Examinateur :	**Mr. MERDES Rachid**	**Professeur, Université de Guelma**
Examinateur :	**Mr. DJAFER Rachid**	**Professeur, Université d'Annaba**
Examinateur :	**Mr. ZENATI Nour Eddine**	**MCA, Université de Souk Ahras**

DEDICACES

Je dédie ce Travail :

A ceux que j'ai de plus chers au monde : mon père et ma mère

En gratitude de leur soutien, leur amour et leur présence continus

Tout au long de mes études et de ma vie.

- que Dieu vous garde -

A mes chers frères

A mes chères sœurs

A ma Fiancée

A tous les membres de ma famille

Je dédie ce modeste travail à toute l'équipe du labo LASEA

À tous mes amis.

DRIOUCHE YOUSSOUF

REMERCIEMENTS

Nous tenons tout d'abord à louer Dieu qui nous a permis d'accomplir ce travail.

Ce travail a été réalisé au *L*aboratoire de *S*écurité *E*nvironnementale et *A*limentaire (*LASEA*), de l'université Badji Mokhtar Annaba. Mes premiers remerciements sont adressés à mon directeur de thèse, le professeur **MESSADI Djelloul**. D'abord, pour m'avoir accueilli dans son équipe et m'avoir permis de travailler sur un sujet aussi passionnant. Ensuite pour m'avoir orienté et encadré avec efficacité tout en me laissant l'initiative et pour ses multiples conseils durant toute la durée qu'a nécessité la réalisation de ce projet de recherche. Enfin je ne le remercierai jamais assez pour la confiance qu'il m'a accordée et surtout pour sa grande patience.

J'exprime également ma profonde gratitude à monsieur **AHMED Djellal** professeur à l'université Badji Mokhtar Annaba, pour nous avoir fait l'honneur de présider le jury de cette thèse.

Mes vifs remerciements vont également à monsieur **MERDES Rachid** professeur à l'université 8 mai 45 Guelma, à monsieur **DJAFER Rachid** professeur à l'université Badji Mokhtar Annaba, et à monsieur **ZENATI Nour Eddine** maître de conférences à l'université de Souk Ahras, pour l'honneur qu'ils nous ont fait en acceptant d'examiner notre travail.

C'est avec beaucoup de gratitude enfin que je remercie tous les membres de l'équipe LASEA pour leur soutien, leur amitié, leur aide. J'ai eu beaucoup de plaisir à partager de bons moments à leurs côtés.

Un grand merci à mes chers amis d'avoir partagé avec moi d'agréables moments. Je tiens à présenter ma reconnaissance et mes remerciements à ma famille, qui est ma source d'inspiration et mon plus grand soutien.

DRIOUCHE YOUSSOUF

Résumé

Le travail présenté dans cette thèse a pour objectifs d'élaborer des modèles QSRR fiables, stables et prédictifs pour la prédiction de l'indice de rétention d'une sérié de composants de l'huile essentielle de Thymus vulgaris. L'ensemble de 36 composés de l'huile essentielle a été divisé en un ensemble de calibrage de 27 composés et un ensemble de validation de 9 composés selon l'algorithme de Kennard et stone (CADEX).

Les modèles QSRR sont développés à l'aide de ces techniques : la régression linéaire multiple (MLR), les réseaux de neurones artificiels (RNA), et les machines à vecteurs supports (SVM).

La stabilité interne des modèles élaborés a été confirmée par plusieurs techniques, à savoir, la validation croisée et la technique dite du Leave-one-out et la Y-Randomisation. Le pouvoir prédictif a été testé avec succès avec une série de test non incluse dans la série d'apprentissage.

Un modèle à quatre descripteurs, avec les valeurs de ces paramètres statistiques (R^2, Q^2, EQMC, EQMP, $EQMP_{ext}$) obtenues attestent de la pertinence des modèles développés, avec une supériorité établie le modèle linéaire (MLR).

Mots clés : huiles essentielles - indice de rétention - QSRR - régression linéaire multiple - supports vecteurs machines - réseaux de neurones artificiels.

Abstract

The work presented in this thesis aims to develop reliable, stable and predictive QSRR models for predicting the retention index of a series of essential oils. The set of 36 essential oil compounds was divided into a training set of 27 compounds and a test set of 9 compounds according to Kennard and stone algorithm.

QSRR models are developed using these techniques: Multiple Linear Regression (MLR), Artificial Neural Networks (ANN), and Support Vector Machines (SVM).

The internal stability of the developed models was verified by several techniques, namely, the cross-validation, the Leave-one-out, and the Y-Randomization. The predictive ability was further performed by using the external test set composed of data not used to develop prediction 'training set'.

A four descriptors model, with the values of these statistical parameters (R^2, Q^2, SDEC, SDEP, SDEPext) obtained attest relevance of the models developed, with a superiority established the linear model (MLR).

Key Words: essential oils - retention index - QSRR- multiple linear regression - support vectors machines - artificial neural networks

الملخص

العمل المقدم في هذه الأطروحة يهدف الى تطوير نماذج QSRR موثوقة ومستقرة وتنبؤية للتنبؤ بمؤشر الاحتفاظ بسلسلة من الزيوت الأساسية. تم تقسيم مجموعة مكونة من 36 مركبًا من الزيوت الأساسية إلى مجموعة معايرة مكونة من 27 مركبًا ومجموعة تحقق من 9 مركبات وفقًا لخوارزمية Kennard et stone (CADEX).

تم تطوير نماذج QSRR باستخدام هذه التقنيات: الانحدار الخطي المتعدد (MLR)، والشبكات العصبية الاصطناعية (ANN)، وشعاع الدعم الالي (SVM).

وللتحقق من الاستقرار الداخلي من النماذج تم استخدام العديد من التقنيات، وهي Leave-one-out وcross-validation، Y-Randomization كما تم اختبار القدرة التنبؤية بنجاح باستخدام مجموعة اختبار خارجي التي لم تدرج في السلسلة الأصلية المستخدمة لتطوير النموذج "مجموعة التدريب".

نموذج مع أربع واصفات، مع قيم هذه المعلمات الإحصائية (R^2، Q^2، EQMC، EQMP، $EQMP_{ext}$) التي تم الحصول عليها تشهد على أهمية النماذج التي تم تطويرها، مع التفوق الذي أنشأ النموذج الخطي (MLR).

الكلمات الدالة: الزيوت الأساسية - مؤشر الاحتفاظ - *QSRR* - الانحدار الخطي المتعدد - شعاع الدعم الالي - الشبكات العصبية الاصطناعية.

SOMMAIRE

Deuxième Partie : Application

SYMBOLES ET ABREVIATIONS

a, b : Paramètres de la droite des moindres carrés.

AG : Algorithme génétique.

CPG : Chromatographie en phase gazeuse.

DFT : Théorie de la fonctionnelle de la densité.

DIF : Détecteur à ionisation de flamme.

e_i : Résidu : différence entre les valeurs observée (y_i) et estimée ($\hat{\mathbf{y}}_\mathbf{i}$).

ei_std : Résidu standardisé.

EQM : Ecart quadratique moyen.

EQMC : Ecart quadratique moyen calculé sur l'ensemble de calibrage.

EQMP : Ecart quadratique moyen de prédiction.

$EQMP_{ext}$: Ecart quadratique moyen calculé sur l'ensemble de validation externe.

F : Statistique de Fisher.

FIT : Fonction de KUBINYI.

FIV : Facteur d'inflation de la variance.

H E C T : Huile essentielle chémotypée.

HF : Hartree -Fock.

h_{ii} : Eléments diagonaux de la matrice chapeau.

IE : Ionisation électronique.

Ir: Indice de rétention.

LMO : Cross-validation by leave-many-out : Validation croisée par omission d'un ensemble d'observations.

log Ir: Logarithme décimal de l'indice de rétention.

LOO : Cross-validation by leave-one-out : Validation croisée par omission d'une observation.

MM : Mécanique moléculaire.

MQ : Mécanique quantique.

MVS (SVM) : Machine à vecteur support (Support vector machine).

N : Dimension de la population (échantillon).

n-p : Nombre de degrés de liberté..

p : Nombre de descripteurs en comptant la constante (Nombre de paramètres).

PRESS : Somme des carrés des erreurs de prédiction.

Q^2_{boot} : Coefficient de prédiction par la technique du bootstrap.

Q^2_{LOO} : Coefficient de prédiction.

QSAR : Relations Quantitatives Structure-Activité.

QSPR : Relations Quantitatives Structure-Propriété.

QSRR : Relations Quantitatives Structure-Rétention.

QSTR : Relations Quantitatives Structure-Toxicité.

R^2 : Coefficient de détermination.

r_i : Résidu studentisé interne.

RLM (MLR) : Régression linéaire multiple.

RMN : Résonance magnétique nucléaire.

RMSE : Racine de l'écart quadratique moyen (Root Mean Squared Error).

RNA : Réseaux de neurones artificiels.

s : Erreur standard.

SCE : Somme des carrés des écarts.

SCT : Somme des carrés totale.

SM : Spectrométrie de masse.

S_{xx} : Somme des carrés des écarts par rapport à la moyenne x.

S_{xy} : Somme des produits rectangles des écarts par rapport à la moyenne x et y.

S_{yy} : Somme des carrés des écarts par rapport à la moyenne y.

t : t de Student.

t_i : Résidu studentisé externe.

T_x : Moyenne arithmétique de la grandeur x.

T_y : Moyenne arithmétique de la grandeur y.

V(x) : Variance de la grandeur x.

$\sqrt{V(y)}$: Erreur standard sur y.

y_i : Valeur observée.

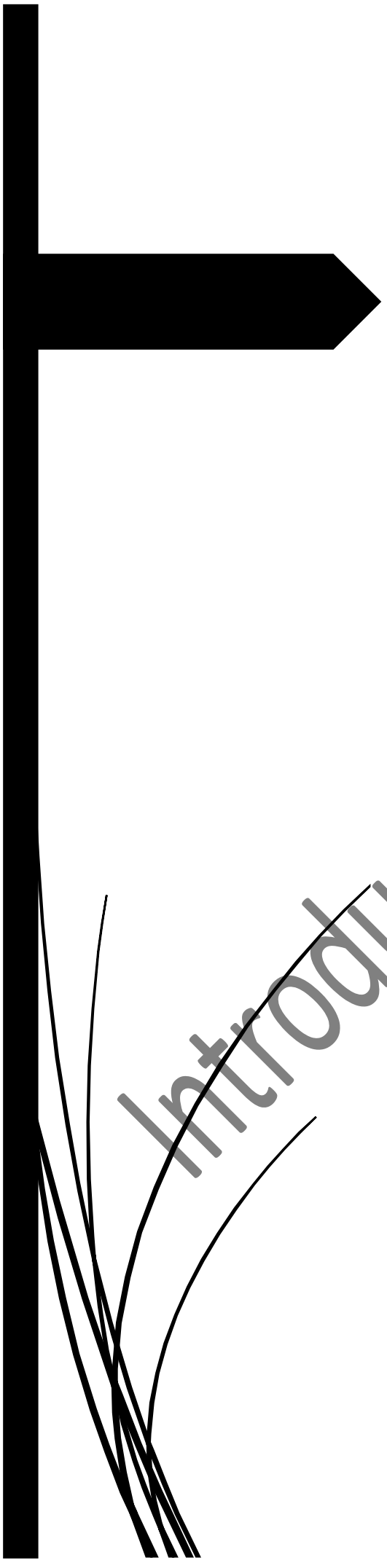
$\hat{y}_{(i)}$: Valeur prédite.

$\hat{y}_i$: Valeur estimée.

Liste des tableaux

Liste des figures

Introduction Générale



Depuis toujours, l'Homme a eu recours aux plantes pour se maquiller, se parfumer, mais aussi pour se soigner sans connaître réellement les propriétés de ces plantes, ni avoir la moindre connaissance scientifique, même sommaire, expliquant leurs vertus. Ce n'est qu'au moyen âge que les huiles essentielles ont été réellement découvertes grâce aux premières distillations et plus tard, grâce aux progrès de la science et tout particulièrement à l'apparition de la chimie. Cette médecine traditionnelle ancestrale est le précurseur de la phytothérapie et de l'aromathérapie d'aujourd'hui **[1]**.

Les huiles essentielles sont des produits de composition complexe, renfermant des substances volatils contenus dans les végétaux obtenus à partir d'une matière première végétale : fleur, feuille, bois, racine, écorce, fruit, ou autre ; soit par entraînement à la vapeur d'eau, soit par extraction mécanique. Le principal procédé d'extraction est la distillation à la vapeur d'eau. Les huiles essentielles sont un assemblage de molécules complexes qui ont toutes des propriétés particulières.

La modélisation moléculaire est largement utilisée **[2]** pour élaborer des modèles fiables permettant de prédire les propriétés physico-chimiques et les activités biologiques. L'une de ces techniques est la modélisation QSAR/QSPR/QSRR (Quantitative Structure-Activity/Property/Retention Relationships) **[3]** qui permet de prédire les propriétés/activités des systèmes chimiques à partir de leurs structures moléculaires.

Dans de nombreux domaines, la prédiction des propriétés physico-chimiques ou activités biologiques des molécules présente un enjeu industriel important car elle permet de réduire les délais et les coûts de productions.

Les méthodes QSAR/QSPR/QSRR sont communément employées dans la littérature et représentent un sous-domaine important de la chemo-informatique. Ces techniques servent à prédire plusieurs propriétés et activités biologiques **[4]** tels que :

a) Des propriétés physico-chimiques :

Prédiction de la solubilité aqueuse, points de fusion, températures d'ébullition, températures critiques, indices de rétention, indices de réfraction, viscosité, coefficient de partage octanol-eau.

b) Des activités biologiques :

Prédictions de la toxicité, prédictions de l'activité anti-inflammatoire, anti-cancer, anti-microbien, antibactérienne, … de différentes familles de composés chimiques.

L'objectif principal de ce travail est d'appliquer la méthodologie QSRR pour développer des modèles fiables pour prédire l'indice de rétention des composants d'huiles essentielles. Nous avons utilisé l'approche QSRR hybride associant algorithme génétique pour la sélection de sous-ensembles de variables significatives parmi quelques 2000 calculées théoriquement, et soit une régression linéaire (MLR), soit une régression non linéaire (RNA, SVM).

Le travail présenté dans ce manuscrit est divisé en deux grandes parties :

- Première partie : Etude bibliographique, avec :
  - Un aperçu général sur les huiles essentielles
  - Aspects théoriques de la modélisation moléculaire
  - Principes et méthodes de modélisations
- Deuxième partie : Application, aux :
  - Résultats obtenus et leurs discussions

Enfin, nous clôturons ce travail par une conclusion générale.

Première partie : Etude bibliographique

I. Les huiles essentielles

I. 1. Historique

I. 2. Définition

I. 3. Intérêt thérapeutique, écologique et économique des huiles essentielles

I. 4. Localisation et structure histologique des huiles essentielles

I. 5. Composition chimique des huiles essentielles

I. 6. Facteurs influençant la composition chimique

I. 7. Les méthodes d'extraction des huiles essentielles

I. 8. Les méthodes d'analyse chimique des huiles essentielles

I. 9. Utilisation des huiles essentielles

I. 10. Les principales propriétés des huiles essentielles

I. 11. Les principales voies d'utilisation des huiles essentielles

I. 12. Toxicité des huiles essentielles

I. 13. Conservation des huiles essentielles

I. Les huiles essentielles :

I. 1. Historique :

De tout temps, le règne végétal a offert à l'homme des ressources naturelles pour son alimentation, son hygiène et sa santé.

Depuis les temps les plus anciens, les parfums de ces mêmes végétaux sont associés à des rites mystiques, artistiques et esthétiques.

Déjà, En Chine, l'empereur Chen Nong (2800 av. J.-C.), médecin érudit, consigne son savoir relatif aux plantes médicinales dans un livre, le Pen Ts'ao, qui recense plus de 1000 plantes médicinales utiles **[5]**.

Il semble que ce sont les Egyptiens, dont l'histoire remonte à plus de 4000 ans qui furent les premiers à tirer parti du règne végétal dans un souci esthétique et spirituel. L'essence de térébenthine était déjà utilisée et tout porte à penser que certains parfums étaient déjà obtenus sous forme d'huile distillée.

Plus tard la civilisation arabe dont Baghdad, Bassora et Damas étaient les principaux centres commerciaux, développa le commerce des épices et des aromates et donna une grande impulsion à l'art de distillation.

C'est Gerber (721-815) qui mentionna le premier de façon écrite la description de la distillation.

L'alambic est incontestablement associé à Avicenne (980-1037), tout comme le vase florentin est associé à Giovanni Baptista della Porta (1535-1615), dont le célèbre ouvrage « De destillatione » parut en 1567, mentionne les connaissances avancées des Arabes dans le domaine de la distillation.

Hermann Boerhave (1668-1738) fut l'un des premiers à décrire les huiles essentielles d'un point de vue chimique **[6]**.

L'aromathérapie tomba ensuite dans l'oubli, et il a fallu attendre le XXème siècle pour que les scientifiques recommencent à s'y intéresser. En 1928 René-Maurice Gatte Fossé « chimiste français » publia un ouvrage « aromathérapie » décrivant la relation entre la structure biochimique de l'huile essentielle et son activité antimicrobienne.

En 1929, Sevelinge un pharmacien en France, étudia les huiles essentielles en médecine vétérinaire et confirma le potentiel antimicrobien élevé de ces substances aromatiques **[7]**.

En 1975, Franchomme en France, aromatologue, mit en évidence l'importance du chémotype (ou race chimique de l'espèce) **[8]**.

L'ère industrielle, après un début empirique, développa de nouvelles techniques de distillation.

Il existe aujourd'hui approximativement 3000 huiles dont environ 300 sont réellement commercialisées, destinées principalement à l'industrie des arômes et des parfums. Mais la tendance actuelle des consommateurs à rechercher une alimentation plus naturelle a entraîné un regain d'intérêt des scientifiques pour ces substances. Depuis deux décennies des études ont été menées sur le développement de nouvelles applications.

I. 2. Définition :

Plusieurs définitions disponibles d'une huile essentielle convergentes sur le fait que les huiles essentielles, communément appelées « essences végétales », sont des produits de composition généralement assez complexe, renfermant les principaux odorants volatils contenus dans les différentes parties des végétaux **[9]**. Elles diffèrent des huiles fixes (huile d'olive, ...) et des graisses végétales par leur caractère volatil ainsi que leur composition chimique **[10]**.

Plus récemment, la norme **[11]** a donné la définition suivante d'une huile essentielle : produit obtenu à partir d'une matière première végétale soit par entraînement à la vapeur soit par procédés mécaniques à partir de l'épicarpe des citrus soit par distillation sèche. L'huile essentielle est ensuite séparée de la phase aqueuse par des procédés physiques pour les deux premiers modes d'obtention. Elle peut subir des traitements physiques n'entraînant pas de changement significatif de sa composition **[12]**.

I. 3. Intérêt thérapeutique, écologique et économique des huiles essentielles :

Les huiles essentielles possèdent de nombreuses activités biologiques. En phytothérapie, elles sont utilisées pour leurs propriétés antiseptiques contre les maladies infectieuses d'origine bactérienne (par exemple contre les bactéries endocanalaires **[13]** ou au niveau de la microflore vaginale **[14]**) ou d'origine fongique contre les dermatophytes **[15]**. Cependant, elles possèdent également des propriétés cytotoxiques **[16]** qui les rapprochent donc des antiseptiques et désinfectants en tant qu'agents antimicrobiens à large spectre. Dans les domaines phytosanitaire et agroalimentaire, les huiles essentielles ou leurs composés actifs pourraient également être employés comme agents de protection contre les champignons phytopathogènes **[17]** et les microorganismes envahissant les denrées alimentaires **[18]**.

Les huiles essentielles jouent un rôle écologique dans les interactions végétales, végétale-animales et pourraient même constituer des supports de communication par des transferts de messages biologiques sélectifs **[19]**. En effet, elles contribuent à l'équilibre des écosystèmes, attirent les abeilles et des insectes responsables de la pollinisation, protègent les végétaux contre les herbivores et les rongeurs, possèdent des propriétés antifongiques, antibactériennes, allopathiques dans les régions arides et peuvent servir de solvants bioactifs des composés lipophiles **[20, 21]**.

Traditionnellement, les huiles essentielles sont présentes dans le processus de fabrication de nombreux produits finis destinés aux consommateurs. Ainsi, elles sont utilisées dans l'agroalimentaire (gâteaux, biscuits, soupe, sauce, chewing gum, chocolats, bonbons…) pour aromatiser la nourriture. Elles sont également utilisées dans l'industrie de la parfumerie, de la cosmétique et de la savonnerie. On les utilise aussi dans la fabrication des adhésifs (colle, scotch …), et celle de la nourriture pour animaux, dans l'industrie automobile, dans la préparation des sprays insecticides. L'homéopathie et l'aromathérapie sont des exemples courants d'usage d'huiles essentielles en médecine douce, et leur popularité s'est accrue d'une façon considérable ces dernières années **[22]**.

I. 4. Localisation et structure histologique des huiles essentielles :

Toutes les parties des plantes aromatiques peuvent contenir de l'huile essentielle **[23]**.

- Les fleurs bien sûr, exemples : orange, rose, lavande ; le bouton floral (girofle) ou les bractées (ylang-ylang "arbre de la famille des Annonacées, originaire d'Asie du Sud-Est'').
- Les feuilles le plus souvent, exemples : eucalyptus, menthe, thym, laurier, sarriette, sauge, aiguilles de pin, de sapin.
- Les organes souterrains, exemples : racines (vétiver), rhizomes (gingembre, acore).
- Les fruits, exemples : fenouil, anis, épicarpes des citrus.
- Les graines : noix de muscade, coriandre.
- Le bois et les écorces, exemple : cannelle, santal, bois de rose.

Les huiles essentielles sont produites par diverses structures spécialement différenciées dont le nombre et les caractéristiques sont très variables.

- Les poils sécréteurs épidermiques rencontrés souvent chez les Lamiacées, Géraniacées et Verbénacées. Ils produisent les essences dites superficielles.

- Les organes sécréteurs sous-cutanés comprenant des cellules et des poches sécrétrices qui sont généralement disséminées au sein du tissu végétal chez les Myrtacées, Rutacées, ainsi que des canaux sécréteurs chez les Apiacées.

La composition chimique d'une huile essentielle peut varier considérablement :

- Dans une même plante selon les organes (feuille, fleur, fruit, bois).
- Dans l'année selon la saison pour une même plante.
- Selon les conditions de culture pour une même espèce végétale (ensoleillement, humidité, longueur du jour, fertilité du sol).
- Selon les races chimiques (ou chémotypes) pour une même espèce (l'exemple classique est le thym avec 7 races chimiques).

I. 5. Composition chimique des huiles essentielles :

Dans les plantes, les huiles essentielles n'existent quasiment que chez les végétaux supérieurs. Elles sont produites dans le cytoplasme des cellules sécrétrices et s'accumulent en général dans des cellules glandulaires spécialisées, situées en surface de la cellule et recouvertes d'une cuticule. Elles peuvent être stockées dans divers organes : fleurs, feuilles, écorces, bois, racines, rhizomes, fruits ou graines **[24]**.

Les huiles essentielles sont constituées principalement de deux groupes de composés odorants distincts selon la voie métabolique empruntée ou utilisée. Il s'agit des terpènes (mono et sesquiterpènes), prépondérants dans la plupart des essences, et des composés aromatiques dérivés du phénylpropane **[25]**.

I. 5. 1. Les monoterpènes :

Les monoterpènes (Figure I. 1) sont les plus simples constituants des terpènes dont la majorité est rencontrée dans les huiles essentielles (90%) **[26]**. Ils comportent deux unités isoprène (C_5H_8), selon le mode de couplage « tête-queue ». Ils peuvent être acycliques, monocycliques ou bicycliques. A ces terpènes se rattachent un certain nombre de produits naturels à fonctions chimiques spéciales.

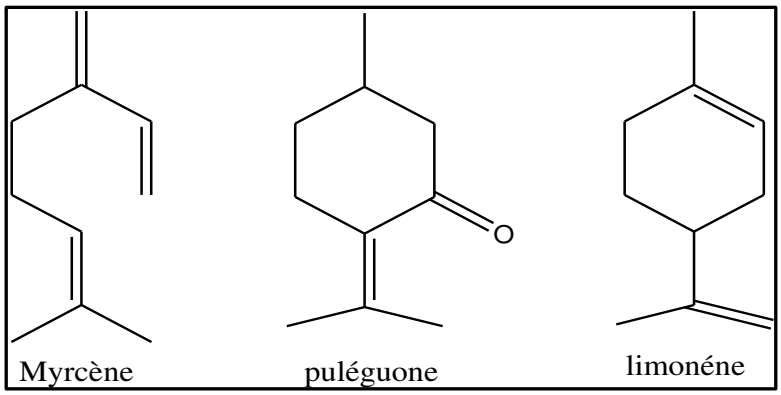


Figure I. 1 : Exemples de quelques monoterpènes.

I. 5. 2. Les sesquiterpènes :

Ce sont des dérivés d'hydrocarbures en $C_{15}H_{22}$. Il s'agit de la classe la plus diversifiée des terpènes qui se divisent en plusieurs catégories structurelles, acycliques, monocycliques, bicycliques, tricycliques, polycycliques (Figure I. 2). Ils se trouvent sous forme d'hydrocarbures ou sous forme d'hydrocarbures oxygénés comme les alcools, les cétones, les aldéhydes, les acides et les lactones dans la nature.

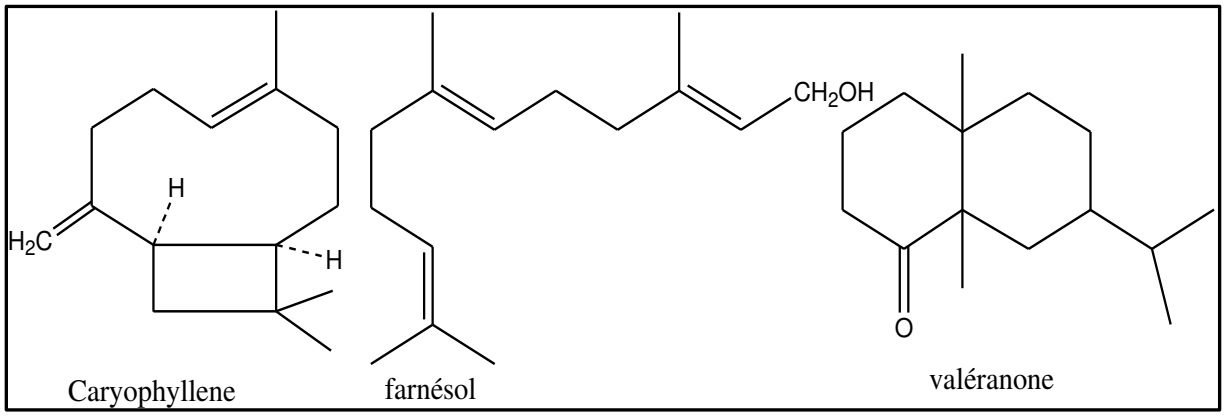


Figure I. 2 : Exemples de quelques sesquiterpènes.

I. 5. 3. Les composés aromatiques :

Une autre classe de composés volatils fréquemment rencontrés est celle des composés aromatiques dérivés du phénylpropane **[25]**. Cette classe comporte des composés odorants bien connus comme la vanilline, l'eugénol, l'anéthole, l'estragole (Figure I. 3) et bien d'autres. Ils sont davantage fréquents dans les huiles essentielles d'Apiaceae (persil, anis, fenouil, etc.) et sont caractéristiques de celles du clou de girofle, de la vanille, de la cannelle, du basilic, de l'estragon, etc **[27]**.

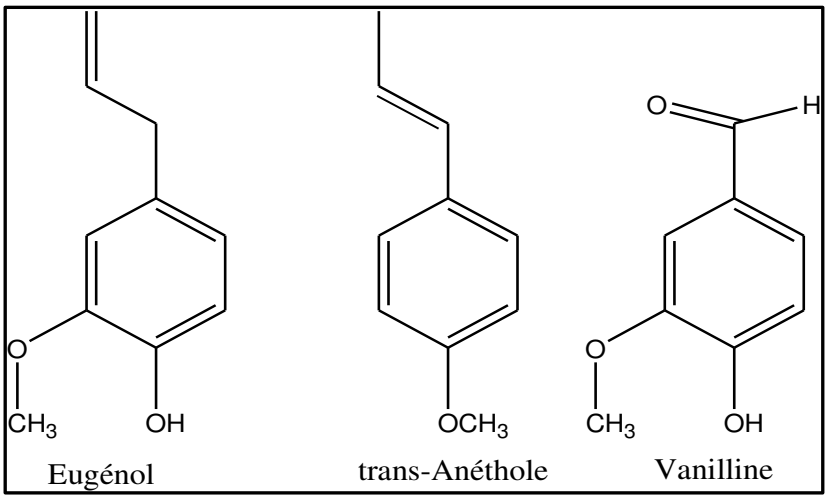


Figure I. 3 : Exemples de composés aromatiques.

I. 6. Facteurs influençant la composition chimique :

Il existe beaucoup de facteurs externes pouvant influencer la composition chimique de l'huile essentielle : la température, le taux d'humidité, la durée d'ensoleillement, la composition du sol, la partie de la plante utilisée, le cycle végétatif de la plante, la méthode utilisée pour l'extraction, sont autant de facteurs susceptibles d'exercer les modifications chimiques. Outre la composition, ces facteurs peuvent également avoir un impact sur la teneur en huile essentielle, par exemple : les *citrus* ont une teneur importante en huile essentielle lorsque la température est élevée. Les fleurs de *Chrysanthemum caronarum* sont riches en huile essentielle sous l'effet de fertilisants **[9]**.

I. 7. Les méthodes d'extraction des huiles essentielles :

Les huiles essentielles sont obtenues avec des rendements très faibles (de l'ordre de 1%) ce qui en fait des substances fragiles, rares, et précieuses. Ainsi les différentes techniques d'extraction des huiles essentielles ou extraits aromatiques doivent d'une part, tenir compte de ces caractéristiques et d'autre part, apporter des performances quantitatives satisfaisantes. Une forte demande toujours plus exigeante basée sur différents phénomènes physiques : la distillation, l'extraction ou la séparation, ces techniques d'extraction seront présentées selon le principe sur lequel elles sont basées et classées en deux catégories distinctes selon le produit final obtenu : une huile essentielle ou un extrait aromatique.

I. 7. 1. Hydrodistillation :

Il s'agit de la méthode la plus simple et de ce fait la plus anciennement utilisée. Le principe de l'hydrodistillation correspond à une distillation hétérogène qui met en jeu l'application de deux lois physiques (loi de Dalton et loi de Raoult) **[28]**.

Le procédé consiste à immerger la matière première végétale dans un ballon lors d'une extraction au laboratoire ou dans un alambic industriel rempli d'eau placé sur une source de chaleur. Le tout est ensuite porté à l'ébullition. La chaleur permet l'éclatement des cellules végétales et la libération des molécules odorantes qui y sont contenues. Ces molécules aromatiques forment avec la vapeur d'eau, un mélange azéotropique. Les vapeurs sont condensées dans un réfrigérant et les huiles essentielles se séparent de l'eau par différence de densité. Au laboratoire, le système équipé d'une cohobe (la phase aqueuse est réutilisée et réinjectée dans le ballon, c'est ce que l'on appelle le cohobage) généralement utilisé pour l'extraction des huiles essentielles est le Clevenger (figure I. 4).

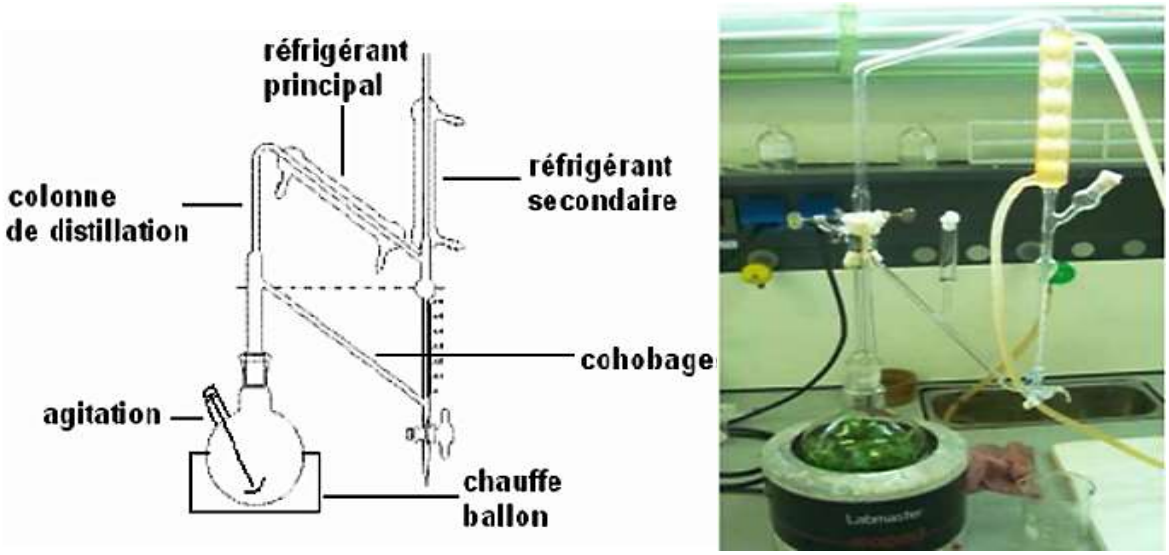


Figure I. 4 : Montage d'hydrodistillation pour l'extraction des huiles essentielles **[29]**.

Les eaux aromatiques ainsi prélevées sont ensuite recyclées dans l'hydrodistillateur afin de maintenir le rapport plante/eau à son niveau initial.

La durée d'une hydrodistillation peut considérablement varier, pouvant atteindre plusieurs heures selon le matériel utilisé et la matière végétale à traiter. La durée de la distillation influe non seulement sur le rendement mais également sur la composition de l'extrait.

I. 7. 2. Entraînement à la vapeur d'eau :

L'entraînement à la vapeur d'eau (Figure I. 5) est l'une des méthodes officielles pour l'obtention des huiles essentielles. A la différence de l'hydrodistillation, cette technique ne met pas en contact direct de l'eau et la matière végétale à traiter. De la vapeur d'eau fournie par une chaudière traverse la matière végétale située au-dessus d'une grille. Durant le passage de la vapeur à travers le matériel, les cellules éclatent et libèrent l'huile essentielle qui est vaporisée sous l'action de la chaleur pour former un mélange « eau + huile essentielle ». Le mélange est ensuite véhiculé vers le condenseur et l'essencier avant d'être séparé en une phase aqueuse et une phase organique : l'huile essentielle. L'absence de contact direct entre l'eau et la matière végétale, puis entre l'eau et les molécules aromatiques évite certains phénomènes d'hydrolyse ou de dégradation pouvant nuire à la qualité de l'huile.

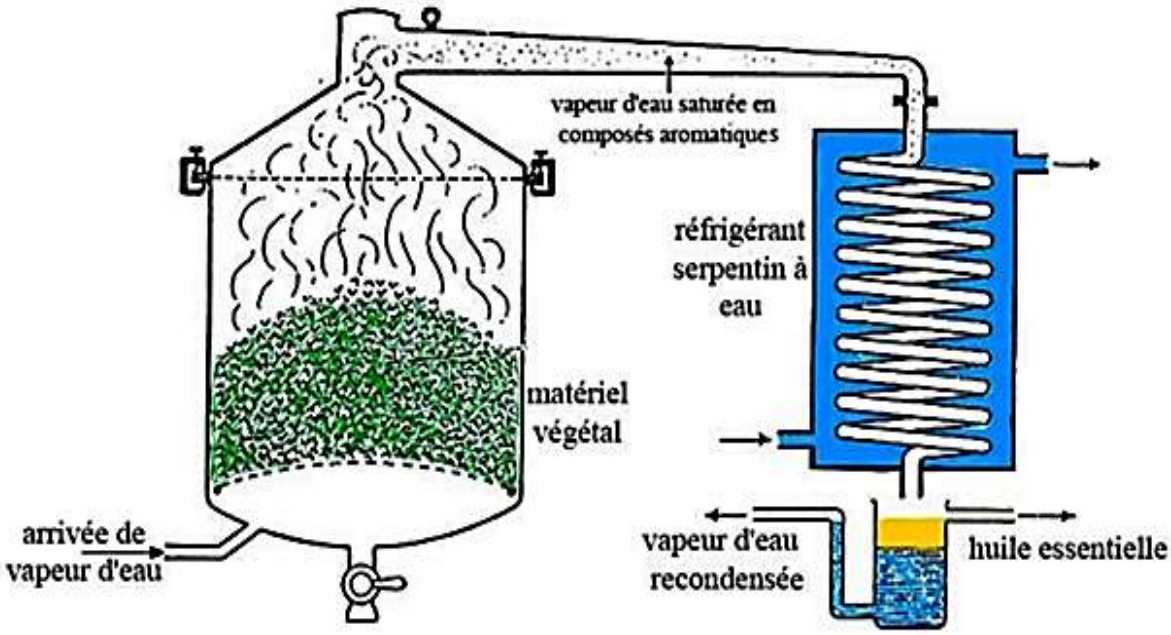


Figure I. 5 : Montage d'entraînement à la vapeur d'eau pour l'extraction des huiles essentielles **[29]**.

I. 7. 3. Hydrodiffusion :

L'hydrodiffusion est une variante de l'entraînement à la vapeur d'eau (Figure I. 6). Dans le cas de l'hydrodiffusion, le flux de vapeur n'est pas ascendant mais descendant. Cette technique exploite ainsi l'action osmotique de la vapeur d'eau. Le principe de cette méthode réside dans l'utilisation de la pesanteur pour dégager et condenser le mélange « vapeur d'eau – huile essentielle » dispersé dans la matière végétale **[30]**.

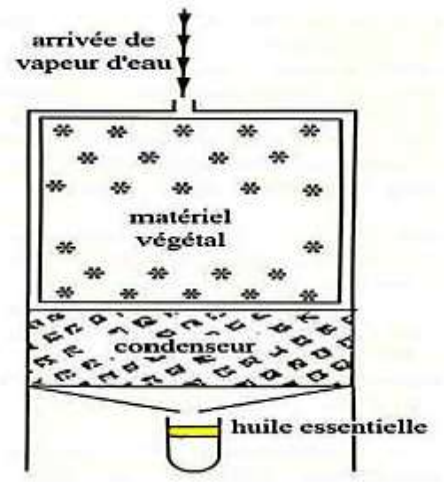


Figure I. 6 : Montage d'hydrodiffusion pour l'extraction des huiles essentielles **[29]**.

L'hydrodiffusion présente l'avantage de ne pas mettre en contact le matériel végétal et l'eau. De plus, l'hydrodiffusion permet une économie d'énergie due à la réduction de la durée de la distillation et donc à la réduction de la consommation de vapeur.

I. 7. 4. Extraction par du CO_2 supercritique :

La technique est fondée sur la solubilité des constituants dans le dioxyde de carbone à l'état super-critique. Grâce à cette propriété, le dioxyde de carbone permet l'extraction dans le domaine liquide (supercritique) et la séparation dans le domaine gazeux. Le dioxyde de carbone est liquéfié par refroidissement et comprimé à la pression d'extraction choisie. Il est ensuite injecté dans l'extracteur contenant le matériel végétal, puis le liquide se détend pour se convertir à l'état gazeux pour être conduit vers un séparateur où il sera séparé en extrait et en solvant (Figure I. 7).

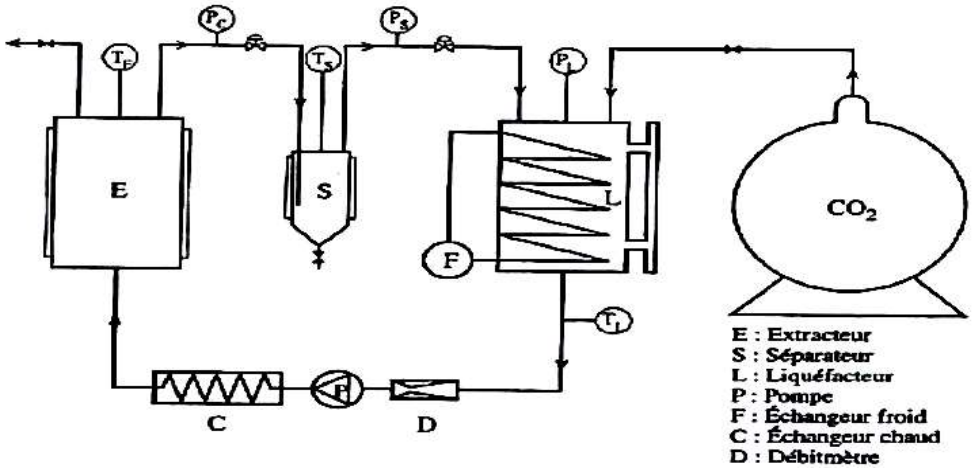


Figure I. 7 : Montage d'extraction par le CO_2 supercritique **[29]**.

L'avantage de cette méthode est la possibilité d'éliminer et de recycler le solvant par simple compression-détente. De plus les températures d'extraction sont basses dans le cas du dioxyde de carbone et non agressives pour les constituants les plus fragiles **[31]**. Cette technique est utilisable pour les essences difficilement distillables.

I. 7. 5. Extraction assistée par micro-onde :

Cette technique d'extraction a été développée au cours des dernières décennies à des fins analytiques **[32]**. Le procédé consiste à irradier par micro-ondes de la matière végétale broyée en présence d'un solvant absorbant fortement les micro-ondes (le méthanol) pour l'extraction de composés polaires ou bien en présence d'un solvant n'absorbant pas les microondes (hexane) pour l'extraction de composés apolaires. L'ensemble est chauffé sans jamais atteindre l'ébullition durant de courtes périodes entrecoupées par des étapes de refroidissement (Figure I. 8).

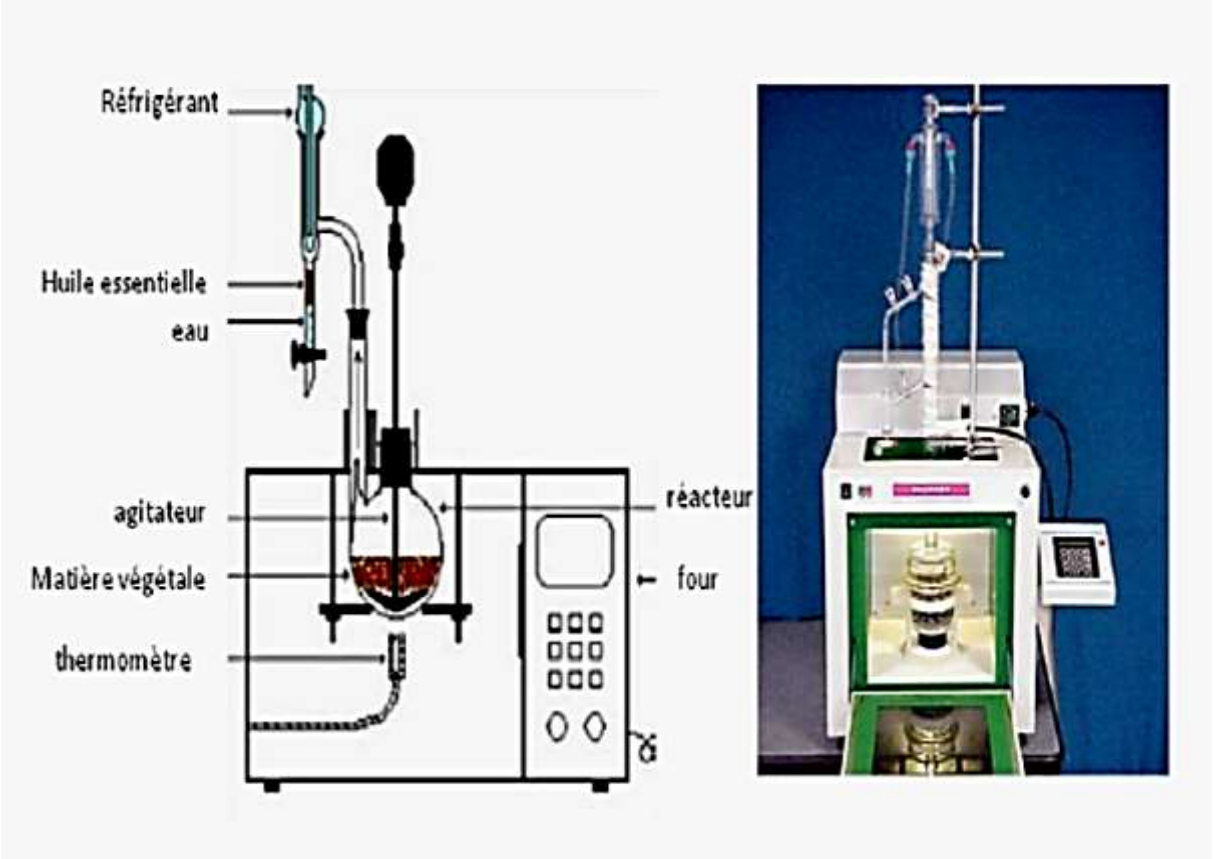


Figure I. 8 : Montage d'extraction assistée par micro-onde : schéma **[29]**.

L'avantage essentiel de ce procédé est de réduire considérablement la durée de distillation et d'obtenir un bon rendement d'extrait.

I. 7. 6. L'expression à froid :

Le procédé d'extraction par expression à froid est assurément le plus simple mais aussi le plus limité. Il est réservé à l'extraction des composés volatils dans les péricarpes des hespéridés ou encore d'agrumes qui ont une très grande importance pour l'industrie des parfums et des cosmétiques. Cependant ce sont des produits fragiles en raison de leur composition en terpènes. Il s'agit d'un traitement mécanique qui consiste à déchirer les péricarpes riches en cellules sécrétrices. L'essence libérée est recueillie par un courant d'eau et reçoit tout le produit habituel de l'entraînement à la vapeur d'eau, d'où la dénomination d'huile essentielle. **[23]**

I. 7. 7. L'extraction par solvants volatils :

La technique d'extraction « classique » par solvant, consiste à placer dans un extracteur un solvant volatil et la matière végétale à traiter. Grâce à des lavages successifs, le solvant va se charger en molécules aromatiques, avant d'être envoyé au concentrateur pour y être distillé à pression atmosphérique.

L'extraction par solvant organique volatil reste la méthode la plus pratiquée. Les solvants les plus utilisés à l'heure actuelle sont l'hexane, le cyclohexane, l'éthanol, le méthanol, le dichlorométhane et l'acétone **[33]**. Le solvant choisi, en plus d'être autorisé devra posséder une certaine stabilité face à la chaleur, la lumière ou l'oxygène, sa température d'ébullition sera de préférence basse afin de faciliter son élimination, et il ne devra pas réagir chimiquement avec l'extrait. L'extraction est réalisée avec un appareil de Soxhlet ou un appareil de Lickens-Nickerson (Figure I. 9).

Ces solvants ont un pouvoir d'extraction plus élevé que l'eau si bien que les extraits ne contiennent pas uniquement des composés volatils mais également bon nombre de composés non volatils tels que des cires, des pigments, des acides gras et bien d'autres substances **[34]**.

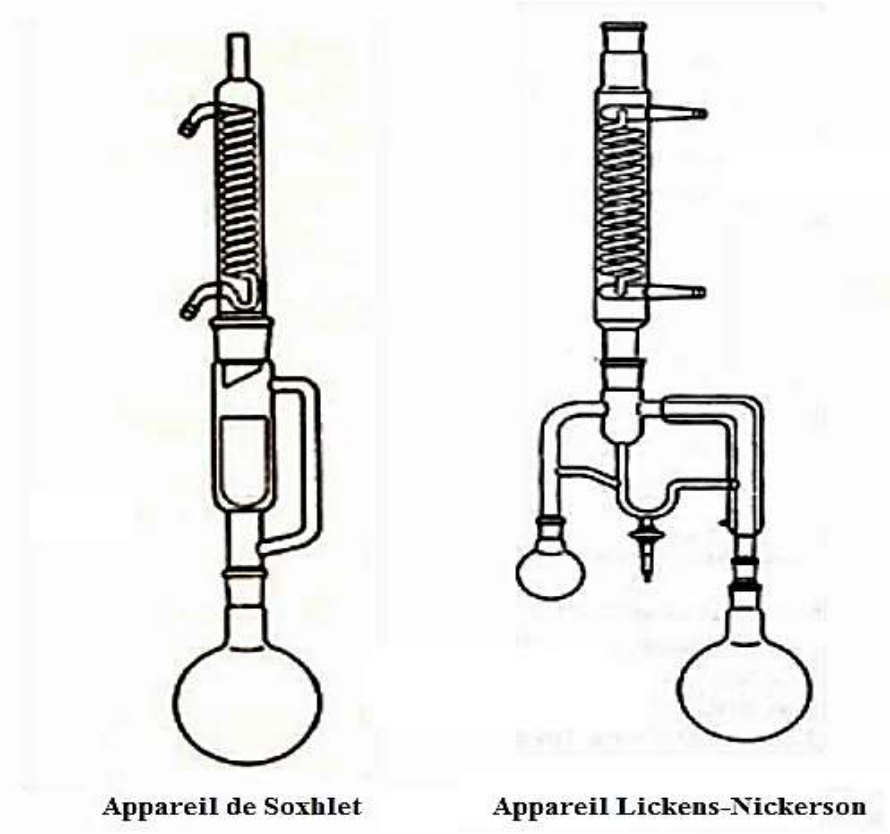


Figure I. 9 : Montage d'extraction par solvant **[29]**.

L'emploi restrictif de l'extraction par solvants organiques volatils se justifie par son coût, les problèmes de sécurité et de toxicité, ainsi que la règlementation liée à la protection de l'environnement. Cependant, les rendements sont généralement plus importants par rapport à la distillation et cette technique évite l'action hydrolysante de l'eau ou de la vapeur d'eau.

I. 8. Les méthodes d'analyse des huiles essentielles

Les huiles essentielles sont des matières premières pour les secteurs de l'industrie pharmaceutique, de l'agroalimentaire et de la cosmétique. La connaissance parfaite de la composition chimique de ces substances permettrait aux professionnels des secteurs précités de pouvoir contrôler leur qualité et de les valoriser. L'identification des composants d'une huile essentielle reste une opération délicate qui nécessite la mise en œuvre de plusieurs techniques qui sont dans certains cas complémentaires **[35]**.

La technique incontournable pour individualiser les constituants d'un mélange reste la chromatographie en phase gazeuse (CPG). Son couplage à un détecteur à ionisation de flamme (DIF) permet la quantification des constituants et le calcul de leurs indices de rétention.

La CPG est souvent combinée avec une technique d'identification spectrale, généralement la spectrométrie de masse (SM) ou la Spectrométrie Infrarouge par Transformée de Fourier (IRTF). Une nouvelle voie d'analyse est le recours à la RMN du carbone-13 décrite par Formacek et Kubeczka et développée par Casanova *et coll*. **[36]**. Elle permet d'identifier les constituants d'un mélange complexe sans individualisation et sans séparation préalable.

I. 8. 1. La chromatographie en phase gazeuse (CPG) :

La chromatographie en phase gazeuse (CPG) est une méthode d'analyse par séparation qui s'applique aux composés gazeux ou susceptibles d'être vaporisés par chauffage sans décomposition. La CPG est la technique usuelle pour l'analyse des huiles essentielles. Elle permet d'obtenir la séparation de composés volatils de mélanges très complexes et l'analyse quantitative des résultats à partir d'un volume d'injection réduit **[37]**.

Pour chacun des composés, des indices de rétention expérimentaux peuvent être obtenus. Ils sont calculés à partir du temps de rétention du composé, encadré dans une gamme étalon d'alcanes ou plus rarement d'esters méthyliques linéaires, à température constante (indice de Kováts) ou en programmation de température (indice de rétention de van den Dool et Kratz) **[38, 39]**. Ces indices sont ensuite comparés avec ceux des produits de référence (mesurés au laboratoire ou décrits dans la littérature). Toutefois l'identification peut être incertaine car il est fréquent d'observer des variations expérimentales, parfois importantes, lorsque l'on compare les indices de rétention obtenus au laboratoire et ceux de la littérature (en particulier sur colonne polaire). C'est pourquoi la comparaison des indices sur deux phases stationnaires de polarité différente est nécessaire. On mesure deux indices : l'un sur phase apolaire, l'autre sur phase polaire. Malgré tout, ceci ne peut suffire à une identification complète, sans l'apport du couplage entre la CPG et une technique d'identification spectroscopique : en général la spectrométrie de masse (CPG/SM).

I. 8. 2. La spectrométrie de masse (SM) :

Le spectromètre de masse permet l'identification et la quantification des composés. Il existe de nombreux types de spectromètres de masse ; tous ont en commun trois éléments : une source, un analyseur et un détecteur.

La source est la partie du spectromètre de masse où sont produits des ions gazeux à partir des molécules introduites. En couplage avec le chromatographe en phase gazeuse, les

composés séparés arrivent dans le spectromètre de masse à fin de détection. Ils sont ionisés par un mécanisme d'impact électronique "ionisation électronique" (IE). La source est maintenue à une température élevée (généralement comprise entre 100 et 250°C) pour éviter la condensation des substances **[40]**. Les ions sont ensuite dirigés vers la partie analytique de l'appareil. Dans le spectromètre de masse, les ions sont séparés selon leur ratio « masse/Charge », à l'aide d'un champ magnétique ou électrique **[41]**. Le faisceau d'ions ayant traversé l'analyseur de masse est détecté et transformé en un signal utilisable.

I. 8. 3. Chromatographie en Phase Gazeuse couplée à la Spectrométrie de Masse (CPG/SM) :

Le couplage CPG/SM peut être utilisé avec divers types d'analyseurs de masses et selon plusieurs modes d'ionisation. Le mode d'ionisation par impact électronique (SM-IE) est la technique la plus utilisée en routine et notamment dans le domaine des huiles essentielles. Il permet de connaître, dans la grande majorité des cas, la masse moléculaire d'un composé et d'obtenir des informations structurales relatives à une molécule à partir de sa fragmentation assez reproductibles et donc utilisables dans une banque de données **[42]**. Dans la source d'ionisation les molécules sont bombardées à l'aide d'électrons, conduisant ainsi à la formation d'ions en phase gazeuse. Les ions sont ensuite dirigés vers l'analyseur de l'appareil. Il existe plusieurs types d'analyseurs de masse dont les plus utilisés pour l'analyse des huiles essentielles sont le «quadripôle» et le «piège à ions» (ou «ion trap» en anglais). Le quadripôle ainsi que l'«ion trap» utilisent la stabilité des trajectoires pour séparer les ions selon leur rapport masse sur charge (m/z).

En ce qui concerne l'«ion trap» (ou quistor), il est constitué d'une électrode circulaire, couverte de deux calottes sphériques. Conceptuellement, on peut voir cet appareillage comme un quadripôle circulaire **[43]**. La superposition des tensions continues et alternatives permet d'obtenir une sorte de «quadripôle à trois dimensions» dans lequel les ions sont gardés captifs «piégés» sur une trajectoire formant une sorte de huit à trois dimensions. Les ions de différentes masses sont présents simultanément dans la trappe, et on cherchera à les expulser en fonction de leur masse pour obtenir le spectre. Cette technique consiste donc à produire des ions directement dans la trappe par impact électronique. Il n'y a pas de source séparée. Les ions sont formés par un flux d'électrons de courte durée et piégés au moyen de radiofréquences **[44]**.

Le faisceau d'ions ayant traversé l'analyseur de masse, est ensuite détecté et transformé en un signal utilisable. Pour ce faire, il existe différents types de détecteurs capables de transformer un courant ionique faible en un signal mesurable. Toutefois, les détecteurs les plus courants sont les multicapteurs d'électrons ou de photons, permettant l'augmentation de l'intensité du signal détecté.

Finalement, l'outil informatique enregistre les données provenant du spectromètre de masse et les convertit en valeurs de masses et d'intensités de pics puis en courant ionique total. Il permet l'examen des données enregistrées et leur manipulation : spectres de masse, chromatogrammes reconstitués, soustraction d'un spectre par rapport à un autre, calcul d'une moyenne sur plusieurs spectres, etc.…

Les spectres de masse ainsi obtenus sont ensuite comparés avec ceux des produits de référence contenus dans les bibliothèques informatisées commerciales disponibles (NIST/EPA/NIH Mass Spectral Library, Wiley Registry of Mass Spectral Data), contenant plusieurs milliers de spectres.

I. 8. 4. Chromatographie en Phase Gazeuse couplée à un Détecteur à Ionisation de Flamme (CPG/DIF) :

Le détecteur à ionisation de flamme (DIF) est le plus populaire des détecteurs. Son principe de fonctionnement est le suivant. Lorsque le soluté est brûlé dans une flamme d'hydrogène, il se forme par combustion du CO et CO_2 qui sont ensuite ionisés en CO+ CO_2+. Ces ions sont collectés par une électrode, souvent en forme de grille cylindrique centrée autour de l'axe de la flamme et portée à un potentiel variant entre 100 et 300 V. Le courant ionique est ensuite amplifié et enregistré. C'est un détecteur universel qui répond à tout composé organique : (il ne répond pas aux gaz permanents ainsi qu'à un certain nombre de gaz minéraux et organiques : CO, CO_2, COS, CS_2, SO_2, H_2O, NH_3, NO, N_2O, HCOOH, HCHO, $SiCl_4$). Ce détecteur est très sensible mais nécessite des débits gazeux très stables. Le courant très faible qui résulte de la détection est fortement amplifié et transformé en une tension mesurable par un électromètre. L'aire du pic varie selon la quantité de composé élué dans une relation linéaire sur un très large domaine de concentrations. Comme par ailleurs la réponse (signal/ masse) du détecteur DIF est assez comparable pour de nombreux composés organiques (contrairement au cas de la détection en SM où les réponses peuvent être très différentes), la CPG/DIF est très utile pour l'analyse quantitative et notamment pour l'analyse

semi-quantitative des huiles essentielles (par normalisation directe des aires du chromatogramme sans aucun étalonnage préalable) **[45]**.

I. 8. 5. L'indice de rétention :

L'un des problèmes de la CPG est le manque de reproductibilité des temps de rétention d'un appareil à l'autre ou d'une colonne à l'autre, même si elles sont de nature identique. Pour résoudre ce problème lié aux phénomènes complexes qui interviennent pendant l'élution (variation des conditions opératoires), Kováts a proposé l'utilisation d'un indice de rétention (IK). En considérant que la montée de température du four est linéaire sur la plage de température étudiée, van den Dool et Kratz **[46]** ont donc utilisé ces indices qui sont pratiquement indépendants des paramètres et des conditions d'analyse par chromatographie gazeuse à température programmée **[47]**.

L'indice de rétention est une grandeur caractéristique de chaque composé et du type de colonne. Deux types de phases stationnaires sont généralement utilisés ce qui permet de résoudre le plus souvent les problèmes de coélutions.

Sur colonne "apolaire", les composés sont élués approximativement dans l'ordre de leur point d'ébullition.

Dans le cas d'une colonne " polaire " les composés les plus " polaires " seront retenus plus facilement et donc auront un indice de rétention plus élevé.

Quel que soit le type de colonne, les constituants d'une même famille sont élués dans le même ordre. Les IK sont calculés par comparaison entre les temps de rétention (t_r) du composé étudié et ceux d'une série d'alcanes linéaires permettant un « étalonnage » du chromatogramme.

Ces IK sont définis par la relation :

$$\mathrm{IK}_A = 100n + \left[100 * \frac{\log t_r[A] - \log t_r[Cn]}{\log t_r[C(n+1)] - \log t_r[Cn]} \right] \qquad (1)$$

Où : IK est l'indice de Kováts.

n est le nombre de carbones de la paraffine précédant immédiatement le composé étudié.

t_r [A] est le temps de rétention du composé étudié.

t_r [Cn] est le temps de rétention de la paraffine précédant immédiatement le composé étudié.

t_r [C(n+1)] est le temps de rétention de la paraffine à n+1 atomes de carbone suivant immédiatement le composé.

I. 9. Utilisation des huiles essentielles :

Ces produits naturels présentent un grand intérêt comme matière première destinée à différents secteurs d'activité :

a. En pharmacie :

Les huiles essentielles peuvent être utilisées :

- Comme aromatisant des médicaments destinés à la voie orale **[48]**.
- Pour leurs actions physiologiques (Menthe, Verveine, Camomille) **[49]**.

b. Dans l'industrie :

- Parfumerie et cosmétologie :

De nombreux parfums sont toujours d'origine naturelle et certaines huiles essentielles constituent des bases de parfums.

Exemples : Rose, Jasmin, Vétiver, Ylang-ylang, etc.… **[49]**.

- Alimentation :

Les huiles essentielles (huile de citron, de menthe, de girofle) sont très utilisées dans l'aromatisation des aliments (jus de fruits, pâtisserie) **[48, 49]**.

Quel que soit le secteur d'activité, l'analyse des huiles essentielles reste une étape importante qui, malgré les progrès constants des différentes techniques de séparation et d'identification, demeure toujours une opération délicate qui nécessite la mise en œuvre simultanée ou successive de diverses techniques **[50]**.

I. 10. Les principales propriétés des huiles essentielles : [51-53]

Les huiles essentielles possèdent de nombreuses propriétés.

I. 10. 1. Anti-infectieuses :

- Antibactériennes :

Les molécules aromatiques possédant l'activité antibactérienne la plus importante sont les phénols contenus par exemple dans l'huile essentielle de clou de girofle.

-Antivirales :

Les virus sont assez sensibles aux huiles essentielles à phénol et à monoterpénol. Plus d'une dizaine d'huiles essentielles possèdent des propriétés antivirales. Nous pouvons citer l'huile essentielle de Ravintsara, l'huile essentielle de Bois de Hô, ou l'huile essentielle de Cannelle de Ceylan.

-Antifongiques :

Les huiles essentielles utilisées pour leurs propriétés antifongiques sont les mêmes que celles citées précédemment cependant la durée du traitement sera plus longue. Par exemple, les huiles essentielles de Cannelle, de Clou de girofle ou de Niaouli sont des antifongiques.

- Antiparasitaires :

Les molécules aromatiques avec les phénols ont une action puissante contre les parasites.

Le thym à linalol, la sarriette des montagnes sont d'excellentes huiles essentielles antiparasitaires.

- Antiseptiques :

Les propriétés antiseptiques et désinfectantes sont souvent retrouvées dans les huiles essentielles possédant des fonctions aldéhydes ou des terpènes comme l'huile essentielle d'*Eucalyptus radiata*.

- Insecticides :

Certaines huiles essentielles sont insectifuges ou insecticides comme celles possédant des fonctions aldéhydes comme le citronnellal contenu dans l'Eucalyptus citronné ou la citronnelle.

I. 10. 2. Anti-inflammatoires :

Les huiles essentielles avec les aldéhydes ont des propriétés actives contre l'inflammation par voie interne comme l'huile essentielle de gingembre.

I. 10. 3. Régulatrices du système nerveux :

-Antispasmodiques :

Les huiles essentielles possédant des esters ou des éthers présentent une action sur les spasmes des muscles lisses ou striés comme l'huile essentielle d'hélichryse.

-Calmantes, anxiolytiques :

Les aldéhydes type citrals contenue par exemple dans l'huile essentielle de Mélisse ou celle de verveine citronnée favorisent la détente et le sommeil.

- Analgésiques, antalgiques :

Les huiles essentielles les plus connues pour leur action antalgiques sont les huiles essentielles d'eucalyptus citronné, de gingembre, de lavande vraie.

I. 10. 4. Drainantes respiratoires :

-Expectorantes :

Les huiles essentielles riches en oxyde (1,8-cinéole) comme l'huile essentielle d*'Eucalyptus globulus* ou de Romarin agissent sur les glandes bronchiques et sur les cils de la muqueuse bronchique.

- Fluidifiantes :

Les huiles essentielles possédant des cétones (comme la verbénone contenue dans l'huile essentielle de Romarin) ont une action mucolytique en dissolvant les sécrétions accumulées au niveau de la muqueuse.

I. 10. 5. Digestives :

Les huiles essentielles de cumin (avec la molécule de cuminal), d'anis étoilé ou par exemple d'estragon ont une action digestive et apéritive. Elles permettent la stimulation de la sécrétion des sucs digestifs. L'huile essentielle de menthe poivrée atténue les nausées.

I. 10. 6. Cicatrisantes :

Les huiles essentielles cicatrisantes sont les huiles essentielles de Ciste *(Cistus ladaniferus)*, de Lavande vraie *(Lavandula vera)*, d'Immortelle *(Helichrysum italicum)*, de myrrhe *(Commiphora myrrha).* On utilise souvent un mélange de plusieurs huiles essentielles cicatrisantes avec une huile végétale comme l'huile d'amande douce.

I. 11. Les principales voies d'utilisation des huiles essentielles : [52, 54]

I. 11. 1. La diffusion atmosphérique :

Lors de la diffusion dans l'atmosphère, il faut prendre soin de choisir des huiles essentielles labélisées biologiques, pures, et adaptées afin d'éviter les allergies et les contre-

indications. Certaines huiles essentielles peuvent être irritantes pour les muqueuses respiratoires. Il faut éviter de diffuser en continu dans une pièce close et toute la nuit en présence d'une personne qui dort, mais plutôt une quinzaine de minutes, une à trois fois par jour. Le diffuseur doit être placé de façon à ne pas diffuser directement vers le visage ou les yeux.

Il faut utiliser un diffuseur qui ne chauffe pas les huiles essentielles afin qu'elles ne s'oxydent pas.

Cette voie d'administration est préférée dans certaines indications comme pour les huiles essentielles utilisées pour une indication respiratoire comme l'*eucalyptus globulus*, le pin.

I. 11. 2. La voie interne :

La voie interne peut être utilisée avec beaucoup de précaution.

- La voie orale :

L'ingestion ne doit jamais se faire pure : il faut toujours les diluer avec de l'huile végétale ou par exemple dans du miel car celles-ci ne sont pas solubles dans l'eau et laisser fondre sous la langue. Il existe des capsules à avaler déjà prêtes avec une base d'huile végétale. Il est préférable de ne jamais ingérer plus de trois gouttes d'une même huile essentielle plus de trois fois par jour.

-La voie rectale :

La voie rectale, avec l'emploi de suppositoires est le mode d'utilisation préconisé dans les infections broncho-pulmonaires. Cette voie permet une absorption rapide et efficace des principes actifs des huiles essentielles en évitant le circuit digestif.

-La voie gynécologique :

Elle permet une action rapide localement avec l'emploi d'ovules vaginaux fabriqués sur le même modèle que les suppositoires en aromathérapie.

I. 11. 3. La voie externe :

- La voie cutanée :

La voie cutanée peut être utilisée dès trois ans en effleurage. Elle est beaucoup utilisée en aromathérapie.

L'huile essentielle est appliquée pure ou en mélange avec une huile végétale préférentiellement au niveau des poignets ou du plexus solaire.

- Le bain :

On peut également mettre quelques gouttes d'huile essentielle dans un bain. Là encore, la dilution avec une huile végétale hydrosoluble est recommandée pour éviter tout risque de réaction cutanée du fait de leur insolubilité et ainsi de leur contact avec la peau en trop grande concentration.

Les huiles essentielles sont toujours insolubles dans l'eau, pour cette raison, il faut utiliser un dispersant en quantité quatre fois supérieure à celle de l'huile essentielle pour disperser le tout dans le bain.

I. 12. Toxicité des huiles essentielles :

Les huiles essentielles ne sont pas des produits qui peuvent être utilisées sans risque. Certaines huiles essentielles sont dangereuses lorsqu'elles sont appliquées sur la peau, en raison de leur pouvoir irritant (les huiles riches en thymol, ou en carvacrol), allergène (huiles riches en cinnamaldéhyde) ou photo-toxique (huiles de *citrus* contenant des furacoumarines), d'autres huiles essentielles ont un effet neurotoxique (les cétones comme l'α-thujone sont toxiques pour les tissus nerveux). La toxicité des huiles essentielles est assez mal connue. Il manque des données sur leurs éventuelles propriétés mutagènes et cancérigènes. La plupart du temps, sous le terme de toxicité sont décrites des données expérimentales accumulées en vue d'évaluer le risque que représente leur emploi. Il existe quelques huiles essentielles dont certains composés sont capables d'induire la formation de cancer, c'est le cas par exemple de dérivés d'allybenzène ou de propenylbenzène comme le safrole, l'estragole, la β-arasone, et le methyl-eugénol. Des chercheurs ont mis en évidence l'activité hépatocarcinogénique de ces composés chez les rongeurs. Le safrole et l'estragole, sont métabolisés au niveau du foie des rats en dérivés hydroxylés puis en esters sulfuriques électrophiles qui sont capables d'interagir avec les acides nucléiques et les protéines. Ces résultats sont controversés, car il existe des différences chez l'homme dans le processus de métabolisation de ces composés. Le safrole par exemple est métabolisé en dihydroxysafrole et trihydroxysafrole non cancérigènes **[55]**.

I. 13. Conservation des huiles essentielles : [56]

Les huiles essentielles de bonne qualité peuvent se conserver plusieurs années sous certaines conditions, jusqu'à cinq ans pour les H.E.C.T par exemple. Seules les essences de *citrus* se gardent un peu moins longtemps (trois ans).

Les huiles essentielles sont volatiles, il ne faut donc pas oublier de bien fermer les flacons. Il est préférable de les conserver dans un flacon en aluminium ou en verre teinté (brun, vert, ou bleu) et de les garder à l'abri de la lumière à une température ambiante jusqu'à vingt degrés.

Il existe des normes spécifiques sur l'emballage, le conditionnement et le stockage des huiles essentielles **[57]** ainsi que sur le marquage des récipients contenant des huiles essentielles **[58]**.

Première partie : Etude bibliographique

II. Aspects théoriques de la modélisation moléculaire

II. Aspects théoriques de la modélisation moléculaire :

II. 1. Introduction :

Toute recherche théorique est sous-tendue par deux motivations essentielles : la compréhension et la prévision, c'est-à-dire la compréhension de ce qui a déjà été fait et la prévision de ce qui est éventuellement réalisable. La prévision répond à des questions du type : "*Que se passerait-il si… ?*", ou "*Est-ce qu'on pourrait faire… ?*" ou encore "*Quelle serait la valeur de… ?*". La réponse traditionnelle serait de faire l'expérience. Mais à une époque où le prix des calculs par ordinateur baisse continuellement, tandis que celui des produits chimiques, des appareils, de la main-d'œuvre qualifiée, etc… ne cesse de croître, il est de plus en plus intéressant d'exploiter les modèles théoriques de toutes sortes afin d'aider à la conception de nouvelles espèces chimiques **[59]**.

L'ambition d'un chimiste théoricien est d'être capable de prédire, confirmer ou réinterpréter l'expérience à l'aide de la modélisation moléculaire. En effet, la persévérance des chercheurs, et surtout la puissance de leurs moyens informatiques jouent en faveur de la chimie théorique, et son champ d'application.

La recherche et la synthèse de nouveaux composés chimiques et biochimiques sont aujourd'hui souvent associées à une étude par modélisation moléculaire. La modélisation moléculaire, est une méthodologie couramment utilisée. Depuis un peu plus d'une trentaine d'années, elle s'est imposée progressivement comme un outil de choix pour la découverte et la conception orientée de nouvelles molécules actives. Auparavant, n'étaient pratiqués que des tests biologiques systématiques sur un grand nombre de molécules et, souvent, seule la chance permettait de mettre en évidence une piste intéressante **[60]**.

La modélisation moléculaire est un outil destiné aux chercheurs préoccupés par la structure et la réactivité des molécules. La connaissance de la structure des édifices moléculaires permet de comprendre ce qui est réalisé dans une transformation physique, chimique ou biologique. Elle peut permettre aussi de prévoir de telles transformations.

La compréhension comme la prévision sont considérablement facilitées lorsque l'on peut visualiser les structures. Une molécule est correctement décrite par sa géométrie et ses propriétés thermodynamiques. La visualisation doit rendre compte de l'ensemble de ces caractéristiques. La question essentielle est de représenter une molécule sur l'écran de la façon la plus proche possible de la "réalité" **[61]**.

II. 2. Méthodes de la modélisation moléculaire :

La modélisation moléculaire est une application des méthodes théoriques et des méthodes de calcul pour résoudre des problèmes impliquant la structure moléculaire et la réactivité chimique. Ces méthodes peuvent être relativement simples et utilisables rapidement ou au contraire elles peuvent être extrêmement complexes et demander des centaines d'heures de temps d'ordinateur, même sur un super-ordinateur. En plus, ces méthodes utilisent souvent des moyens infographiques très sophistiqués qui facilitent grandement la transformation de quantités impressionnantes de nombres en quelques représentations graphiques facilement interprétables.

Différentes approches sont envisageables dans le cadre des outils de modélisation moléculaire. Si celles de mécanique classique, économique en termes de temps de calcul, permettent de traiter des systèmes moléculaires de grande taille, les méthodes quantiques (*ab-initio*, semi-empiriques ou théorie de la fonctionnelle de la densité) sont, quant à elles, capables de calculer les propriétés électroniques des systèmes.

II. 2. 1. La mécanique quantique (MQ) :

La chimie quantique applique les principes de la mécanique quantique aux systèmes moléculaires pour tenter de résoudre l'équation de Schrödinger **[62]**. En effet, le comportement électronique et nucléaire des molécules, étant responsable des propriétés chimiques, peut être décrit de façon réaliste à partir de cette équation. Différentes méthodes de résolution ont alors été développées. En particulier, le développement grandissant des moyens informatiques ont permis le développement de ces méthodes **[63]**. On distingue trois approches :

- Les méthodes *ab-initio* : elles visent à la résolution de l'équation électronique de Schrödinger pour déterminer la fonction d'onde approchée du système étudié.

- La théorie de la fonctionnelle de la densité (DFT) : elle recherche la densité électronique la plus proche possible en partant du principe que la densité électronique d'un système d'électrons détermine toutes les propriétés de ce système.

- Les méthodes semi-empiriques : elles sont une simplification des méthodes *ab-initio* et sont paramétrées de façon à reproduire des résultats expérimentaux. Les méthodes semi-empiriques sont surtout utilisées pour des systèmes moléculaires de très grande dimension ou pour une première optimisation de structures moléculaires.

II. 2. 1. 1. Principe de base de la mécanique quantique :

La mécanique quantique décrit la matière comme un ensemble de noyaux atomiques autour desquels gravitent des électrons, eux-mêmes décrits explicitement par leur probabilité de présence en un point et représentés par des fonctions d'onde. En d'autres termes, l'application des lois de la mécanique ondulatoire aux électrons, permet de déterminer l'état électronique d'un système d'atomes, mais aussi l'ensemble de ses propriétés observables (structurales : géométries, angles, longueurs... ; énergétiques : énergies de liaison, d'excitation... ; spectroscopiques : fréquences de vibration, spectres UV-visible, IR et microonde... ; électroniques, magnétiques et réactionnelles : barrières d'activation...). Les bases du calcul quantique ont été posées en 1925 par Heisenberg, Born et Jordan, puis finalisées en 1926 par Schrödinger et sa fameuse équation (2), dont le formalisme permet de décrire rigoureusement la nature microscopique de la matière **[63]**.

$$\hat{H}\Psi = E\Psi \qquad (2)$$

où $\hat{H}$ est l'opérateur *hamiltonien* et E l'énergie du système.

L'hamiltonien $\hat{H}$ total d'une molécule comportant N noyaux et n électrons, est défini par la somme de cinq termes (terme cinétique des électrons, terme cinétique des noyaux, terme de répulsions électrons - électrons, terme de répulsions noyaux-noyaux et terme d'attractions électrons - noyaux).

Dans le cas général des systèmes d'intérêt chimique, qui sont le plus souvent polyatomiques et multi-électroniques, cette équation ne peut être résolue analytiquement. Des approximations ont donc été proposées.

L'approximation de Born et Oppenheimer **[64]** établit que la grande différence de masse et donc de vitesse qui existe entre noyaux et électrons implique que leurs mouvements peuvent être étudiés séparément. On peut alors considérer le comportement des électrons dans une molécule en supposant que les noyaux occupent les positions fixes dans l'espace. On aboutit alors à l'équation de Schrödinger électronique :

$$\hat{H}_{el}\Psi_{el} = E_{el}\Psi_{el} \qquad (3)$$

Cet hamiltonien électronique $\hat{H}_{el}$ comporte trois contributions ; la première est relative à l'énergie cinétique des électrons, la seconde à l'attraction entre noyaux et électrons et la dernière correspond à la répulsion coulombienne entre électrons. On écrit :

$$\hat{H} = -\sum_{i=1}^{n} \frac{1}{2} \nabla_i^2 - \sum_{i=1}^{n} \sum_{A=1}^{M} \frac{Z_A}{r_{iA}} + \sum_{i=1}^{n} \sum_{j>1}^{n} \frac{1}{r_{ij}} \quad (4)$$

où ∇_i est l'opérateur énergie cinétique correspondant à l'électron i, n désigne le nombre d'électrons du système, M est le nombre de noyaux, Z_A est le numéro atomique de l'atome A et r_{ij} est la distance entre les électrons i et j.

L'hamiltonien du système s'obtient en ajoutant à l'hamiltonien électronique le terme de répulsion entre noyaux $\hat{U}_{NN}$:

$$\hat{H} = \hat{H}_{el} + \hat{U}_{NN} \quad (5)$$

$$\hat{U}_{NN} = \sum_{A=1}^{M} \sum_{B>A}^{M} \frac{Z_A Z_B}{R_{AB}} \quad (6)$$

et R_{AB} désigne la distance entre les noyaux A et B.

Ce dernier terme peut être considéré comme constant. En conséquence, on peut écrire :

$$E = E_{el} + \hat{U}_{NN} \quad (7)$$

On peut exprimer l'hamiltonien électronique sous la forme de deux contributions ; un opérateur de cœur $\hat{H}$ et un opérateur de répulsion inter-électronique :

$$\hat{H}_{el} = \hat{H}^c + \sum_{i=1}^{n} \sum_{j>1}^{n} \frac{1}{r_{ij}} \quad (8)$$

Avec :

$$\hat{H}^c = -\sum_{i=1}^{n} \frac{1}{2} \nabla_i^2 + V_{Ni} \quad (9)$$

où i et j se rapportent à 2 électrons du système et où V_{Ni} est l'opérateur associé à l'interaction attractive noyaux-électrons.

Dans l'approximation orbitale, la fonction d'onde électronique Ψ_{el} définie ci-dessus est représentée par un *déterminant de Slater* **[65, 66]**, construit à partir de fonctions mono électroniques produits d'une orbitale spatiale et d'une fonction de spin : les spin-orbitales χ. Une telle fonction a l'avantage de vérifier le principe d'exclusion de Pauli **[67]**.

L'opérateur $\hat{H}^c$ est une somme de termes mono électroniques. L'énergie a pour expression :

$$E = \langle \Psi_{el} | H_{el} | \Psi_{el} \rangle = \sum_{k=1}^{n} H_k^c + \sum_{k=1}^{n} \sum_{l<k}^{n} \left(J_{kl} - K_{kl} \right) \quad (10)$$

où Ψ est la fonction d'onde multi électronique (n électrons) dont le déterminant de Slater est construit à partir de n spin-orbitales. Les méthodes dites du champ auto-cohérent ou SCF (Self Consistent Field) utilisent le principe variationnel pour résoudre l'équation (10). Le meilleur ensemble de spin-orbitales est celui qui minimise l'énergie, tout en vérifiant le principe d'orthonormalité des autres spin-orbitales.

Cette approche mène à la définition de l'opérateur, ou hamiltonien, de Hartree-Fock qui vérifie la relation :

$$F(1) = \hat{H}^c(1) + \sum_{i=1}^{N} \left[\hat{J}_i(1) - \hat{K}_i(1) \right] \quad (11)$$

où $\hat{H}^c$ est l'hamiltonien de cœur relatif à un électron, $\hat{J}_i(1)$ et $\hat{K}_i(1)$ sont respectivement les opérateurs coulombien et d'échange.

Les équations de Hartree-Fock:

$$F \chi_k = e_k \chi_k \quad (12)$$

définissent un ensemble de fonctions permettant de construire un déterminant de Slater qui approche le mieux la fonction d'onde multiélectronique du système étudié. En pratique, pour résoudre ces équations, il faut connaître les spin-orbitales solution de l'équation (12) et qui définissent les opérateurs $\hat{J}$ et $\hat{K}$. C'est donc un processus itératif avec une estimation initiale de la matrice de densité, jusqu'à ce que le système ait atteint sa cohérence interne, d'où le nom de méthode du champ auto-cohérent.

L'application de l'approximation CLOA (Combinaison Linéaire d'Orbitales Atomiques) aux fonctions propres de l'opérateur de Hartree-Fock mène aux équations de Roothaan **[68]** :

$$FC = SCE \quad (13)$$

où F est la matrice de Fock, C est la matrice des coefficients des orbitales atomiques dans les orbitales moléculaires, S est la matrice de recouvrement des orbitales atomiques et E est la matrice diagonale des énergies.

II. 2. 1. 2. Les méthodes *ab-initio* :

Les méthodes *ab-initio* sont des méthodes non empiriques, toutes les intégrales sont rigoureusement calculées et il n'y a pas d'approximation à faire sauf celle de Born-Oppenheimer et l'approximation OM-CLOA. Dans les méthodes *ab-initio*, toutes les particules (noyau et électrons) sont traitées explicitement. On n'utilise aucun paramètre empirique dans le calcul de l'énergie.

Les méthodes *ab-initio* se divisent en deux sous familles : les méthodes Hartree – Fock **[69, 70]**, et les méthodes post Hartree-Fock **[71]**. La principale différence entre ces deux méthodes est que les interactions électroniques sont négligées dans les méthodes HF et réintroduites dans les méthodes post HF. Ces méthodes ne peuvent être appliquées qu'à des systèmes de quelques dizaines d'atomes pour les méthodes HF et d'une dizaine d'atomes seulement pour les méthodes post HF **[72]**.

II. 2. 1. 3. La théorie de la fonctionnelle de la densité (DFT) :

La théorie de la fonctionnelle de la densité est basée sur le postulat proposé par Thomas et Fermi qui dit que les propriétés électroniques peuvent être décrites en terme de fonctionnelles de la densité électronique, en appliquant localement des relations appropriées a un système électronique homogène **[73]**.

Hohenberg et Kohn, en 1964 **[74, 75]**, ont repris la théorie de Thomas-Fermi et ont montré qu'il existe une fonctionnelle de l'énergie $E[\rho(r)]$ associée à un principe variationnel, ce qui a permis de jeter les bases de la théorie de la fonctionnelle de la densité.

La théorie de la fonctionnelle de la densité est basée sur le théorème Hohenberg-Kohn **[76]**, qui établit que l'énergie d'un système dans son état fondamental est une fonctionnelle de la densité électronique de ce système, $\rho(r)$, et que toute densité, $\rho'(r)$, autre que la densité réelle conduit nécessairement à une énergie supérieure. Ainsi contrairement aux méthodes précédentes, la théorie de la fonctionnelle de la densité ne consiste pas à chercher une fonction d'onde complexe, Ψ , à 3N-dimensions décrivant le système à étudier, mais plutôt une simple fonction à trois dimensions : la densité électronique totale ρ **[77]**. Il existe trois types de fonctionnelles énergies d'échange-corrélation : les fonctionnelles locales, les fonctionnelles à correction du gradient et les fonctionnelles hybrides.

II. 2. 1. 4. Les méthodes semi-empiriques :

Dans les méthodes *ab-initio* la quasi-totalité du temps de calcul est consommée par les calculs des intégrales, et dans le but de réduire ce temps de calcul, il est nécessaire de simplifier les équations de Roothaan. Une méthode semi-empirique est une méthode dans laquelle une partie des calculs nécessaires aux calculs Hartree-Fock est remplacée par des paramètres ajustés sur des valeurs expérimentales (l'hamiltonien est toujours paramétré par comparaison avec des références). En général toutes ces méthodes sont très précises pour des familles de produits données voisines de celles utilisées pour la paramétrisation. Selon la nature des approximations utilisées **[78]**, on distingue plusieurs variantes :

- **CNDO/2 :** (Complete Neglect of Differential Overlep) 1ere méthode semi empirique, elle a été proposée par Pople, Santry et Segal en 1965 **[79]**. Méthode présentant certains défauts comme l'ignorance de la règle de Hund.
- **INDO:** (Intermediate Neglect of Differential Overlap). Proposée par Pople, Beveridge et Dobosh en 1967 **[80]**. Elle permet de distinguer entre les états singulets et les états triplets d'un système en conservant les intégrales d'échange.
- **Méthode NDDO:** (Neglect of Diatomic Differential Overlap). Proposée par Pople, Santry et Segal en 1965 **[79]**. Toutes les intégrales biélectroniques bicentrées sont retenues.
- **MINDO/3 :** Proposée par Bingham, Dewar et Lo en 1975 **[81]**. Paramétisation effectuée en se référant aux résultats expérimentaux et non pas aux résultats *ab-initio*, de plus l'algorithme d'optimisation utilisé est très efficace (Davidon-Fletcher-Powel). Cependant, elle surestime la chaleur de la formation des systèmes insaturés et sous-estime celle des molécules contenant des atomes voisins ayant des paires libres.
- **MNDO:** (Modified Neglect of Diatomic Overlap). Proposée par Dewar et Thiel en 1977 **[82]**. Méthodes basées sur l'approximation NDDO (Neglect of Diatomic Differential Overlap) qui consiste à négliger le recouvrement différentiel entre orbitales atomiques sur des atomes différents. Cette méthode ne traite pas les métaux de transition et présente des difficultés pour les systèmes conjugués.
- **AM 1 :** (Austin Model 1) Proposée par Dewar, Zoebisch, Healy et Stewart en 1985 **[83]**. Il a tenté de corriger les défauts de MNDO.
- **PM 3 :** (Parametric Method 3). Proposée par Stewart en 1989 **[84]**. Présente beaucoup de points en commun avec AM1, d'ailleurs il existe toujours un débat concernant les mérites relatifs de paramétrisation de chacune d'elles.

- **SAM 1:** (Semi *ab-initio* Model 1). La méthode la plus récente proposée par Dewar en 1993 **[85]**. Elle inclut la corrélation électronique.
- **PM 6:** (Parametric Method 6). Stewart a développé, en 2007, une nouvelle méthode s'appuyant sur PM3 nommée PM6 dans laquelle a été incorporé un nouveau paramétrage coeur-coeur avec un accent sur les composés d'intérêt biologique **[86]**. Pour cela, ils ont modifié l'interaction coeur-coeur par une fonction de Voityuk **[87]** qui permet de prendre en compte la répulsion de deux atomes non chargés grâce à l'incorporation d'un terme diatomique. De plus, les paramètres pour le traitement des orbitales d ont été ajoutés ce qui permet d'avoir, désormais, 80 atomes paramétrés pour cette méthode et de pouvoir ainsi traiter les métalloprotéines. Malgré les améliorations apportées, la méthode PM6 échoue pour la description des interactions non-covalentes notamment en ce qui concerne la dispersion et la représentation des liaisons hydrogènes **[88]** en sous estimant la force de ces interactions.

II. 2. 2. La mécanique moléculaire (MM) :

La mécanique moléculaire est apparue en 1930 **[89]**, mais s'est développée à partir des années 1960, avec les progrès d'accessibilité et de performance des ordinateurs. Elle permet de déterminer l'énergie d'une molécule en fonction de ses coordonnées atomiques et de chercher des minima de l'énergie correspondant à des conformères stables **[90, 91]**.

Les techniques de modélisation basées sur la mécanique quantique souffrent d'un inconvénient majeur : elles sont très coûteuses en termes de temps de calcul et ne sont dès lors applicables qu'à des systèmes moléculaires de taille restreinte. Au final, le temps nécessaire au traitement d'un système par les méthodes *ab-initio* est environ proportionnel à la quatrième puissance du nombre d'électrons qu'il contient. L'utilisation de ces techniques peut s'avérer problématique pour l'étude d'objets macromoléculaires tels qu'une enzyme en interaction avec un inhibiteur.

Par contre, la mécanique moléculaire **[92]** considère l'énergie d'un système uniquement en fonction de ses positions atomiques. Cette approximation repose elle aussi sur les travaux de Born et Oppenheimer. En effet, l'approximation de Born-Oppenheimer, en découplant les mouvements des noyaux et des électrons d'une molécule, postule que ces derniers peuvent s'adapter de manière quasi instantanée à la position des noyaux. Le fait d'ignorer les mouvements des électrons épargne ainsi un temps de calcul considérable.

En particulier, la mécanique moléculaire permet l'étude d'une gamme étendue de propriétés en décrivant l'énergie d'un système par la somme d'une série de contributions rendant compte des interactions intra et intermoléculaire. Pour chacune des contributions, des pénalités énergétiques sont appliquées lorsqu'une variable (par exemple une longueur de liaison ou un angle de valence) s'écarte de sa valeur de référence. L'ensemble de ces termes et des paramètres destinés à décrire chaque type d'atome rencontré constitue un champ de forces qui comporte généralement cinq contributions principales **[93]**.

$$E_{stériq} = E_{élongation} + E_{flexion} + E_{torsion} + E_{van\ der\ Waals} + E_{électro} \qquad (14)$$

Chacun de ces termes possédant une position d'équilibre préférentielle (longueur de liaison, angle de liaison...). La recherche de l'énergie minimale par optimisation de la géométrie joue un rôle primordial. L'énergie de la molécule est exprimée sous la forme d'une somme de contributions associées aux écarts de la structure par rapport à des paramètres structuraux de référence. A cet égard, la mécanique moléculaire ressemble aux modèles de type "tiges et boules", mais elle est beaucoup plus quantitative.

L'idée directrice de cette méthode est d'établir, par le choix des fonctions énergétiques et des paramètres qu'elles contiennent, un modèle mathématique, le **"champ de force"**, qui représente aussi bien que possible les variations de l'énergie potentielle avec la géométrie moléculaire. Cependant, il n'existe pas encore de modèle unique permettant de simuler tous les aspects du comportement moléculaire, mais un ensemble de modèles **[94]**. Alors pour atteindre un minimum local sur la surface de potentiel dans un temps minimum, il faut représenter toutes les variations possibles de l'énergie potentielle avec la géométrie moléculaire. Cependant, aucune méthode ne peut garantir de façon absolue l'obtention de la plus basse énergie : c'est-à-dire le minimum global ou absolu. **Presque toutes les méthodes de minimisation ont un point en commun : on commence à un endroit donné de l'hypersurface énergie-coordonnées et on accède au minimum local le plus proche**.

II. 2. 2. 1. Champ de force en mécanique moléculaire :

Pour trouver la géométrie optimum d'un ensemble d'atomes, il faut minimiser 3 coordonnées cartésiennes par atome (pour une protéine de 1000 atomes = 3000 coordonnées cartésiennes). Donc il faut trouver le minimum d'une fonction (l'énergie) dans un espace à quelques milliers de variables.

Le problème a donc consisté à choisir une fonction potentielle analytiquement simple qui représente ces coordonnées de la molécule. Cette fonction doit être simple pour pouvoir

être calculée rapidement, et suffisamment précise pour simuler de façon acceptable les propriétés structurales et thermodynamiques des macromolécules. Un champ de force est constitué de plusieurs fonctions d'énergie potentielle qui dérivent des interactions intramoléculaires entre atomes liés et non liés.

$$E_{Total} = \underbrace{E_{élongation} + E_{flexion} + E_{torsion}}_{E_{liés}} + \underbrace{E_{van\ der\ Waals} + E_{électro}}_{E_{non\text{-}lié}} + ... \qquad (15)$$

Une des principales difficultés de MM est de choisir le champ de force adéquat pour la modélisation du système moléculaire d'intérêt. La spécificité de chaque champ de force va dépendre du nombre de termes présents dans l'équation générale, termes qui augmentent en général avec la complexité du champ de force.

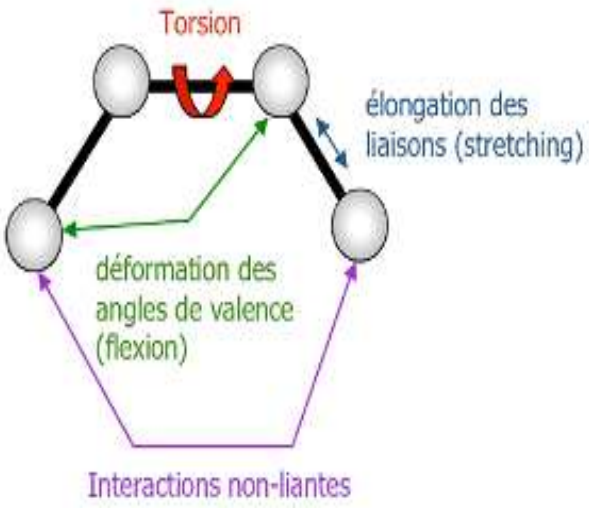


Figure II. 1 : Interactions intramoléculaires entre atomes liés et non liés.

II. 2. 2. 2. Energie d'interaction entre atomes liés :

Les potentiels utilisés sont du même type que ceux utilisés en analyse vibrationnelle. La déformation du squelette est décrite par l'élongation des liaisons et distorsion des angles de valence.

Ces déformations des liaisons et des angles sont représentées par un oscillateur harmonique avec une sommation sur toutes les liaisons covalentes ou sur tous les angles entre atomes liés par liaisons "covalentes" **[95]**.

II. 2. 2. 2. 1. Energie d'élongation (stretching) :

L'élongation des liaisons, $E_{stretch}$, est un terme destiné à réguler la distance entre deux atomes liés de façon covalente. A l'origine, cette contribution a été exprimée sous forme d'un potentiel de Morse ou, de manière à simplifier les calculs, par un potentiel harmonique (en x²) issu de la loi de Hooke qui décrit l'énergie associée à la déformation d'un ressort. Par analogie, les champs de force sont souvent comparés à des modèles assimilant les atomes à des boules reliées entre elles par des ressorts (figure II. 2).

$$E(L) = \frac{1}{2} K_r (L - L_0)^2 \qquad (16)$$

où :

K_r : est la constante d'élongation ou constante de Hooke.

L_o : la longueur de la liaison de référence.

L : la longueur de la liaison dans le modèle.

Figure II. 2 : Energie d'élongation entre deux atomes liés.

II. 2. 2. 2. 2. Energie de flexion (bending) :

La fluctuation des atomes autour de leur position d'équilibre engendre une déformation des angles de valence (figure II. 3).

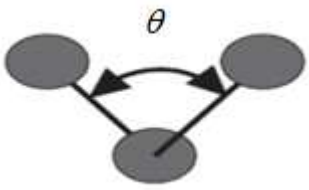


Figure II. 3 : Energie de déformation des angles de valence.

Ce phénomène est régi par une énergie de flexion qui peut s'exprimer sous les mêmes formes que précédemment à savoir, pour la plus simple :

$$E(\theta) = \tfrac{1}{2} K_\theta (\theta - \theta_0)^2 \quad (17)$$

K_θ : constante de flexion.

θ_o: angle de valence de référence.

Θ : angle de valence dans la molécule.

Le couple (K_θ, θ_o) représente ici encore un sous-ensemble du champ de force.

II. 2. 2. 2. 3. Energie de torsion :

La variation des angles dièdres (angles de torsion), $E_{torsion}$, impose l'utilisation d'un terme périodique. La torsion correspond à la rotation autour d'une liaison simple. L'énergie potentielle s'exprime en fonction de l'angle de rotation Φ (angle dièdre) (figure II. 4).

Le terme énergétique représentant la déformation des angles dièdres par une fonction développée en série de Fourier et il est calculé par la formule **[96]** :

$$E(\phi) = 1/2\left[V_1(1+\cos\phi) + V_2(1-\cos 2\phi) + V_3(1+\cos 3\phi)\right] \quad (18)$$

V_1, V_2, V_3 sont les constantes du potentiel de l'énergie de torsion.

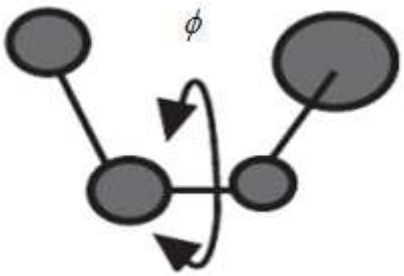


Figure II. 4 : Energie de torsion.

II. 2. 2. 3. Energie d'interaction entre atomes non liés :

Les deux derniers termes traduisent les interactions entre atomes non liés de manière covalente, seuls les atomes séparés de plus de trois liaisons sont considérés comme pouvant avoir des interactions non liées. Ce choix se justifie par le fait que les interactions, entre atomes séparés par une, deux ou trois liaisons, sont suffisamment bien décrites par les termes de déformation des longueurs de liaisons et d'angles formés par deux liaisons covalentes.

II. 2. 2. 3. 1. Energie de van der Waals :

Le quatrième terme de la fonction d'énergie est le potentiel de Lennard-Jones qui traduit les interactions de van der Waals. Il est constitué d'un terme répulsif en r_{ij}^{-12} représentant le principe d'exclusion de Pauli et la répulsion internucléaire à courte distance. Il contient également un terme attractif en r_{ij}^{-6} représentant les interactions d'origine dipolaire dues aux fluctuations des nuages électroniques de chacun des atomes (forces de London) **[97]**. La distribution asymétrique des électrons dans les orbitales induit des dipôles instantanés. Ces dipôles oscillent et produisent une force attractive : la force de dispersion de London. À très courte distance, la répulsion entre les deux nuages électroniques est très forte.

$$E_{ij} = \sum_i \sum_j -\frac{A_{ij}}{r_{ij}^6} + \frac{B_{ij}}{r_{ij}^{12}} \qquad (19)$$

Il s'agit d'une double somme sur tous les atomes chargés mais ne considérant que les paires d'atomes non-liés par des liens covalents (on exclut les paires prises en considération pour les liens covalents, les angles entre les liens et les angles de torsions).

r_{ij} : La distance entre les deux atomes.

A_{ij} et B_{ij} constantes de van der Waals.

Chaque atome possède un rayon de van der Waals caractéristique.

La distance de contact ou distance optimale entre deux atomes, c'est-à-dire la distance pour laquelle l'énergie de van der Waals est minimum, correspond à la somme de leurs rayons de van der Waals.

II. 2. 2. 3. 2. Interactions électrostatiques :

Enfin le cinquième terme est le potentiel coulombien qui traduit les interactions électrostatiques entre les différentes paires d'atomes **[78]**. Il s'agit des interactions entre deux atomes portant chacun une charge.

L'énergie de cette interaction est décrite par la loi de Coulomb :

$$E_{elect} = \sum \frac{q_i q_j}{D r_{ij}} \qquad (20)$$

q_i, q_j: sont les charges portées par les atomes.

r_{ij} : la distance entre les deux atomes.

D : la constante diélectrique du milieu.

Cette constante diélectrique est une propriété macroscopique du milieu environnant les charges ; son évaluation n'est pas aisée. Elle prend les valeurs 1,0 et 78,8 dans le vide et dans l'eau respectivement, alors que des valeurs de D entre 4,0 et 7,0 sont employées pour simuler un environnement de protéine **[67]**.

II. 2. 2. 3. 3. Energie de liaison hydrogène :

Les liaisons hydrogène sont le résultat des interactions électrostatiques (70%) et de van der Waals (30%) entre un atome électronégatif (généralement un atome d'oxygène ou d'azote) portant un doublet d'électrons libre et un atome d'hydrogène porté par un atome électronégatif. Les deux atomes sont distants d'environ 3 Å. L'énergie des liaisons hydrogène est de l'ordre de 3 $kcal.mol^{-1}$. Les glucides polaires peuvent ainsi former des liaisons hydrogène entre eux ou avec des molécules d'eau. Ils se dissolvent donc facilement dans l'eau : ils sont hydrophiles.

Les phénomènes de répulsion et de délocalisation électronique interviennent. Plusieurs types de fonctions d'énergie potentielle ont été développés pour tenir compte de la directivité de la liaison hydrogène. Actuellement, les fonctions les plus utilisées permettant d'exprimer ces interactions dans des systèmes moléculaires importants sont souvent simplifiées :

La fonction $E_H = A/r_{ij}^{12} - B/r_{ij}^{10}$

La fonction $E_H = A'/r_{ij}^{12} - B'/r_{ij}^{6}$

Les coefficients A, B, A', B' sont spécifiques des liaisons hydrogène **[78]**.

II. 2. 2. 4. Quelques champs de force :

Différents champs de force sont proposés dans la littérature, ils se distinguent les uns des autres par les termes dans le développement de l'expression de l'énergie de la molécule. Chacun a un domaine d'application spécifique de sorte que le choix d'un champ de force dépend des propriétés et de l'application du système que l'on veut étudier.

Type de composé : carbohydrate, complexe métallique.

Environnement : gaz, solution.

Type d'interaction à étudier : liaison hydrogène, ….

Les champs de force les plus répandus sont :

- **MM2, MM3 et MM4 :** Il a été développé par Allinger en 1976 et c'est le champ de force le plus utilisé par la communauté des chimistes organiciens **[98]**. Il a été conçu au début pour les molécules simples (alcanes, alcènes et alcynes non conjugués, les composés carbonylés, les sulfures, les amines…), mais ses versions améliorées MM3 **[99]** et MM4 **[100]** lui permet de traiter des molécules organiques de plus en plus complexes.
- **MM+ :** Est une extension du champ de force MM2, avec l'ajout de quelques paramètres additionnels **[101]**. MM+ est un champ de force robuste, il a l'aptitude de prendre en considération les paramètres négligés dans d'autres champs de force et peut donc s'appliquer pour des molécules plus complexes tels que les composés inorganiques **[102]**.
- **AMBER :** (Assisted Model Building with Energy Refinement), est un champ de force de mécanique moléculaire mis en point par Kollman **[103]**. Ce champ de force a été paramétré pour les protéines et les acides nucléiques. Il a été utilisé pour les polymères et pour d'autres petites molécules **[104]**.
- **CHARMM (Bio+) :** (Chemistry Harvard Macromolecular Mechanic), développé par Brooks B R et Karplus M **[105]**. Il utilise une fonction d'énergie empirique pour les systèmes macromoléculaires et molécules biologiques (protéines, acides nucléiques, …). Son concept est semblable à celui d'AMBER.

II. 3. Domaine d'application de la modélisation moléculaire :

On peut diviser l'application de la MM en trois catégories :

- Soit pour obtenir une géométrie à laquelle on attache de l'intérêt. Cette situation se présente lorsque la modélisation guide l'interprétation des résultats provenant des études de structure par rayons X ou par diffraction électronique, ou lorsqu'il s'agit de modéliser une molécule pour les besoins de l'infographie.
- Dans l'interprétation des effets stériques sur la réactivité ou bien de la stabilité relative des isomères en tant qu'énergie stérique ou de tension.
- Quand aucune liaison n'est rompue, ni formée et qu'aucun intermédiaire chargé n'intervient, l'interconversion conformationnelle se prête particulièrement bien à une description par la MM. On peut obtenir grâce à cette analyse des informations

structurales sous forme d'un profil énergétique (en fonction d'un angle dièdre par exemple) ou des cartes énergétiques 3D.

En conclusion, on peut dire que la mécanique moléculaire aujourd'hui est à la porte de tous les chercheurs.

La mécanique moléculaire ne peut pas encore rivaliser avec la mécanique quantique dans beaucoup de domaines qui lui sont propres, mais elle reste une méthode de choix dans l'interprétation de phénomènes sous contrôle stérique et dans le calcul de structure.

La mécanique moléculaire permet de passer en revue de grosses molécules (produits pharmaceutiques, colorants, etc.) pour établir des relations entre structure et réactivité et ainsi faire un tri avant de passer au stade expérimental. La différence du temps de calculs par la mécanique moléculaire par rapport aux autres méthodes quantiques est d'environ de quelques puissances de dix, cette différence augmente en fonction de la taille de la molécule.

Première partie : Etude bibliographique

III. Principes et méthodes de modélisations

III. 1. Relations Quantitatives Structures-Activités/Propriétés/Rétentions

III. 2. Les Descripteurs moléculaires

III. 3. Méthodes de sélection des ensembles de calibrage et de test

III. 4. Développement de modèles QSAR/QSPR/QSRR

III. 5. Paramètres d'évaluation de la qualité de l'ajustement

III. 6. Domaine d'application

III. 7. Les Logiciels utilisés dans nos études QSXR

III. Principes et méthodes de modélisations :

La connaissance des propriétés et des activités est d'une importance capitale pour pouvoir classer et utiliser les composés chimiques. La caractérisation expérimentale complète est difficile, voire impossible, pour des raisons de temps, de coût, de dangerosité de certains essais ou d'éthique (limitations des essais sur les animaux). L'utilisation des méthodes alternatives à l'expérience est devenue plus qu'indispensable. Parmi ces méthodes, on trouve les méthodes de modélisation moléculaire qui permettent de justifier les données expérimentales disponibles et prédire les propriétés/activités pour des composés nouveaux ou des composés pour lesquels les données expérimentales ne sont pas disponibles. Parmi ces méthodes de modélisation les plus utilisées, on peut citer les méthodes QSXR (X= Activité, Propriété, Rétention, Toxicité), QSAR (Quantitative Structure-Activity Relationships), QSPR (Quantitative Structure-Property Relationships), QSRR (Quantitative Structure-Retention Relationships) et QSTR (Quantitative Structure-Toxicity Relationships). Ces méthodes s'appuient sur le principe que les propriétés physico-chimiques et les activités biologiques des molécules dépendent fortement de leurs structures chimiques **[106]**.

III. 1. Relations Quantitatives Structure-Activité/Propriétés/Rétention :

III. 1. 1. Généralités sur la modélisation QSAR/QSPR/QSRR :

Une relation QSAR/QSPR/QSRR est un modèle ou une formule mathématique qui permet de relier, d'une manière quantitative, la structure d'une molécule à une propriété ou à une activité donnée. Les méthodes QSAR/QSPR/QSRR sont de plus en plus utilisées, du fait de la croissance des moyens de calculs (machines, logiciels, …). Récemment, on assiste à la mise en place d'un nouveau règlement européen REACH (Registration, Evaluation, Authorisation and Restriction of Chemicals) **[107]** qui recommande l'utilisation des méthodes alternatives pour limiter le recours à l'expérimentation.

En fait, les premiers développements dans le sens de telles méthodologies sont plutôt anciens. Dès 1868, Crum-Brown et Fraser ont postulé l'existence de relations entre les activités physiologiques et les structures chimiques en reliant les changements d'activité biologique à des modifications structurales simples, ne disposant alors pas de moyen pour caractériser les structures chimiques en termes quantitatifs.

Hansch et Fujita établirent, en 1964, les premières corrélations entre les propriétés physico- chimiques (logP, pKa, paramètres stériques et électroniques) et l'activité biologique

(activités enzymatiques, pharmacologiques) **[108]**. En 1971, ils réalisent une étude de relation structure- activité sur différentes familles d'antifongiques : benzoquinone, sels d'alkylpyridinium, imidazoles et phénols. Ils observent que quels que soient la famille et le champignon utilisé, l'activité antifongique dépend du logP (coefficient de partage octanol-eau) expérimental ou calculé **[109]**.

Ces études ont été extrapolées aux techniques séparatives en corrélant les propriétés physico- chimiques des analytes avec les temps de rétention obtenus expérimentalement : c'est l'étude quantitative des relations structure temps de rétention noté QSRR **[110]**.

Au cours des décennies passées, les Relations Quantitatives Structure-Activité/ Propriétés/Rétention/Toxicité (QSAR/QSPR/QSRR/QSTR) sont devenues un puissant outil théorique, alternatif à la mécanique quantique, pour la description et la prédiction des propriétés des systèmes moléculaires complexes dans différents environnements.

III. 1. 2. Principe des méthodes QSAR/QSPR/QSRR :

Le principe des méthodes QSAR/QSPR/QSRR est d'établir une relation mathématique reliant de manière quantitative des propriétés moléculaires, appelées descripteurs, avec une observable macroscopique (activité biologique, toxicité, propriété physico-chimique, etc.), pour une série de composés chimiques similaires à l'aide de méthodes d'analyses de données. La forme générale d'un tel modèle est la suivante :

$$\text{Activité/Propriété} = f\left(D_1, D_2, \ldots D_n, \ldots\right) \qquad (21)$$

D_1, D_2, …D_n sont des descripteurs des structures moléculaires.

L'objectif d'une telle méthode est d'analyser les données structurales afin de détecter les facteurs déterminants pour la propriété /activité mesurée. Pour ce faire, différents types d'outils statistiques peuvent être employés :

- Régressions linéaires simples et multiples **[111]**
- Régressions aux moindres carrés partielles (PLS) **[112]**
- Arbres de décision **[113]**
- Réseaux de neurones **[114-116]**
- Algorithmes génétiques **[117]**
- Vecteurs Machines **[116]**

Une fois cette relation établie et validée, elle peut alors être employée pour la prédiction de la propriété /activité de nouvelles molécules, pour lesquelles les valeurs expérimentales ne sont pas disponibles. De tels modèles peuvent être également utilisés pour mieux comprendre les mécanismes et les modes d'action.

III. 1. 3. Méthodologie générale d'une étude QSXR :

La méthodologie générale d'une étude QSXR est la suivante :

a- Constituer une base de données à partir des mesures expérimentales fiables de la propriété ou de l'activité de chaque composé.

b- Sélectionner les descripteurs en relation avec la propriété ou l'activité étudiée.

c- Diviser cette base de données, aléatoirement, selon l'ordre de réponses, ou par choix orienté selon l'algorithme Kennard et Stone (CADEX) **[118]** (le choix utilisé dans cette thèse) ou selon l'algorithme DUPLEX, en un ensemble d'apprentissage (training set) qui contient généralement les 2/3 de la base de données et un ensemble de test (test set) constitué par le 1/3 restant.

d- Etablir des modèles mathématiques en utilisant la série d'apprentissage.

e- Caractériser les modèles élaborés par leurs indices de validation internes et vérifier leur robustesse par un test d'hasardisation (randomisation) de la variable dépendante Y (réponse).

f- Valider les modèles élaborés en utilisant la série de tests et calculer leurs paramètres statistiques de validation externe.

g- Elaborer le domaine d'applicabilité du modèle retenu.

h- Explorer et exploiter les modèles validés pour comprendre les mécanismes et les modes d'action.

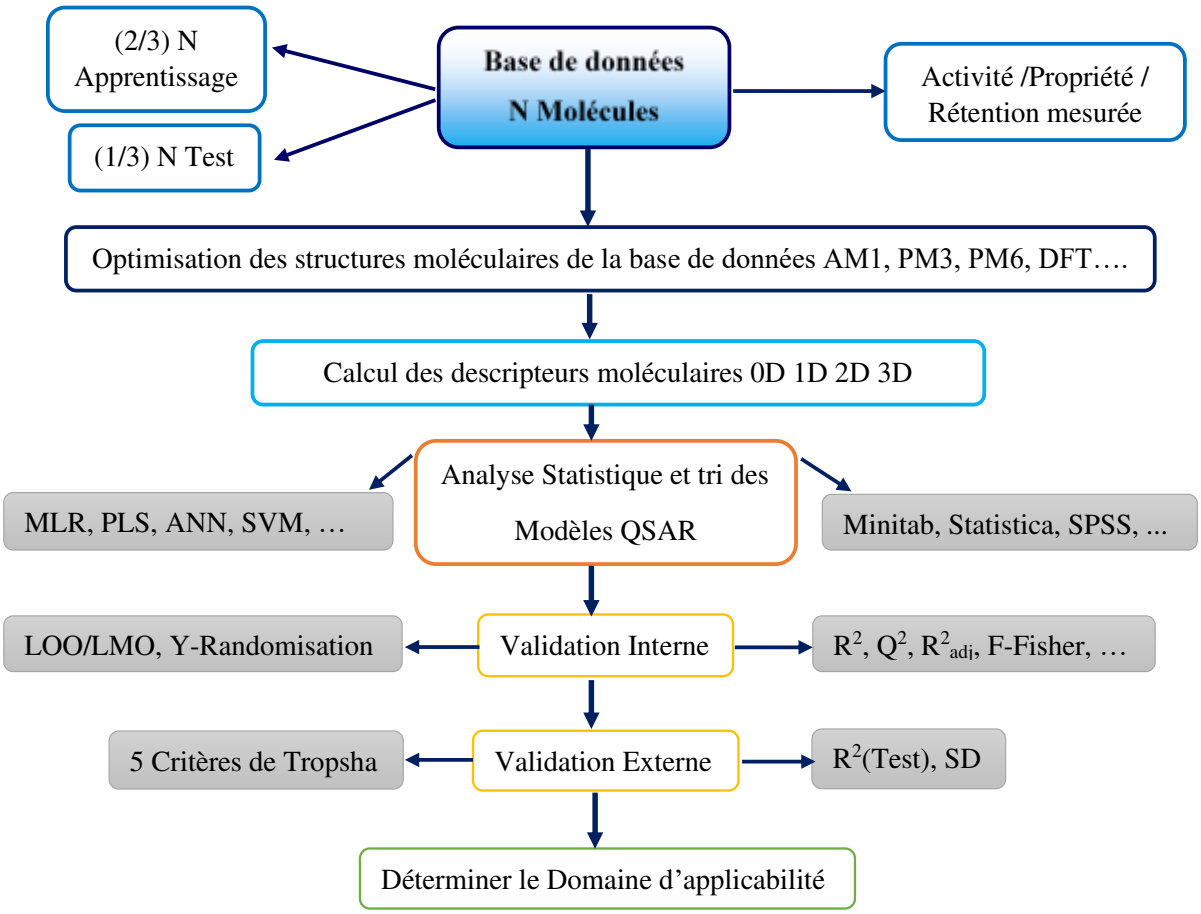


Figure III. 1. Méthodologie générale d'une étude QSXR.

III. 1. 4. Importance de la base de données :

Un modèle QSRR est dépendant des données expérimentales de référence, le choix de la base de données est un point critique de son développement. Dans la plupart des cas, les données expérimentales sont issues de la littérature, soit des produits de synthèse ou bien des produits d'extraction à partir de plantes.

Pour être de qualité, une base de données doit être composée de données expérimentales aussi fiables que possible, puisque les barres d'erreurs sur celles-ci se propageront dans le modèle final, étant donné que les paramètres de ce dernier sont ajustés par rapport à ces données. Il est donc important de choisir des données présentant des incertitudes faibles, afin de limiter les barres d'erreur expérimentales. De plus, les données doivent être obtenues suivant un protocole expérimental unique. En effet, les conditions expérimentales ont une forte influence sur les valeurs obtenues. La définition de la propriété en termes de conditions expérimentales est d'ailleurs un point important de la démarche **[4]**.

III. 2. Les Descripteurs moléculaires :

Toute molécule peut être décrite par des valeurs que l'on appelle descripteurs. Ces valeurs sont le résultat d'un ensemble de procédures mathématiques qui transforment l'information chimique, codée dans une représentation symbolique de la molécule, en un nombre utilisable. Toutefois, on utilise dans quelques cas des mesures expérimentales des propriétés en tant que descripteurs. Au cours des dernières décennies, la recherche a lancé un challenge pour développer des descripteurs capables de décrire les structures moléculaires de la manière la plus exhaustive possible. On dénombre aujourd'hui des milliers de descripteurs qui sont calculés par des logiciels spécifiques disponibles **[119, 120]**. Les descripteurs moléculaires sont fréquemment classés par rapport à la dimensionnalité de la représentation moléculaire sur laquelle ils sont calculés **[121]**.

III. 2. 1. Descripteurs 0D :

Les plus simples des descripteurs sont ceux de dimension zéro (0 D). Ils sont directement dérivés de la formule brute de la molécule, ces descripteurs constitutionnels sont :

- La masse molaire
- Les nombres absolus et relatifs d'atomes (C, H, O, S, N, F, Cl, Br, I, P…)
- Les nombres absolus et relatifs de liaisons (simples, doubles, aromatiques. . .)
- Les nombres absolus et relatifs de cycles (aromatiques ou non)
- etc.

On peut remarquer que ces descripteurs ne permettent pas de distinguer les isomères de constitution ; cependant si on développe des modèles avec seulement ce type de descripteurs, on tomberait alors dans le problème d'interprétation des mécanismes d'interaction mis en jeu pour la propriété étudiée. En dépit de cela ces descripteurs sont très utilisés du fait de leur extrême simplicité non seulement d'un point de vue conceptuel mais surtout calculatoire.

III. 2. 2. Descripteurs 1D :

Les descripteurs unidimensionnels (1D) reflètent des propriétés générales c'est-à-dire ils mentionnent l'absence ou la présence de certains éléments spécifiques d'une molécule, tels que :

- Les groupes fonctionnels
- Les fragments des atomes centraux
- etc…

III. 2. 3. Descripteurs 2D :

Les descripteurs bidimensionnels (2D) se réfèrent à la façon dont les atomes sont connectés dans la molécule. Cependant ils caractérisent la structure moléculaire selon sa taille, son degré de ramification, et sa forme générale. Parmi ces descripteurs on a :

- Les indices de formes et de connectivité : indices de Kier et Hall **[122]**
- Les descripteurs topologiques : matrices de distance et de connectivité **[123]**
- Les descripteurs de charges partielles : charge partielle positive (négative) totale, aire de la surface de van der Waals positive (négative), aire de la surface de van der Waals polaire (hydrophobe) …
- etc... .

III. 2. 4. Descripteurs 3D :

Les descripteurs tridimensionnels (3D) obtenus de la géométrie moléculaire (3D) décrivent des objets tridimensionnels et se répartissent en 2 groupes : ceux qui ne dépendent que des coordonnées internes de la molécule et ceux qui dépendent de son orientation absolue **[124]**. Parmi tous ces descripteurs on peut citer :

- Les descripteurs de l'énergie potentielle : valeur de l'énergie potentielle et composantes de cette énergie : van der Waals, électrostatiques, atomes « hors-du plan », torsion, etc...
- Des descripteurs de formes et de volumes : aire et volume de van der Waals, surface accessible au solvant, moment d'inertie, globularité (molécule sphérique, plane ou linéaire)
- Des descripteurs du moment dipolaire : orientation et intensité
- Les descripteurs électroniques : densité électronique, distribution de charges
- Les descripteurs quantiques basés sur les orbitales moléculaires
- etc... .

La géométrie moléculaire doit être optimisée d'abord à un certain niveau approprié selon la théorie de la mécanique quantique **[125]**.

III. 3. Méthodes de sélection des ensembles de calibrage et de test :

La sélection d'échantillons représentatifs est une étape importante dans une procédure d'élaboration de modèles QSAR/QSPR/QSRR. En effet, si les jeux d'étalonnage et de validation ne couvrent pas les mêmes domaines de variation, la validation du modèle ne sera pas correcte. Les échantillons d'étalonnage doivent donc répondre à certains critères ; on identifie 3 règles d'optimalité pour les échantillons de calibrage :

- les échantillons retenus doivent présenter une variabilité maximale ;

- la plage de variation des valeurs doit être la plus grande possible, mais limitée aux valeurs rencontrées dans la pratique ;

- les échantillons doivent être uniformément répartis.

Dans La littérature, cette tâche a été exécutée en utilisant beaucoup et différentes méthodes de sélection d'échantillons, chacune avec ses avantages et ses inconvénients. Plusieurs méthodes de sélection d'échantillons (algorithme de Kennard-Stone, algorithme DUPLEX, sélection aléatoire des échantillons, OPTISIM, Répartition uniforme des échantillons sur la variable dépendante, etc...) peuvent être utilisées **[126]**.

III. 3. 1. Algorithme CADEX :

C'est une technique séquentielle qui maximise les distances euclidiennes entre les nouveaux échantillons sélectionnés et ceux qui le sont déjà. Elle commence par situer les deux échantillons les plus éloignés l'un de l'autre, qui sont retirés de la base de données initiales et affectés à l'ensemble de calibrage.

Pour chaque échantillon non sélectionné (éch i), l'algorithme :

- calcule la distance vers chaque échantillon déjà sélectionné ;

- attribue à (éch i) la plus petite des distances.

L'échantillon (éch i) associé à la plus grande distance est donc le plus éloigné de tous les échantillons déjà sélectionnés ; c'est donc lui qui est sélectionné.

La procédure est répétée jusqu'à l'obtention du nombre d'échantillons désirés pour l'ensemble de calibrage. Le fait de sélectionner les échantillons les plus éloignés les uns des autres introduit une grande diversité dans l'ensemble de calibrage ; l'obtention d'une répartition uniforme est un autre avantage de cette technique **([118]**.

III. 3. 2. Algorithme DUPLEX :

Une version améliorée appelée DUPLEX a été proposée par Snee **[127]** ; il est largement utilisé dans le domaine de la chimiométrie, y compris plusieurs applications ANN **[128, 129]**. Cependant, la complexité de calcul de cet algorithme peut interdire son utilisation sur de grands ensembles de données. Par ailleurs, selon un travail récent de Ren *et al.* **[130]**, DUPLEX est l'une des meilleures méthodes pour diviser les données en un ensemble d'apprentissage et un ensemble de test, qui mesure la distance entre tous les échantillons par la distance euclidienne.

Cet algorithme commence avec la liste des n observations, les ℓ régresseurs étant standardisés à l'unité selon :

$$z_{ij} = \frac{x_{ij} - \bar{x}_j}{s_j \sqrt{n-1}} \quad i = 1,\ldots n \;;\; j = 1, .., \ell \tag{22}$$

Où

s_j : Ecart-type du j ème régresseur.

$\bar{x}_j$: Moyenne du j ème régresseur.

x_{ij} : Valeur du régresseur j pour la i ème observation.

n : Nombre d'observations.

Les régresseurs standardisés sont alors orthonormalisés en factorisant le produit à gauche de la matrice $\mathbf{Z} = \left(z_{ij}\right)$ par sa transposée $\mathbf{Z'}$, sous la forme :

$$\mathbf{Z'.Z} = \mathbf{T'.T} \tag{23}$$

$\mathbf{T}$ est une matrice (ℓ × ℓ) triangulaire supérieure unique, dont les éléments peuvent être obtenus par la méthode de Cholesky **[131]**. On opère alors la transformation :

$$\mathbf{W} = \mathbf{Z.T^{-1}} \tag{24}$$

qui conduit à un nouvel ensemble de variables w orthogonales et de variance unité. Celles-ci sont utilisées pour calculer la distance euclidienne, entre les C_n^2 paires de points. Les 2 points les plus éloignés sont sélectionnés pour l'ensemble de calibrage, puis parmi les points restants, les 2 plus éloignés sont sélectionnés pour la validation (ensemble de test). Puis parmi les points restants, le plus éloigné des points de calibrage précédemment sélectionnés est

sélectionné pour le calibrage. Puis parmi les points restants, le plus éloigné des points de validation précédemment sélectionnés est sélectionné pour la validation. Puis l'algorithme continue à placer les points restants, alternativement dans l'ensemble de calibrage et dans l'ensemble de validation, jusqu'à ce que les n points soient affectés. Les ensembles de calibrage et de validation n'étant pas forcément de même taille, l'algorithme DUPLEX peut séparer les données dans n'importe quel rapport souhaité. De telles séparations sont réalisées en utilisant l'algorithme jusqu'à ce que l'ensemble de validation contienne le nombre de points requis, puis en versant les points non assignés dans l'ensemble de calibrage. L'utilisation de l'algorithme DUPLEX suppose que le nombre d'observations, n, est tel que : $n \geq 2\ \ell + 25$, ℓ désignant le nombre de régresseurs ; l'ensemble de validation devant contenir 15 éléments au minimum.

Par conséquent, il garantit que la composition de l'ensemble de calibrage et de l'ensemble de test ne présente pas, en même temps, un déséquilibre des deux ensembles de données **[132]**.

III. 3. 3. Choix aléatoire :

L'échantillonnage aléatoire simple (au hasard) est la méthode la plus courante pour le fractionnement des données dans le développement des modèles, où les données sont sélectionnées avec une probabilité uniforme. L'échantillonnage au hasard simple est facile à réaliser et peut être efficacement exécuté dans un seul passage sur les données en utilisant des algorithmes tels que l'algorithme de Knuth**[133]**. Cependant, le problème avec cette approche est qu'il y a une chance que la scission de données souffre de la variance, ou de partialité, en particulier lorsque les données ne sont pas réparties uniformément **[134]**.

III. 4. Développement de modèles QSAR/QSPR /QSRR:

III. 4. 1. Sélection d'un sous-ensemble de variables par algorithme génétique (GA-VSS) :

Les algorithmes génétiques fournissent des solutions aux problèmes n'ayant pas de solutions calculables en temps raisonnable de façon analytique ou algorithmique. Selon cette méthode, des milliers de solutions (génotypes) plus au moins bonnes sont créées au hasard puis sont soumises à un procédé d'évaluation de la pertinence de la solution mimant l'évolution des espèces : les plus "adaptés", c'est- à- dire les solutions au problème qui sont

les plus optimales survivent davantage, que celles qui le sont moins et la population évolue par générations successives en croisant les meilleures solutions entre elles et les faisant muter, puis en relançant ce procédé un certain nombre de fois afin d'essayer de tendre vers la solution optimale.

Les algorithmes génétiques constituent une méthode de choix pour la sélection de sous- ensembles de variables explicatives.

Dans un algorithme génétique adapté à l'optimisation, une solution potentielle est considérée comme un individu dans une population. La valeur de la fonction de coût associée à une solution mesure « l'adaptation » de l'individu associé à son environnement. Un algorithme génétique simule l'évolution, sur plusieurs générations, d'une population initiale dont les individus sont mal adaptés au moyen d'opérateurs génétiques de reproduction et de mutation. Après un certain nombre de générations, la population est constituée d'individus bien adaptés, autrement dit des solutions supposées « bonnes » au problème d'optimisation.

Dans le présent travail, la sélection des descripteurs a été réalisée par algorithme génétique, dans la version MobyDigs de Todeschini **[135]**, en maximisant Q^2_{LOO}.

III. 4. 2. Méthodes utilisées pour le développement de modèles QSAR/QSPR/QSRR

L'application pratique des gammes des descripteurs moléculaires dans le développement de modèles QSAR/QSPR/QSRR n'est pas une tâche aisée **[136]**. Tout d'abord, un très grand nombre (>10000) de descripteurs moléculaires, de différentes complexités et de conceptions diverses ont été imaginés et proposés au cours des (60 dernières) années. Ensuite, pendant ce temps, aucune règle stricte n'a été établie, ni même proposée, pour la sélection de descripteurs adaptés parmi la myriade de descripteurs disponibles. Ce choix a souvent été basé sur l'intuition chimique des chercheurs, ou en se pliant à la tradition.

Une autre difficulté dans la sélection des descripteurs QSAR/QSPR/QSRR découle de la non standardisation des gammes de descripteurs. Les gammes empiriques des constantes d'induction, de résonance et d'effet stérique des constituants, ou les échelles empiriques d'effets de solvant comportent des erreurs intrinsèques liées aux erreurs respectives des mesures expérimentales. Par ailleurs, les méthodes quanto- mécaniques appliquées aux calculs des descripteurs moléculaires et aux distributions de charges liés aux OM sont souvent basées sur différents paramètres semi- empiriques, ou l'utilisation de différents ensembles de

base dans les calculs *ab-initio*. Naturellement, un descripteur construit à l'aide de différentes méthodes expérimentales ou théoriques, pour divers composés, ne peut être utilisé pour le calcul d'un modèle QSAR/QSPR unique. Une approche systématique pour la sélection de gammes de descripteurs pour le calcul de modèles QSAR/QSPR/QSRR est basée sur la discrimination statistique entre de larges ensembles de descripteurs.

Dans ce qui suit nous passerons en revue diverses approches utilisées pour le développement des " meilleures" équations QSRR dans de grands espaces de descripteurs.

En dernier ressort, les modèles QSAR/QSPR/QSRR peuvent être développés selon des modèles mathématiques différents, généralement en relation avec l'analyse statistique multivariée. Le premier modèle, et le plus largement utilisé, consiste en une équation (multi) linéaire obtenue par régression des données expérimentales en fonction d'un ensemble de descripteurs pré- sélectionnés (ou d'un seul), en utilisant la méthode des moindres carrés ordinaires (MCO). Dans quelques cas, les modèles physiques ou chimiques connus du phénomène étudié laissent prévoir certaines formes mathématiques non linéaires (exponentielles ou logarithmiques) de la dépendance entre les données expérimentales et les descripteurs moléculaires. Les modèles QSAR/QSPR/QSRR peuvent alors être établis à l'aide de la technique de régression par les moindres carrés non linéaires. D'autres modèles ont été développés en utilisant l'analyse factorielle ou l'analyse en composantes principales. L'intérêt de ces méthodes est qu'elles évacuent le problème de multicolinéarité inhérent aux méthodes de régression linéaires. Cependant, l'interprétation des équations QSAR/QSPR/QSRR est alors entravée par la nature formelle des composantes principales. Une alternative aux méthodes très classiques de régression linéaire multiple (RLM) et d'analyse en composantes principales (ACP) est la technique de régression par les moindres carrés partiels (MCP ou PLS) **[112, 137-141]**.

On a également appliqué les méthodes modernes de l'intelligence artificielle au développement de modèles QSAR/QSPR/QSRR **[142-146]**. Ces méthodes comprennent : les Réseaux de neurones (RNA), Machines à vecteurs supports (SVM), et d'autres méthodes globales d'optimisation.

Nous présenterons dans ce qui suit une courte vue d'ensemble des différentes méthodes mathématiques utilisées pour développer notre modèle.

III. 4. 2. 1. La régression linéaire multiple :

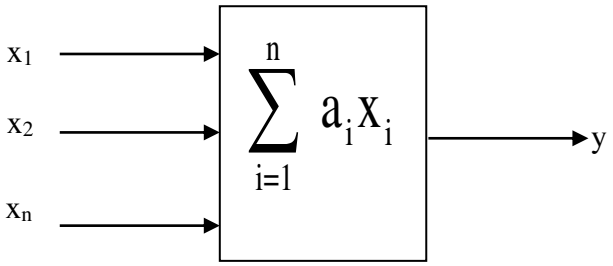
L'étude d'un phénomène peut, le plus souvent, être schématisée de la manière suivante : on s'intéresse à une grandeur y, que nous appellerons par la suite réponse ou variable expliquée, qui dépend d'un certain nombre de variables x_1; x_2; ….x_n que nous appellerons facteurs ou variables explicatives.

La régression est une des méthodes les plus connues set les plus appliquées en statistique pour l'analyse de données quantitatives. Elle est utilisée pour établir une liaison entre une variable quantitative et une ou plusieurs autres variables quantitatives, sous la forme d'un modèle. Si on s'intéresse à la relation entre deux variables, on parlera de régression simple en exprimant une variable en fonction de l'autre. Si la relation porte entre une variable et plusieurs autres variables, on parlera de régression multiple. La mise en œuvre d'une régression impose l'existence d'une relation de cause à effet et entre les variables prises en compte dans le modèle **[147]**.

La régression multi-linéaire (MLR, pour Multiple Linear Regression) **[148]** est la méthode la plus simple et la plus communément employée pour le développement de modèles prédictifs. Elle repose sur l'hypothèse qu'il existe une relation linéaire entre une variable dépendante y (ici, la propriété) et une série de n variables indépendantes xi (ici, les descripteurs). L'objectif est d'obtenir une équation de la forme suivante :

$$y = a_0 + a_1x_1 + a_2x_2 + \ldots + a_nx_n \qquad (25)$$

où a_i sont les coefficients de la régression.



La détermination de l'équation (26) se fait alors à partir d'une base de données de p échantillons pour laquelle à la fois les variables dépendantes et la variable indépendante sont connues. Il s'agit donc de considérer un système de p équations.

$$\begin{aligned}\hat{y}_1 &= a_0 + a_1 x_{1,1} + a_2 x_{2,1} + \ldots + a_n x_{n,1} + \varepsilon_1 \\ \hat{y}_2 &= a_0 + a_1 x_{1,2} + a_2 x_{2,2} + \ldots + a_n x_{n,2} + \varepsilon_2 \\ \hat{y}_p &= a_0 + a_1 x_{1,p} + a_2 x_{2,p} + \ldots + a_n x_{n,p} + \varepsilon_p\end{aligned} \quad (26)$$

où les résidus ε_i représentent l'erreur du modèle, constituée par l'incertitude sur la variable dépendante y_i d'une part, sur les variables indépendantes x_i d'autre part, mais aussi par les informations contenues dans les variables indépendantes mais non exprimées via les variables dépendantes.

Ce système d'équations peut être écrit sous la forme matricielle suivante :

$$\begin{pmatrix} y_1 \\ . \\ . \\ . \\ y_p \end{pmatrix} = \begin{pmatrix} 1 & x_{1,1} & \ldots & x_{n,1} \\ . & . & . & . \\ . & . & . & . \\ . & . & & . \\ 1 & x_{1,p} & \ldots & x_{1,p} \end{pmatrix} \begin{pmatrix} a_0 \\ . \\ . \\ . \\ a_p \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ . \\ . \\ . \\ \varepsilon_p \end{pmatrix} \quad (27)$$

soit de manière condensée :

$$\mathbf{Y} = \mathbf{XA} + \varepsilon \quad (28)$$

La méthode consiste alors à choisir les coefficients du vecteur **A** en faisant en sorte de minimiser la somme des carrés des écarts entre les valeurs prédites et les valeurs réelles sur l'intégralité de la base de données et ceci sous couvert de certaines hypothèses de départ.

En premier lieu, les variables indépendantes xi, comme leur nom l'indique, sont supposées indépendantes entre elles et leur incertitude est négligeable. Ensuite, les différents échantillons yi sont supposés indépendants entre eux et suivent une distribution normale.

L'erreur ε est elle-même supposée suivre une distribution normale, centrée en 0. Enfin, par nature, la dépendance de y vis-à-vis des xi est supposée linéaire.

La valeur prédite de la variable dépendante est alors :

$$\hat{y}_i = \hat{a}_0 + \hat{a}_1 x_{1,i} + \ldots + \hat{a}_n x_{n,i} \quad (29)$$

Les résidus peuvent donc être définis comme la différence entre les valeurs prédites et observées de y.

$$\varepsilon_i = y_i - \hat{y}_i \quad (30)$$

Il s'agit alors de trouver les coefficients $\hat{a}_i$ afin de minimiser la somme des carrés de ces résidus pour l'intégralité de la base de données.

$$\min\left[\sum(\varepsilon_i)^2\right] = \min\left[\sum(y_i - \hat{y}_i)^2\right] = \min\left[\sum\left(y_i - \hat{a}_0 - \hat{a}_1 x1_{,i} - \ldots - \hat{a}_n x_{n,i}\right)^2\right] = \min\left(Y - X\hat{A}\right)^T\left(Y - X\hat{A}\right) \quad (31)$$

Les coefficients peuvent être obtenus à partir de l'équation matricielle suivante :

$$\hat{A} = \left(X^T X\right)^{-1} X^T Y \quad (32)$$

Bien entendu, la régression multi-linéaire souffre de certains désavantages. Le principal découle de sa linéarité. Elle est donc défaillante pour la mise en évidence de dépendances non-linéaires. Cela dit, elle n'en reste pas moins une méthode simple et efficace dans la plupart des cas. De plus, pour peu que les variables indépendantes soient choisies de manière raisonnée, les équations obtenues peuvent être interprétées d'un point de vue phénoménologique **[149]**.

III. 4. 2. 2. Méthode des réseaux de neurones artificiels : [150-152]

Dans cette technique on cherche à s'inspirer du traitement de l'information effectuée par le cerveau humain où le neurone est l'unité fonctionnelle de base du système nerveux. Il est constitué principalement de trois parties qui ont un rôle bien défini : les dendrites qui collectent les signaux en provenance d'autres neurones et les font converger vers le soma, ce dernier recueille et traite l'information et l'axone qui transmet le signal traité vers l'extérieur.

Ainsi, les RNA sont des systèmes de calculs qui imitent les systèmes biologiques par l'utilisation des interconnections entre de simples neurones artificiels. D'un point de vue technique, chaque neurone est connecté à d'autres par des liens directs.

Chaque lien est associé à un poids qui représente l'information utilisée par le réseau pour résoudre le problème. La sortie du neurone est calculée à l'aide de l'équation :

$$O_j = f\left(\sum_{i=1}^{n} w_{ij} x_i + b_j\right) \quad (33)$$

Où O_j est la sortie du $j^{ème}$ neurone, f est la fonction d'activation, b_j est l'erreur du $j^{ème}$ neurone, w_{ij} le poids synaptique du synapse i (signal d'entrée) vers le neurone j et n le nombre de signaux d'entrée du $j^{ème}$ neurone (Figure III. 2 (a)). L'apprentissage du RNA se fait en ajoutant les poids de connexion en utilisant des algorithmes qui minimisent l'écart entre la

sortie obtenue et la sortie désirée. Il est, donc, évident que la rapidité et la qualité de l'apprentissage des RNA dépendent de la fonction d'activation *f*, de la structure du réseau et, bien évidemment, de la méthode de correction de l'erreur *bj* en réévaluant les poids w_{ij}. Généralement les neurones sont arrangés en couches pour donner un réseau. Ce dernier a une couche d'entrée et un ou plusieurs couches cachées.

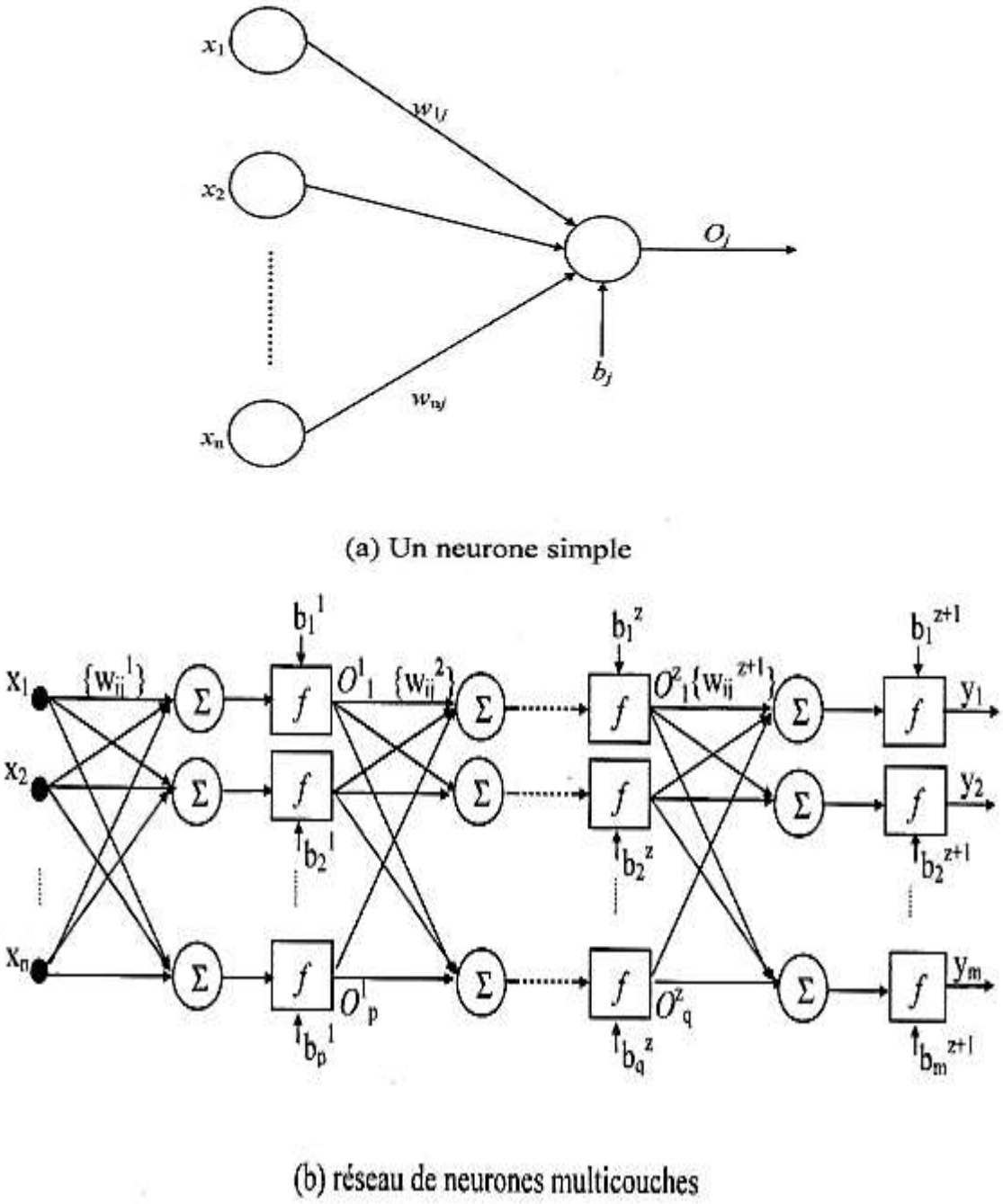


Figure III. 2. Schéma fonctionnel d'un réseau de neurones.

III. 4. 2. 3. Machines à vecteurs supports SVM :

La régression par Machines à Vecteurs de Support (SVM) **[153]** consiste à trouver la fonction f(x) qui a au plus une déviation " par rapport aux exemples d'apprentissage (xi; yi), pour i = 1,…, N, et qui est la plus plate possible. Cela revient à ne pas considérer les erreurs inférieures à ε et à interdire celles supérieures à ε **[154]**. Maximiser la platitude de la fonction permet de minimiser la complexité du modèle qui influe sur ses performances en généralisation. En effet, la théorie de l'apprentissage **[153]** permet de borner l'erreur de généralisation par une somme de deux termes : l'un dépendant de la complexité du modèle et l'autre dépendant de l'erreur sur les données d'apprentissage **[155]**. Les méthodes SVMs sont basées sur le contrôle de la complexité du modèle lors de l'apprentissage.

Dans la méthode SVM, différents hyperparamètres apparaissent : C, qui représente le compromis entre la complexité du modèle et l'erreur sur les données d'apprentissage ; λ, qui correspond à la largeur du tube d'insensibilité ; les éventuels paramètres de la fonction noyau k (σ,γ, ...). Ces hyperparamètres sont en général réglés en fonction d'une estimation de l'erreur de généralisation qui peut être évaluée sur un jeu indépendant de données de validation ou par validation croisée **[156]**. Cela implique de réaliser l'apprentissage pour différentes valeurs et d'estimer leurs performances. Dans le cas d'une estimation de l'erreur de généralisation par validation croisée, cette procédure peut se révéler très coûteuse en temps de calcul.

III. 4. 2. 4. Corrélation valeurs calculées – valeurs mesurées : [157]

III. 4. 2. 4. 1. Rappels de statistiques :

L'espérance mathématique d'une variable aléatoire discrète x, notée E(x), est définie par la relation.

$$E(x)=\sum_i x_i p_i \tag{34}$$

où p_i est la probabilité associée au résultat x_i.

La dispersion de la variable aléatoire x est décrite par la variance V(x) définie par la relation :

$$V(x)=\sum_i \left[x_i - E(x) \right]^2 p_i \tag{35}$$

qui peut être réécrite

$$V(x)=E(x)\left[x-E(x)\right]^2 p_i \tag{36}$$

la racine carrée de la variance définit l'écart-type :

$$\sigma_j=\left[V(x)\right]^{0,5} \tag{37}$$

Considérons maintenant les écarts x-E(x) et y-E(y). Si la probabilité pour que l'écart x-E(x) admette une valeur donnée n'est en aucune façon altérée par la valeur prise par l'écart y-E(y), et réciproquement, les variables x et y sont dites statistiquement indépendantes.

Si, par ailleurs, la probabilité pour un certain écart x-E(x) est altérée par la valeur prise par y-E(y), et réciproquement, les variables x et y sont statistiquement dépendantes.

La covariance de deux variables aléatoires est couramment utilisée comme mesure de l'association statistique. Formellement la covariance des variables aléatoires x et y, notée cov(x,y), est définie comme l'espérance mathématique du produit de [x-E(x)] par [y-E(y)].

$$cov(x,y)=E\left[x-E(x)\right]E\left[y-E(y)\right] \tag{38}$$

Le coefficient de corrélation de x et y, noté r(x,y) est défini par :

$$r(x,y)=\frac{Cov(x,y)}{\sigma_x.\sigma_y} \tag{39}$$

III. 4. 2. 4. 2. Estimation de la droite de régression :

Les mécanismes de rétention chromatographique n'étant pas établis de manière définitive, il n'est pas sûr qu'un modèle puisse englober tous les paramètres de rétention essentiels, ce qui affectera les valeurs calculées qui présenteront des degrés de précision bien moindres que ceux des valeurs expérimentales. Pour cette raison nous choisirons les valeurs mesurées comme variables indépendantes, c'est- à- dire comme abscisses, assumant implicitement qu'elles sont connues avec une meilleure précision que les valeurs calculées.

Les calculs se font à partir d'un échantillon formé par l'ensemble des paires de valeurs données par :

$$\begin{matrix} x_1\ x_2\ x_3\ \ldots.\ x_N \\ y_1\ y_2\ y_3\ \ldots.\ y_N \end{matrix} \tag{40}$$

x_i : variable indépendante (valeur mesurée – expérimentale).

y_i : variable dépendante (valeur fournie par le modèle).

Il s'agit d'établir la droite de régression :

$$Y = a + bx \tag{41}$$

représentant les paires de valeurs x_i/ y_i par la méthode des moindres carrés. Le calcul des paramètres a et b et l'analyse de variance qui s'ensuit nécessitent les grandeurs suivantes **[158]** :

$$T_x = \sum_{i=1}^{N} x_i \quad \text{(somme des } x_i\text{)} \tag{42}$$

$$T_y = \sum_{i=1}^{N} y_i \quad \text{(somme des } y_i\text{)} \tag{43}$$

$$\overline{x} = T_x / N \quad \text{(moyenne de } x_i\text{)} \tag{44}$$

$$\overline{y} = T_y / N \quad \text{(moyenne de } y_i\text{)} \tag{45}$$

$$S_{xx} = \sum_{i=1}^{N} \left(x_i - \overline{x}\right)^2 = \sum_{i=1}^{N} x_i^2 - \frac{T_x^2}{N} \tag{46}$$

$$S_{yy} = \sum_{i=1}^{N} \left(y_i - \overline{y}\right)^2 = \sum_{i=1}^{N} y_i^2 - \frac{T_y^2}{N} \tag{47}$$

$$S_{xy} = \sum_{i=1}^{N} \left(x_i - \overline{x}\right)\left(y_i - \overline{y}\right) = \sum_{i=1}^{N} x_i y_i - \frac{T_x T_y}{N} \tag{48}$$

En utilisant ces quantités on trouve :

$$b = \frac{S_{xy}}{S_{xx}} \tag{49}$$

$$a = \overline{y} - b\overline{x} \tag{50}$$

Pour la droite de régression on obtient la relation des moindres carrés :

$$Y = a + bx = \overline{y} + b\left(x - \overline{x}\right) \tag{51}$$

On calcule également la somme des carrés autour de la régression ($SQ_{a.reg}$) selon :

$$SQ_{a.reg} = \sum_{i=1}^{N} \left(y_i - Y_i\right)^2 \tag{52}$$

La valeur Y_i peut être obtenue en substituant les valeurs x_i dans (51).

On introduit également la somme totale des carrés SQ_{tot} et la somme des carrés due à la régression $SQ_{d.reg}$.

$$SQ_{tot} = \sum_{i=1}^{N} \left(y_i - \overline{y} \right)^2 = S_{yy} \tag{53}$$

$$SQ_{d.reg} = S_1^2 = \sum_{i=1}^{N} \left(\overline{Y}_i - \overline{y} \right)^2 = b^2 S_{xx} = \frac{S_{xy}^2}{S_{xx}} \tag{54}$$

$$SQ_{a.reg} = SQ_{tot} - SQ_{d.reg} \tag{55}$$

SQ calculées selon (53) à (55) et les carrés moyens correspondants DQ :

$$\frac{SQ}{\phi} = S_y^2 = DQ \tag{56}$$

sont habituellement résumés dans les tableaux d'analyse de variance :

Tableau III. 1 : Analyse de variance.

Source	**SQ**	ϕ	**DQ**
Due à la régression	S_{xy}^2 / S_{xx}	1	S_{xy}^2 / S_{xx}
Autour de la régression	$S_{yy} - \frac{S_{xy}^2}{S_{xx}}$	N-2	$\frac{S_{yy} - \frac{S_{xy}^2}{S_{xx}}}{N-2}$
Totale	S_{yy}	N-1	$\frac{S_{yy}}{N-1}$

(57)

La moyenne de la somme des carrés $D_{a.reg}$ est une mesure de la dispersion des valeurs y_i autour de la régression : c'est la variance V(y).

$$V(y) = S_2^2 = \frac{S_{yy} - \frac{S_{xy}^2}{S_{xx}}}{N-2} \tag{58}$$

L'erreur est définie par :

$$\sqrt{V(y)} = SE(y) \tag{59}$$

On a :

$$F_{obs} = \frac{S_1^2}{S_2^2} \tag{60}$$

Le rapport F est un critère qui permet de voir si (a+bx) est une bonne approximation de x (=x_{MLR} ; x_{RNA} ; x_{SVM}).

La valeur $SE(y)$ peut être utilisée pour les calculs des déviations standards $SE(a)$ et $SE(b)$ des paramètres a et b :

$$SE(a) = S_2(a) = SE(y) \times \sqrt{\frac{1}{N} + \frac{\bar{x}^2}{S_{xx}}} \tag{61}$$

$$SE(b) = S_2(b) = \frac{SE(y)}{\sqrt{S_{xx}}} \tag{62}$$

Les quantités $SE(a)$ et $SE(b)$ peuvent être utilisées comme mesure de la précision avec laquelle les paramètres a et b sont obtenus si le nombre de paires x_i/y_i, c'est-à-dire le degré de liberté $\phi = N - 2$, est pris en compte.

Les limites de confiance sup. et inf. a_s/a_i et b_s/b_i sont définies comme suit :

$$\begin{aligned} a_{\sup} &= a + t_{p,\phi}.\, SE(a) \\ a_{\inf} &= a - t_{p,\phi}.\, SE(a) \\ b_{\sup} &= b + t_{p,\phi}.\, SE(b) \\ b_{\inf} &= b - t_{p,\phi}.\, SE(b) \end{aligned} \tag{63}$$

Où $t_{p,\phi}$ (t de student) est fonction du degré de liberté ϕ et de la probabilité P de trouver les valeurs pertinentes entre les limites de confiance. La probabilité désirée pour les limites de confiance doit toujours être spécifiée à l'avance.

L'établissement d'une régression se conclut avec les calculs des paramètres a et b et la détermination de leurs intervalles de confiance.

Les limites de confiance peuvent ensuite être utilisées pour tester certaines hypothèses. Ainsi, on peut se demander si la droite de la régression linéaire sous-tendant l'ensemble des paires $x_i y_i$:

α / passe par l'origine des coordonnées

β / possède une pente différente de l'unité.

Dans les calculs de régression réalisés pour corréler des valeurs théoriques et expérimentales, l'intérêt principal concerne les limites de confiance d'une valeur individuelle.

Le calcul est effectué comme suit : pour une valeur x donnée, l'erreur standard :

$$SE(Y) = SE(y)\sqrt{1+\frac{1}{N}+\frac{\left(x-\bar{x}\right)^2}{S_{xx}}} \tag{64}$$

est utilisée pour calculer les limites de confiance supérieure et inférieure :

$$\begin{aligned} Y_{sup} &= Y + t_{p,\varphi} SE(Y) \\ Y_{\inf} &= Y - t_{p,\varphi} SE(Y) \end{aligned} \tag{65}$$

pour une probabilité donnée. L'expression (64) inclut non seulement les contributions de la déviation standard due à l'imprécision sur la pente b de la droite de régression et de la position $\bar{y}$ correspondant à $\bar{x}$, mais également la variance des points individuels par rapport à la droite de régression. La déviation standard SE(Y) est une fonction de x et croît avec la distance par rapport à $\bar{x}$. En conséquence, l'extrapolation des valeurs pour des points extérieurs à l'intervalle d'origine du paramètre x fournira des erreurs considérables.

APPLICATION NUMÉRIQUE : EXEMPLE

On considère les 7 paires de valeurs suivantes :

x : 1 2 4 5 6 7 10
y : 1 2 3 4 4 6 8 } N=7

On en déduit : $T_x=35 \quad \bar{x} = 5$

$T_y=28 \quad \bar{y} = 4$

$$S_{xx} = 231 - \frac{35^2}{7} = 56 \;;\; b = \frac{43}{56} = 0,7679$$

$$S_{yy} = 146 - \frac{28^2}{7} = 34 \;;\; a = 4 - 5b = 0,1607$$

$$S_{xy} = 183 - \frac{28 \times 28}{7} = 43$$

Y=0,1607+0,7679x

Tableau III. 2 : Analyse de la variance (application).

Source	SQ	ϕ	DQ
Due à la régression	33,0179	1	33,0179
Autour de la régression	0,9821	5	0,1964
Total	34,0000	6	

$$\sqrt{V(y)} = \sqrt{0,1964} = SE(y) = 0,4432$$

$$SE(a) = SE(y) \times \sqrt{\frac{1}{N} + \frac{\overline{x}^2}{S_{xx}}} = 0,442 \times \sqrt{\frac{1}{7} + \frac{25}{56}} = 0,3402$$

$$SE(b) = \frac{SE(y)}{\sqrt{S_{xx}}} = \frac{0,4432}{\sqrt{56}} = 0,0592$$

Ce qui permet le calcul de a_{sup} et a_{inf} ainsi que b_{sup} et b_{inf} :

Sachant que pour P=0,95, $t_{0,95\ ;\ 5}$=2,02, il vient :

$$a_{\text{sup}} = 0,1607 + 2,02 \times 0,3402 = +0,8479$$
$$a_{\text{inf}} = 0,1607 - 2,02 \times 0,3402 = -0,5265$$
$$b_{\text{sup}} = 0,7679 + 2,02 \times 0,0592 = +0,8875$$
$$b_{\text{inf}} = 0,7679 - 2,02 \times 0,0592 = +0,6483$$

Calculons les limites de confiance de y pour les positions x=8 et x=15 dans l'exemple traité :

$$SE(Y) = 0.4432\sqrt{1 + \frac{1}{7} + \frac{(x-5)^2}{56}}$$

x=8 : $SE(Y) = 0.4432\sqrt{1 + 0.1429 + 0.1607} = 0.506$

$$Y = 0,1607 + 0,7679 \times 8 = 6,304$$

$$\left.\begin{aligned} Y_{\text{sup}} &= 6,304 + 2,02 \times 0,506 = 7,326 \\ Y_{\text{inf}} &= 6,304 - 2,02 \times 0,506 = 5,282 \end{aligned}\right\} P = 0,9$$

x=15 : $SE(Y) = 0,4432\sqrt{1 + 0,1429 + 1,7858} = 0,759$

$$Y = 0,1607 + 0,7679 \times 15 = 11,679$$

$$\left.\begin{array}{l} Y_{\sup} = 11,679 + 2,02 \times 0,759 = 13,211 \\ Y_{\inf} = 11,679 - 2,02 \times 0,759 = 10,147 \end{array}\right\} P = 0,9$$

Comme on peut le voir les limites de confiance pour la position d'un point additionnel avec x= 8 et spécialement pour x = 15 sont plutôt larges pour le niveau de confiance P = 0,9 requis. Notons que la déviation standard SE(Y) est fonction de x et croît avec l'augmentation de la distance par rapport à $\bar{x}$.

III. 4. 2. 4. 3. Comparaison de deux moyennes de variables appariées : [159]

Emploi de la loi de Student

a. Conditions de validité :

Cette méthode n'est utilisable que pour des résultats appariés. Ceci est réalisé lorsque les conditions expérimentales sont telles qu'à tout résultat x_{1i} de la première série peut correspondre le résultat x_{2i} de la deuxième série.

Il en sera ainsi par exemple si x_{1i} et x_{2i} sont les résultats des deux mesures effectuées à partir du même échantillon. L'ensemble des résultats est donc formé de couples de valeurs non indépendantes. Il en résulte que les effectifs n_1 et n_2 sont égaux.

Cette méthode n'est utilisable que pour comparer des résultats prélevés dans des populations normales. Cependant, (théorème de Laplace – Liapounoff) elle reste valable si les populations ne sont pas normales à condition que les effectifs n_1 et n_2 soit suffisamment élevés.

- Aucune hypothèse sur les variances vraies n'est nécessaire.

Remarque : Cette méthode consiste à contrôler l'hypothèse de nullité de la différence des moyennes m_1 et m_2 ($m_1 - m_2 = d = 0$). Si on désigne par y_i la différence $x_{1i} - x_{2i}$, on sera donc amené à comparer la moyenne y des y_i à 0. Les estimations des variances σ_1^2 et σ_2^2 nécessaires pour effectuer ces comparaisons pourront être calculées soit à partir des écarts-types soit à partir des étendues. Nous rappellerons la méthode utilisant les estimations S_1^2 et S_2^2 .

Cette méthode devra être préférée à toute autre chaque fois que l'on disposera de résultats appariés. En effet elle est plus rapide. Elle est également plus efficace car elle permet de ne pas tenir compte de la variance entre paires de résultats.

b. Exposé de la méthode :

b.1. Détermination de la valeur expérimentale t :

On désigne par t la valeur numérique de la fonction discriminante. Les différences $y_i=x_{1i}-x_{2i}$ ayant été calculées, la valeur expérimentale t est définie par la formule :

$$t = \frac{\bar{y}}{\frac{S_y}{\sqrt{n}}} \tag{66}$$

Elle se calcule, en pratique, à l'aide de la formule :

$$t = \sum y_i \sqrt{\frac{n-1}{n\sum y_i^2 - \left(\sum y_i\right)^2}} \tag{67}$$

b. 2. Détermination des limites $t_{\alpha 1}$ et $t_{1-\alpha 2}$:

Ces limites pourront être déterminées en utilisant la table de la loi de Student. Celle-ci donne en fonction du nombre de degrés de liberté $\nu = n - 1$ et pour α inférieur à 0,05 la valeur positive $t_{1-\alpha}$ ayant la probabilité α d'être dépassée lorsque l'hypothèse H_0 est vraie. Pour le risque α partagé en α_1 et α_2 cette table donne directement la limite $t_{1-\alpha 2}$. La limite $t_{\alpha 1}$ se déduit de la valeur $t_{1-\alpha 1}$.

$$t_{\alpha 1} = - t_{1-\alpha 1} \tag{68}$$

b. 3. Décision :

La valeur expérimentale t est ensuite comparée aux limites $t_{\alpha 1}$ et $t_{1-\alpha 2}$.

- Si l'hypothèse contrôlée est d = 0

elle sera refusée si $t < t_{\alpha 1}$ ou $t > t_{1-\alpha 2}$

elle ne sera pas refusée si $t_{\alpha 1} < t < t_{1-\alpha 2}$

- Si l'hypothèse contrôlée est $d \geq 0$ ($m_1 \geq m_2$) ($\alpha_2 = 0$)

elle sera refusée si $t > t_{1-\alpha 2}$

elle ne sera pas refusée si $t < t_{1-\alpha 2}$

Lorsque l'hypothèse contrôlée est refusée, la probabilité pour que cette conclusion soit fausse est égale à α.

III. 4. 2. 4. 4. Comparaison de deux droites de régression : [160]

Soit : $Y = a + bx = \bar{y} + b\left(x - \bar{x}\right)$

L'équation d'une droite de régression.

Les deux droites comparées seront distinguées par les indices I et II.

Il faut calculer les coefficients ȳ et b, ainsi que les variances résiduelles S_2^2.

Nous nous plaçons dans le cas où l'un des nombres de couples de résultats NI ou NII est inférieur à 30.

III. 4. 2. 4. 5. Comparaison des ordonnées des deux droites au point moyen :

Choisir une valeur de x_0 appartenant au domaine de toutes les droites et aussi près que possible du point moyen pour chacun. Calculer la valeur de Y correspondant sur chaque droite à l'abscisse x_0 et comparer les valeurs de Y de la façon indiquée ci-après. On calcule une estimation S_2^2 de la variance résiduelle commune aux deux droites en faisant une moyenne pondérée de $S_{2\,\mathrm{I}}^2$ et $S_{2\,\mathrm{II}}^2$ le nombre de degrés de liberté :

$$S_2^2 = \frac{\left(N_I - 2\right) S_{2(I)}^2 + \left(N_{II} - 2\right) S_{2(II)}^2}{N_I + N_{II} - 4} \tag{69}$$

Cette estimation est utilisée pour calculer $S_{Y_I}^2$ et $S_{Y_{II}}^2$ par la formule :

$$S_{Y_I}^2 = S_2^2 \left[\frac{1}{N_I} + \frac{\left(x_0 - \bar{x}\right)}{\sum \left(x_{i_I} - \bar{x}_I\right)^2} \right] \tag{70}$$

La formule (71) permet d'en déduire une valeur t.

$$t = \frac{\left|Y_I - Y_{II}\right|}{\sqrt{S_{Y_I}^2 + S_{Y_{II}}^2}} \tag{71}$$

Cette variable suit la loi de Student à ($N_I + N_{II} - 4$) degrés de liberté dans le cas où les deux droites de régression vraies sont confondues.

La valeur expérimentale t est donc comparée à la limite $t_{1-\alpha/2}$ donnée par la table de Student.

Si la valeur de t est supérieure à la limite donnée par la table, on peut admettre, au niveau de confiance choisi, que les deux droites se déplacent parallèlement l'une par rapport à l'autre.

III. 4. 2. 4. 6. Comparaison des pentes des deux droites :

Utiliser l'estimation commune (69) de la variance résiduelle pour calculer :

$$S^2_{b_{(I)}} = \frac{S^2_{2(I)}}{\sum\left(x_{i_I} - \bar{x}_I\right)^2} \tag{72}$$

En déduire t par la formule :

$$t = \frac{\left|b_I - b_{II}\right|}{\sqrt{S^2_{b_I} + S^2_{b_{II}}}} \tag{73}$$

Cette variable suit la loi de Student à ($N_I + N_{II} - 4$) degrés de liberté dans le cas où les deux droites de régression vraies sont confondues. Comparer la valeur de t à la limite $t_{1-\alpha/2}$ donnée par la table de Student. Si le test est significatif, on peut conclure qu'il y a rotation d'une droite par rapport à l'autre.

III. 4. 2. 4. 7. Comparaison des variances résiduelles :

Comparer les valeurs S^2_2 par le test de Snedecor en formant le rapport :

$$F = \frac{S^2_{2\,I}}{S^2_{2\,II}}$$

Comparer ce rapport expérimental aux limites données par la table de Snedecor pour $Y_I = (N_I - 2)$ et $Y_{II} = (N_{II} - 2)$, soit, au niveau de confiance $(1 - \alpha)$:

$$F_{1-\alpha/2}(N_I, N_{II}) \text{ et } F_{\alpha/2}(N_I, N_{II}) = \frac{1}{F_{1-\alpha/2}(N_I, N_{II})}$$

L'hypothèse contrôlée : $\sigma_1^2 = \sigma_2^2$

- sera refusée si $F > F_{1-\alpha_2}$

ou si $F < F_{\alpha_1}$

- ne sera pas refusée si $F_{\alpha_1} < F < F_{1-\alpha_2}$

L'utilisation des indices de rétention permet d'uniformiser la présentation des données de rétention en chromatographie gazeuse.

III. 4. 2. 4. 7. Programmes de calculs, en Matlab :

```
t95N= 1.708 ; % Entrer le t(p,N-2)
y=[]; % Entrer les yi : variable dépendante (valeur fournie par le modèle).
x=[]; % Entrer les xi : variable indépendante (valeur mesurée – expérimentale).
X=[];% Entrer les xi pour les composés de validation externe
Tx=sum(x)
Ty=sum(y)
Xmean = mean(x);
Ymean = mean(y);
dX = length(x);
Vx  = var(x);
Vy  = var(y);
Sxx = Vx*(dX-1)
Syy = Vy*(dX-1)
DET = (x-Xmean).* (y-Ymean);
Sxy = dX * (mean(DET));
b = Sxy/Sxx
a = Ymean-(b*Xmean)
Y= a+b*y        %Droite des moindres carrés.
SQareg = sum((y-Y).*(y-Y))  %la somme des carrés autour de la régression (SQa.reg)
SQdreg = sum((Y-Ymean).*(Y-Ymean)) %la somme des carrés due à la régression SQd.reg.
Due = (Sxy*Sxy)/Sxx
Autour=Syy-Due
Vy=Autour/(dX-2)
SEy=sqrt(Vy)
```

```
SEa=SEy*sqrt((1/dX)+(Xmean*Xmean/Sxx))
SEb=SEy/sqrt(Sxx)
Au = a+(t95N*SEa)
Al = a-(t95N*SEa)
Bu = b+(t95N*SEb)
Bl = b-(t95N*SEb)
% Validation
SEY=SEy*sqrt(1+(1/dX)+(((X-Xmean).*(X-Xmean))/Sxx))
Yv=a+b*X
Yvsup=Yv+(t95N*SEY)
Yvinf=Yv-(t95N*SEY)
tat3=[X Yv SEY Yvsup Yvinf]
stat=[Tx Ty Xmean Ymean Sxx Syy Sxy b a SQareg SQdreg ];
stat1=[Vy SEy SEa SEb Au Al Bu Bl];
tat=[Due Autour Syy];
tat1=[dX-2 dX-1];
tat2 = [Due Vy];
xlswrite ('STAT20190.xls',stat','feuil1','B1' );
xlswrite ('STAT20190.xls',tat','feuil1','B13' );
xlswrite ('STAT20190.xls',tat1','feuil1','C14' );
xlswrite ('STAT20190.xls',tat2','feuil1','D13' );
xlswrite ('STAT20190.xls',stat1','feuil1','B16' );
xlswrite ('STAT20190.xls',tat3,'feuil1','A27' );
plot (x,y,'*', X,Yv,'o' , X,Yvinf, '+',X,Yvsup,'+')
```

III. 5. Paramètres d'évaluation de la qualité de l'ajustement :

L'ajustement des modèles QSRR peut être déterminé par le coefficient de détermination multiple R^2 et la racine de l'erreur quadratique moyenne RMSE (Root Mean-Squared Error).

Ces paramètres sont calculés sur l'ensemble de calibrage et ils sont utilisés pour décider si le modèle possède la qualité prédictive reflétée dans le R^2. L'utilisation de la RMSE montre l'erreur entre la moyenne des valeurs expérimentales et prédites.

- Le coefficient de détermination multiple :

$$R^2 = 1 - \frac{SCE}{SCT} = 1 - \frac{\sum_1^n (y_i - \hat{y}_i)^2}{\sum_1^n (y_i - \bar{y})^2} \quad (74)$$

où $\hat{y}_i$ est la valeur estimée du paramètre physique, et $\bar{y}$ la moyenne des valeurs observées.

- La racine de l'erreur quadratique moyenne de calibrage (désignée également par EQMC) :

$$EQMC = \sqrt{\frac{\sum_{i=1}^{n} \left(y_i - \hat{y}_i \right)^2}{n}} \quad (75)$$

III. 5. 1. Robustesse du modèle :

La stabilité du modèle a été explorée en utilisant la validation croisée, cette dernière est considérée comme une validation interne qui consiste à mesurer sa capacité à corréler la propriété avec les descripteurs quand on modifie légèrement les données (suppression d'une ou plusieurs données). Il existe plusieurs méthodes de validation croisée : LOO (*Leave One Out*) **[161]** et LMO (*Leave Many Out*) **[162]**.

Dans le cas du Leave One Out (LOO), une seule observation du jeu d'entraînement est retirée et les coefficients de la régression sont optimisés sur les n-1 autres données.

La propriété prédite $\hat{y}_{(i)}$ est recalculée à partir de cette nouvelle équation pour le composé isolé. Cette manipulation est effectuée pour les n hydrocarbures du jeu d'entraînement, puis le coefficient de prédiction noté Q^2 est calculé à l'aide de l'équation suivante :

$$Q^2_{LOO} = \frac{SCT - PRESS}{SCT} \quad (76)$$

La somme des carrés des erreurs de prédiction, désignée par l'acronyme PRESS (pour : Predictive Residual Sum of Squares) est calculée par :

$$PRESS = \sum_1^n \left(y_i - \hat{y}_{(i)} \right)^2 \quad (77)$$

Et le EQMP par :

$$\sigma_N = EQMP = \sqrt{\frac{\sum_{i=1}^{n}\left(y_i - \hat{y}_{(i)}\right)^2}{n}} = \sqrt{\frac{PRESS}{n}} \qquad (78)$$

Contrairement à R^2, qui augmente avec le nombre de paramètres de la régression, le facteur Q^2_{LOO} affiche une courbe avec maximum (ou avec palier), obtenu pour un certain nombre de variables explicatives, puis décroît par la suite de façon monotone. Ce fait confère une grande importance au coefficient Q^2_{LOO}. Une valeur de $Q^2_{LOO} > 0{,}5$ est, généralement, considérée comme satisfaisante, et une valeur supérieure à 0,9 est excellente **[163]**.

Dans le cas du Leave Many Out (LMO), un groupe de molécules du jeu d'entraînement est retiré au lieu d'une seule observation. Une faible valeur de Q^2 implique que le modèle n'est pas robuste et ne sera pas prédictif, mais la réciproque n'est pas nécessairement vraie **[164]**. En effet, le modèle est considéré comme robuste quand les différents coefficients de prédiction Q^2 ont des valeurs très proches et quand la différence entre les Q^2 et le R^2 est faible.

III. 5. 2. Test de randomisation :

Ce test permet de mettre en évidence des corrélations dues au hasard. Il consiste à générer un vecteur « propriété considérée » par permutation aléatoire des composantes du vecteur réel (Figure III. 3). On calcule alors sur le vecteur obtenu (considéré comme vecteur expérimental réel) un modèle QSRR, selon la méthode habituelle. Ce procédé est répété plusieurs fois (100 dans notre cas).

Deux méthodes semblent exister : celle qui considère la permutation des descripteurs également **[165, 166]** et celle qui ne le fait pas **[167]**. Dans ce travail, pour une raison pratique (comme la difficulté à automatiser la sélection des descripteurs) la sélection des descripteurs n'a pas été prise en compte lors de la randomisation.

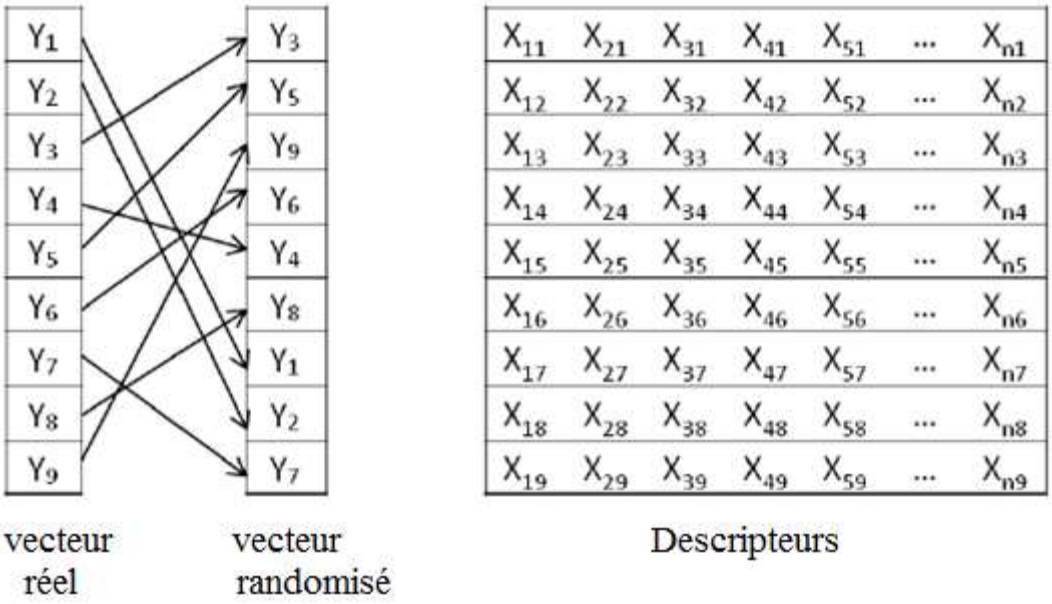


Figure III. 3 : Illustration de la méthode du test de randomisation.

III. 5. 3. Validation externe :

La meilleure façon d'estimer la véritable puissance prédictive d'un modèle QSRR est de comparer les valeurs prédites et observées d'un ou de plusieurs composés « ensemble de validation » qui ne sont pas utilisés dans le développement du modèle **[168, 169]**.

La mesure de la prédictivité la plus utilisée est le $R^2_{CV,ext}$ ou Q^2_{ext} défini par la relation suivante :

$$R^2_{\mathrm{CV,ext}} = 1 - \frac{\sum_{i=1}^{n_{ext}} \left(y_i - \hat{y}_i \right)^2}{\sum_{i=1}^{n_{ext}} \left(y_i - \overline{y}_{tr} \right)^2} \tag{79}$$

De même, l'autre évaluateur (RMSE) de la prédictibilité peut être calculé pour le jeu de validation selon la relation :

$$EQMP_{ext} = \sqrt{\frac{\sum_{i=1}^{n_{ext}} \left(y_i - \hat{y}_{(i)} \right)^2}{n_{ext}}} \tag{80}$$

Une validation externe supplémentaire selon **[164]** est appliquée uniquement à l'ensemble de validation. Selon les critères recommandés de Tropsha *et al*., un modèle QSRR prédictif, doit remplir les conditions suivantes :

1) $Q^2_{EXT} > 0.5$ (81-a)

2) $R^2 > 0.6$ (81-b)

3) $(R^2 - R^2_0)/R^2 < 0.1$ et $0.85 < k < 1.15$ (81-c)

$(R^2 - R'^2_0)/R^2 < 0.1$ et $0.85 < k' < 1.15$ (81-d)

où

$$R = \frac{\sum (y_i - \bar{y})(\tilde{y}_i - \bar{\tilde{y}})}{\sqrt{\sum (y_i - \bar{y})^2 \sum (\tilde{y}_i - \bar{\tilde{y}})^2}} \quad (82\text{-a})$$

$$R^2_0 = 1 - \frac{\sum (y_i - y_i^{r_0})^2}{\sum (y_i - \bar{y})^2} \quad (82\text{-b}) \qquad R'^2_0 = 1 - \frac{\sum (\tilde{y}_i - \tilde{y}_i^{r_0})^2}{\sum (\tilde{y}_i - \bar{\tilde{y}})^2} \quad (82\text{-c})$$

$$k = \frac{\sum (y_i \tilde{y}_i)}{\sum (\tilde{y}_i)^2} \quad (82\text{-d}) \qquad k' = \frac{\sum (y_i \tilde{y}_i)}{\sum (y_i)^2} \quad (82\text{-e})$$

R est le coefficient de corrélation entre les valeurs calculées et expérimentales dans l'ensemble de test; R^2_0 (valeurs calculées par rapport à celles observées) et R'^2_0 (valeurs observées par rapport à celles calculées) sont les coefficients de détermination; k et k 'sont les pentes des droites de régressions passant par l'origine pour les valeurs calculées par rapport aux valeurs observées et observées par rapport à celles calculées, respectivement; $y_i^{r_0}$ et $\tilde{y}_i^{r_0}$ sont définis respectivement par : $y_i^{r_0} = k\,\tilde{y}$ et, $\tilde{y}_i^{r_0} = k'\,y$; les sommations portent sur tous les échantillons de l'ensemble de test.

La validation est en évolution permanente avec l'utilisation de nouveaux coefficients. De manière générale, les coefficients R^2 et Q^2 doivent avoir des valeurs proches de 1 (de préférence supérieures à 0,6) et leur différence doit être faible pour considérer le modèle comme robuste. Cependant, l'évaluation des coefficients doit se faire au regard de la taille de la base de données (notamment pour R^2) et de l'ordre de grandeur de l'incertitude expérimentale (RMSE). Mais d'autres paramètres sont pris en considération pour le choix du modèle comme la possibilité d'interprétation des descripteurs.

III. 6. Domaine d'application :

Le domaine d'application (DA) **[170, 171]** est une région théorique dans l'espace définie par les descripteurs du modèle et la réponse modélisée, pour lequel un modèle QSRR donné devrait faire des prédictions fiables. Dans ce travail, Le domaine d'application structurel a été vérifié par l'approche des leviers (h_{ii}) **[172]**.

Un avertissement sur l'effet de levier important d'un échantillon est, en général, donné pour un h >= h*= 3 (p+1)/n où n est le nombre total d'échantillons dans l'ensemble de calibrage et p le nombre de descripteurs impliqués dans la corrélation.

La présence de valeurs aberrantes en réponse (valeurs aberrantes en **Y**) et les composés structurellement influents (valeurs aberrantes en **X**) a été vérifiée par le diagramme de Williams, qui représente les résidus standardisés en fonction des valeurs des leviers.

III. 7. Les Logiciels utilisés dans nos études QSXR :

Il existe plusieurs logiciels libres ou commerciaux disponibles dans les études QSXR. Ceux-ci comprennent des logiciels spécialisés pour dessiner les structures chimiques, générant des structures 3D, le calcul des descripteurs moléculaires et le développement de modèles QSXR. Les logiciels utilisés dans nos travaux sont :

1/ ChemDraw V.7.0 : Utilisé pour dessiner et structurer les composés chimiques **[173]**.

2/ HyperChem Pro V6.03 : Il offre plusieurs méthodes d'optimisation (PM3, MM+, AM1, etc. …) par lesquelles les molécules sont optimisées, des descripteurs moléculaires peuvent être calculés tels que : le volume moléculaire, les énergies HUMO et LUMO…etc **[174]**.

3/ DRAGON V.6 : Il assure le calcul des descripteurs moléculaires **[175]**.

Les modèles QSXR ont été générés en utilisant les logiciels suivants :

- Pour la régression linéaire multiple, nous avons utilisé Minitab **[176]**.

- Pour les domaines d'applicabilité, nous avons utilisé Matlab **[177]**.

- Pour les réseaux de neurones artificiels et les Machines à vecteurs supports, nous avons utilisé le logiciel Molegro **[178]**.

Deuxième partie : Application

I. Modélisation de l'indice de rétention

I. 1. La régression linéaire multiple

I. 2. Méthode des réseaux de neurones artificiels

I. 3. Machines à vecteurs supports SVM

I. 4. Corrélation valeurs calculées – valeurs mesurées et intervalles de confiance

I. Modélisation de l'indice de rétention :

I. 1. La régression linéaire multiple :

I. 1. 1. Collecte et division des données :

Les données prélevées dans la littérature **[179]** ont été, au préalable, séparées par l'algorithme de Kennard et Stone (CADEX) en un ensemble de calibrage de 27 composés, et un ensemble de validation comprenant les 9 composés restants.

I. 1. 2. Calcul des descripteurs moléculaires :

Nous avons utilisé le logiciel de modélisation moléculaire HyperChem 6,03 pour représenter les molécules puis, à l'aide de la méthode semi- empirique PM3, obtenir les géométries finales. Tous les calculs ont été menés dans le cadre du formalisme RHF (pour Restricted Hartree-Fock ou formalisme de Hartree Fock avec contrainte de spin) sans interaction de configuration.

Les structures moléculaires ont été optimisées à l'aide de l'algorithme Polak- Ribiere avec pour critère une racine du carré moyen du gradient égale à 0,001 kcal/mol. Les géométries ainsi optimisées ont été transférées dans le logiciel informatique DRAGON pour le calcul de plus de 1600 descripteurs. En utilisant les options correspondantes du logiciel DRAGON version 6, nous avons d'abord éliminé les descripteurs à valeurs constantes (écarts types inférieurs à 0,001) qui n'apportent aucune information, ensuite ceux qui sont hautement corrélés ($R \geq 0,95$) et qui véhiculent une information redondante.

I. 1. 3. Calcul du modèle :

Nous avons appliqué la règle qui stipule qu'au minimum, on peut avoir un descripteur ou variable explicative pour cinq (5) observations du jeu de calibrage. Disposant de 27 composés, la limite est de 5 variables. L'évaluation par algorithme génétique a été faite pour les dimensions 1, 2, 3, 4 et 5.

Parmi les modèles obtenus, un modèle à quatre descripteurs a été sélectionné car il représente le meilleur compromis entre ajustement (R^2), prédiction (Q^2_{LOO}, Q^2_{EXT}) et stabilité (Q^2_{BOOT}). Les descripteurs retenus sont reportés dans le tableau I. 1. Ce tableau comporte aussi une définition succincte des descripteurs ainsi que les classes auxquelles ils appartiennent.

Tableau I. 1 : Descripteurs moléculaires.

N°	Descripteur	Classe	Définition
1	piPC02	Nombre de marches et de chemins	Nombre de Chemins moléculaires multiples d'ordre 2.
2	ChiA_B(p)	Descripteurs basés sur des matrices 2D	Indice de type Randic moyen à partir de la matrice de Burden pondérée par la polarisabilité
3	SM2_B(s)	Descripteurs basés sur des matrices 2D	Moment spectral d'ordre 2 à partir de la matrice de Burden pondérée par l'état-I
4	Mor15u	Descripteur MoRSE-3D	Signal 15 / non pondéré

I. 1. 4. Equation et analyse de régression :

Le modèle obtenu a pour équation :

$$\log Ir = 1{,}16 + 0{,}291\ piPC02 + 1{,}94\ ChiA\_B(p) + 0{,}0992\ SM2\_B(s) - 0{,}0342\ Mor15u \quad (83)$$

$$n = 27\ ;\ R^2 = 97{,}81\ \%\ ;\ Q^2_{LOO} = 96{,}91\ \%\ ;\ Q^2_{L(5)O} = 96{,}67\ \%\ ;\ Q^2_{Boot} = 95{,}92\ \%$$

$$F = 245{,}27\ (p = 0{,}000)\ ;\ s = 0{,}013\ \log unit\ ;\ EQMC = 0{,}011\ ;\ EQMP = 0{,}013$$

Le logarithme de l'indice de rétention pour chacun des 27 composés utilisés pour l'ensemble de calibrage (élaboration du modèle) est bien corrélé avec les quatre descripteurs d'où la grande valeur du coefficient de détermination R². Le modèle a de très bonnes aptitudes prédictives confirmées par la valeur de Q². En autre la statistique de Fisher montre qu'il est très significatif. Les écarts quadratiques (EQMP/C) sont faibles et proches.

D'après les valeurs du test t ($|t|$), on peut classer les descripteurs sélectionnés selon leur pourcentage de contribution qui dans ce cas se présente dans l'ordre décroissant suivant :

$$piPC02 > SM2\_B(s) > ChiA\_B(p) > Mor15u$$

Les valeurs des VIF (< 5) suggèrent que ces descripteurs sont faiblement corrélés les uns avec les autres.

Ainsi, le modèle peut être considéré comme une équation de régression optimale.

Tableau I. 2 : Caractéristiques des descripteurs du modèle.

Descripteur	Coef	SE Coef	t	Probabilité-t	VIF
Constant	1,1622	0,1293	8,9900	0,0000	
piPC02	0,2906	0,0163	17,8100	0,0000	1,9500
ChiA_B(p)	1,9410	0,3718	5,2200	0,0000	1,6400
SM2_B(s)	0,0992	0,0121	8,1700	0,0000	1,3340
Mor15u	-0,0342	0,0049	-7,0000	0,0000	1,2160

La valeur de t pour un descripteur est liée à sa signification statistique. Les valeurs absolues élevées de t rapportées indiquent que les coefficients de régression sont significativement plus grands que l'écart type. La probabilité P de t pour un descripteur donne sa signification statistique lorsqu'il est impliqué dans un modèle QSRR global ; elle renseigne sur les interactions entre descripteurs. Les descripteurs auxquels correspondent des probabilités de t inférieures à 0,05 sont considérés comme statistiquement significatifs pour un modèle donné, c'est-à-dire que leur influence sur la variable dépendante n'est pas due au hasard **[180]**. Les valeurs des probabilités de t pour les quatre descripteurs sont nulles, ce qui indique qu'ils sont hautement significatifs.

Les valeurs des facteurs d'inflation de la variance et la matrice de corrélation reproduite dans le tableau I. 3 suivant suggèrent que ces descripteurs sont faiblement corrélés entre eux.

Tableau I. 3 : Matrice de corrélation des descripteurs du modèle.

	log Ir	piPC02	ChiA_B(p)	SM2_B(s)
piPC02	0,860			
	(0)			
ChiA_B(p)	-0,083	-0,467		
	(-0,681)	(-0,014)		
SM2_B(s)	0,646	0,360	0,136	
	(0)	(-0,065)	(-0,497)	
Mor15u	-0,545	-0,291	-0,130	-0,163
	(-0,003)	(-0,141)	(-0,518)	(-0,417)

I. 1. 5. Analyse des résidus et diagnostics d'influence :

Les valeurs de h_i, ainsi que les valeurs des résidus, ordinaires et standardisés, sont présentées dans le tableau I. 4, dont les colonnes 3 et 4 reproduisent les valeurs expérimentales et calculées des indices de rétention (log Ir) des composés considérés,

Tous les résidus ordinaires e_i (colonne 6), sont inférieurs, en valeur absolue, à 3 fois l'erreur standard ($|e_i| < 3S$), soit 3S= 0,039.

Tableau I. 4 : Valeurs des log Ir expérimentales (log Ir_{exp}) et calculées (log Ir_{calc}), des leviers (h_i) ainsi que des résidus ordinaires (e_i) et standardisés ($e_{i\ std}$).

ID	Composés	log Ir Exp	log Ir Calc	Hat (h_i)	e_i	$e_{i\ std}$
1	alpha-Thujene	2,9782	2,9944	0,2370	0,0144	1,7166
2	beta-Pinene	2,9965	2,9896	0,2320	-0,0104	-1,2321
3	beta-Myrcene	3,0026	2,9958	0,2030	-0,0042	-0,4673
4	o-Cymene	3,0224	3,0409	0,1280	0,0209	2,0480
5	1,8-cineole	3,0253	3,0355	0,1230	0,0055	0,5323
6	cis-beta-Terpineol	3,0370	3,0539	0,1340	0,0139	1,3702
7	Terpinolene	3,0426	3,0247	0,1890	-0,0153	-1,6718
8	Linalool	3,0481	3,0578	0,2130	0,0078	0,8836
9	Isopulegol	3,0550	3,0610	0,2120	0,0110	1,2472
10	Camphor	3,0652	3,0746	0,2500	0,0046	0,5629
11	Isoborneol	3,0785	3,0490	0,1130	-0,0310	-2,9546
12	Thymol methyl ether	3,0983	3,1038	0,3860	0,0038	0,6269
13	Verbenone	3,1031	3,1069	0,3040	0,0069	0,9443
14	Dihydrocarvone	3,1116	3,0935	0,2120	-0,0165	-1,8779
15	Thymol	3,1399	3,1350	0,0580	-0,0050	-0,4323
16	Eugenol	3,1464	3,1500	0,5070	0,0000	-0,0008
17	beta-Caryophyllene	3,1667	3,1623	0,0750	-0,0077	-0,6879
18	alpha-Bergamotene	3,1644	3,1701	0,1010	0,0101	0,9418
19	Germacrene D	3,1787	3,1785	0,1490	-0,0015	-0,1563
20	gamma-Elemene	3,1804	3,1769	0,0710	-0,0031	-0,2801
21	beta-Bisabolene	3,1867	3,1735	0,1170	-0,0165	-1,5873

Tableau I. 4 : suite et fin

22	Delta-cadinene	3,1906	3,1908	0,1570	0,0008	0,0863
23	Caryophyllene oxide	3,2009	3,2100	0,1380	0,0100	0,9970
24	Spathulenol	3,2063	3,2103	0,2020	0,0003	0,0323
25	Aromadendrene oxide	3,2101	3,2001	0,1390	-0,0099	-0,9860
26	Muurolol	3,2177	3,2280	0,1800	0,0080	0,8614
27	Bisabolol	3,2201	3,2233	0,1680	0,0033	0,3452

I. 1. 6. Validation externe :

Pour vérifier les capacités prédictives du modèle on a eu recours à sa validation sur un ensemble prévu à cet effet et choisi dès le départ. Cet ensemble de validation, qui n'a pas servi à l'élaboration du modèle, est constitué des composés numérotés de 28 à 36 (tableau I. 5).

$$n_{EXT} = 9 ;\ Q^2_{EXT} = 95{,}46\ \% ;\ EQMP_{EXT} = 0{,}016$$

La valeur de $Q^2_{EXT} = 95{,}46\ \%$ nous informe sur la validité du modèle et sa capacité à prédire des valeurs qui n'ont pas servi à le générer. L'écart quadratique moyen de prédiction externe $EQMP_{EXT} = 0{,}016$ est faible, garantie d'une bonne aptitude à la prédiction.

Tableau I. 5 : Quelques caractéristiques des éléments de l'ensemble de validation externe pour le logarithme de l'indice de rétention.

ID	Composés	log Ir Exp	log Ir Pred	Hat (h_i)	$e_{i\ std}$
28	alpha-Pinene	2,9809	2,9980	0,2360	1,6432
29	Camphene	2,9868	2,9858	0,2130	-0,3784
30	alpha-Phellandrene	3,0073	3,0066	0,1940	-0,3043
31	alpha-Terpinene	3,0137	3,0245	0,2030	1,2934
32	gamma-Terpinene	3,0370	3,0198	0,1890	-1,7839
33	Borneol	3,0745	3,0456	0,1210	-2,0738
34	4-Terpineol	3,0766	3,0772	0,1160	-0,2341
35	alpha-Terpineol	3,1069	3,0850	0,1590	-2,1743
36	Cadinol	3,2164	3,2338	0,2160	1,2405

I. 1. 7. Diagramme de Williams :

Avant qu'un modèle QSRR ne soit exploité pour le criblage de composés, son domaine d'application doit être défini, pour que les prédictions des composés contenus dans ce domaine puissent être considérés comme fiables.

Le diagramme de Williams représenté dans la figure I. 1 permet d'afficher les valeurs des résidus de prédiction standardisés en fonction de leviers (h_i), pour les deux ensembles (calibrage et validation).

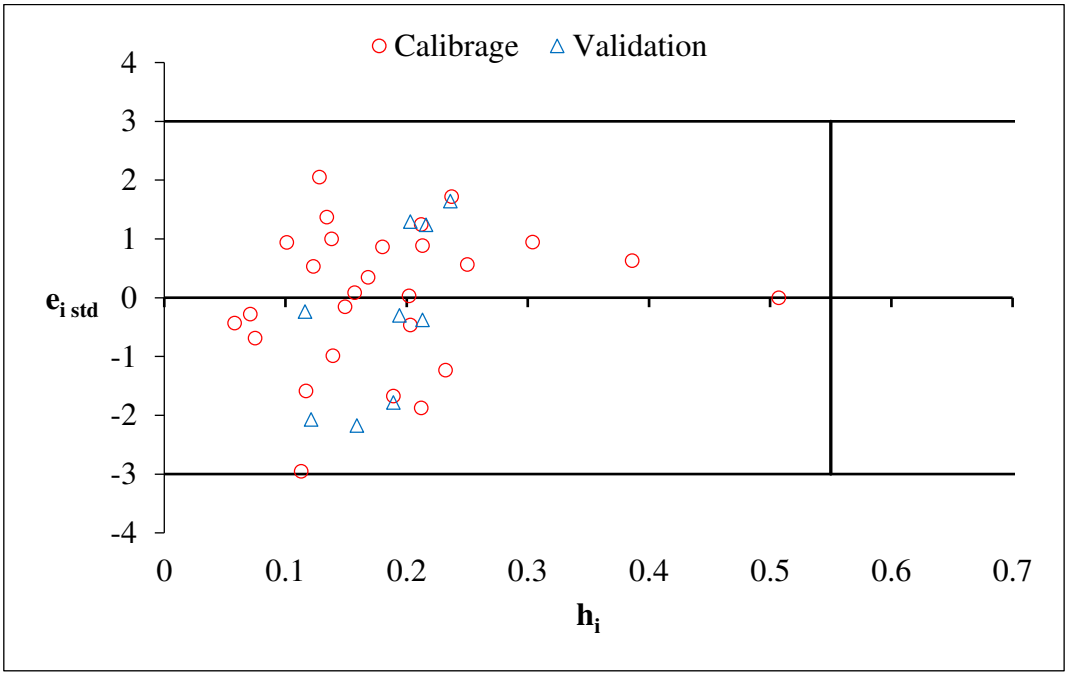


Figure I. 1 : Diagramme de Williams pour les éléments des ensembles de calibrage (27) et de validation (9).

Tous les résidus standardisés de prédiction sont compris entre les limites ± 3, et les valeurs des leviers h_i sont toutes inférieures à la valeur critique

$$h^* = \frac{3\times(p+1)}{n} = \frac{3\times(4+1)}{27} = 0,555.$$

Ainsi, on n'observe ni point influent ($h_i > 0,555$), ni point aberrant ($|e_{istd}| > 3$).

I. 1. 8. Qualité de l'ajustement :

La qualité de l'ajustement peut être vérifiée en procédant par la présentation des valeurs calculées de log Ir, pour l'ensemble de calibrage et des valeurs prédites pour

l'ensemble de validation en fonction de celles expérimentales. La figure I. 2 représente log $Ir_{calc\text{-}pred}$ en fonction de log Ir_{exp} en comparaison à la première bissectrice.

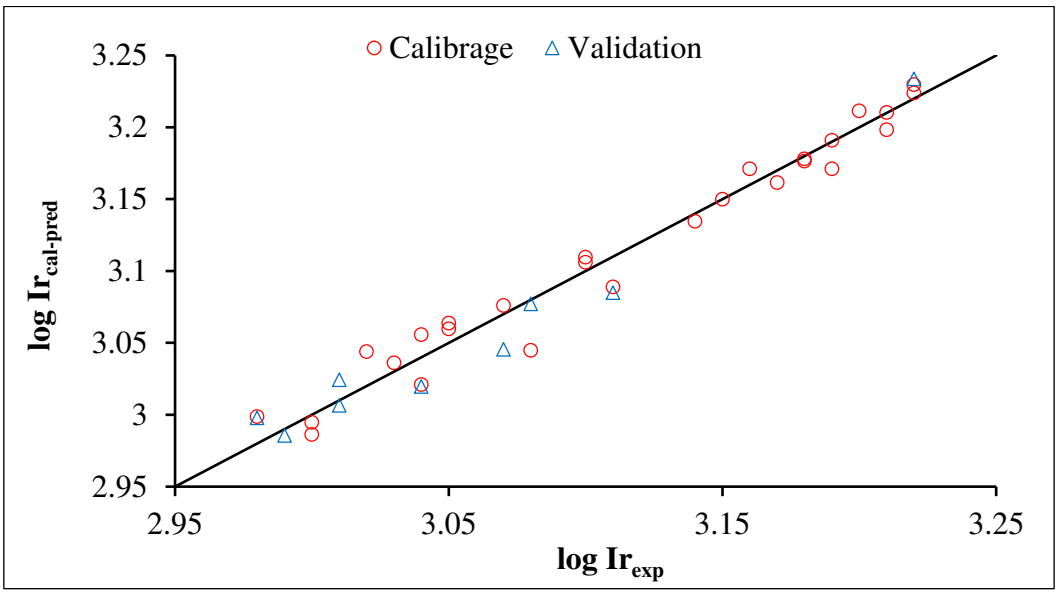


Figure I. 2 : Vérification de l'ajustement des deux ensembles.

La faible dispersion des points autour de la première bissectrice montre que les valeurs prédites (pour l'ensemble de validation) et calculées (pour l'ensemble de calibrage) sont en adéquation avec les valeurs expérimentales.

I. 1. 9. Tests de randomisations :

La validité du modèle a été éprouvée par le test de randomisation de log Ir (Figure I. 3).

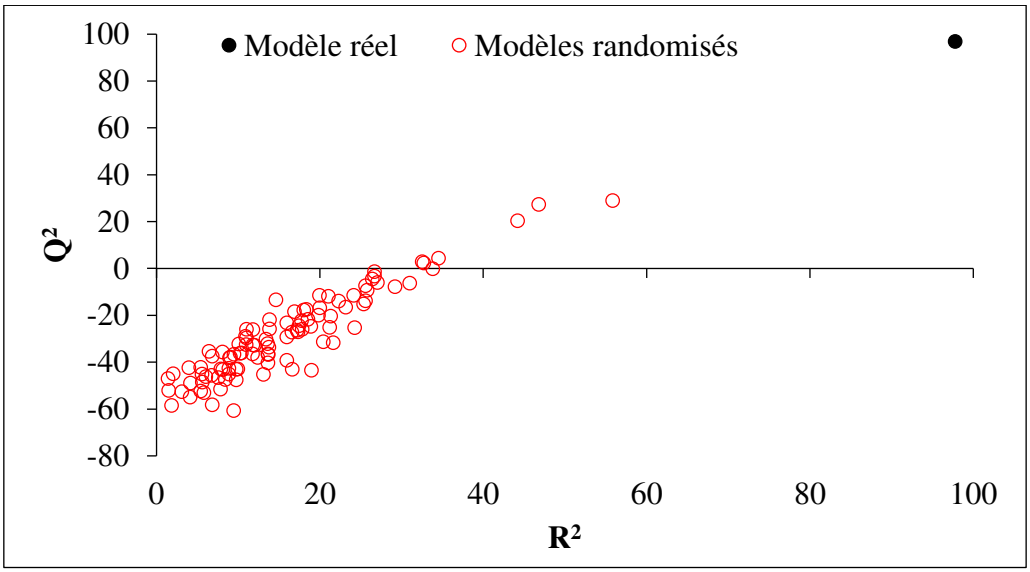


Figure I. 3 : Test de randomisation.

Les 100 modèles pour lesquels nous avons randomisé les valeurs des indices de rétention ont des valeurs de Q^2 ou faibles ou négatives, et des valeurs du coefficient de corrélation multiple (R^2) petites. Seul le cercle noirci a des valeurs élevées et proches pour ces deux statistiques, il représente notre modèle qui, par conséquent, n'est pas dû au hasard.

I. 2. Méthode des réseaux de neurones artificiels :

Le choix du nombre de neurones de la couche cachée est fixé à 3 et le nombre d'itérations à 200. Les graphes suivants explicitent ce choix.

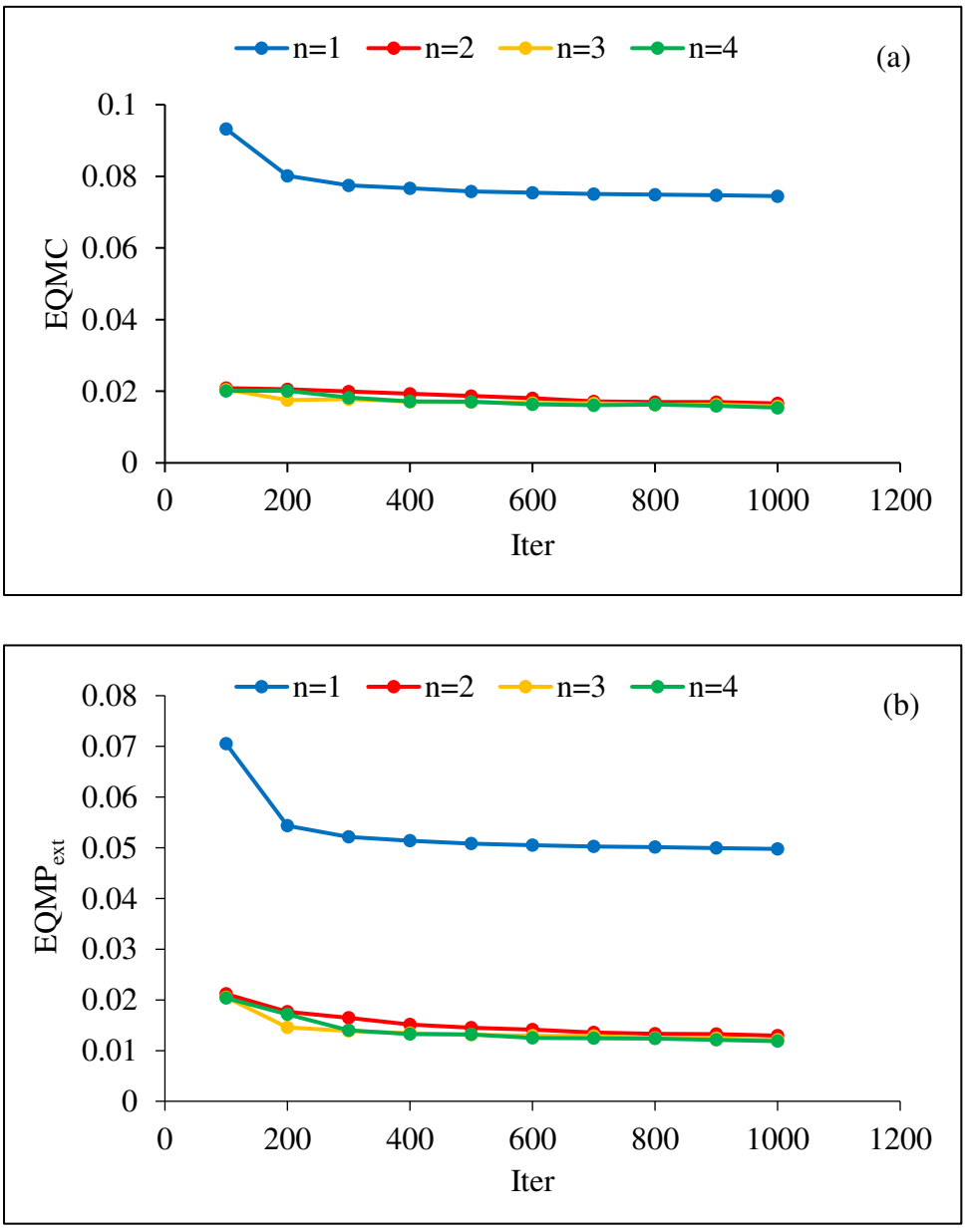


Figure I. 4 : Choix du nombre de neurones de la couche cachée (a) EQMC et (b) $EQMP_{ext}$.

La structure optimale adoptée est reproduite dans le tableau I. 6 :

Tableau I. 6 : Structure optimale adoptée pour le réseau de neurones.

Entrées	04 (les descripteurs)
Sortie	01 (log Ir)
Couche cachée	Une couche cachée
Nombre de neurones dans la couche cachée	3
Algorithme d'apprentissage	Rétro propagation du gradient de l'erreur
Fonction d'apprentissage	Tangente hyperbolique (couche cachée) Linéaire (couche de sortie)

R^2= 96,52%; Q^2_{LOO} = 92,04%; Q^2_{ext} = 94,08%; EQMC = 0,0147; EQMP= 0,0218

$EQMP_{ext}$= 0,0176 ; n= 27 ; n_{ext}= 9

Les valeurs de R^2 montrent la qualité de l'ajustement, alors que la petite différence entre R^2 et Q^2_{LOO} renseigne sur la robustesse du modèle, les valeurs assez proches d'EQMC et $EQMP_{ext}$ signifient que la capacité de prédiction interne du modèle n'est pas trop dissemblable de son pouvoir d'ajustement. La validation statistique externe (Q^2_{ext}) atteste de la bonne capacité prédictive des composés n'ayant pas participé au calcul du modèle.

La qualité de l'ajustement à été vérifiée par la représentation des valeurs calculées du logarithme de l'indice de rétention pour l'ensemble de calibrage, et celles prédites pour l'ensemble de validation en fonction de celles expérimentales. La figure I. 6 représente log $Ir_{cal\text{-}pred}$ en fonction de log Ir_{exp}.

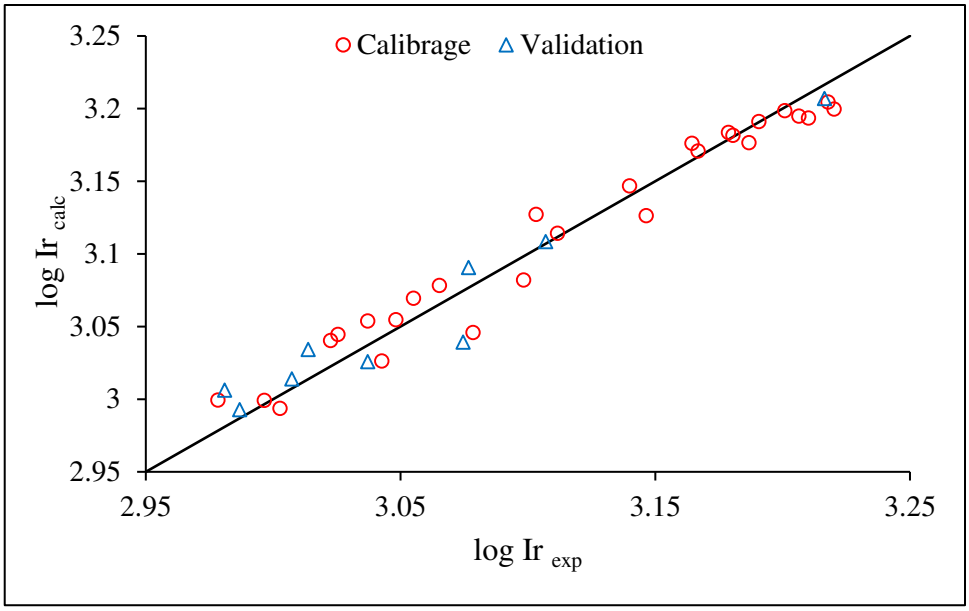


Figure I. 5 : Qualité de l'ajustement.

La figure ci- dessus (figure I. 5) montre que les valeurs calculées et prédites des deux ensembles de calibrage et de validation, sont en bon accord avec les valeurs expérimentales correspondantes, ceci confirmé par la valeur de Q^2_{LOO} = 92,04%.

I. 3. Machines à vecteurs supports SVM :

Après la mise en place du modèle MLR et RNA, une régression SVM a été utilisée pour développer un modèle sur les composés de l'ensemble de calibrage, sur la base du même sous-ensemble de descripteurs.

Dans notre travail, le modèle SVM utilise la fonction de base radiale (RBF). Avec une procédure de réglage fin, nous avons essayé d'obtenir la plus faible racine de l'erreur quadratique moyenne (EQMC) liée au meilleur paramètre de régression en utilisant le leave-one- out (LOO) en tenant compte du $EQMP_{ext}$ de l'ensemble de test.

A fin de déterminer les valeurs optimales des paramètres du model SVM, nous avons fixé la valeur de C à 10, puis nous avons variés les valeurs d'epsilon et gamma. Une fois nous avons obtenus des valeurs d'epsilon et gamma correspond à des valeurs minimales de RMSE et $RMSE_{ext}$, on les fixe et on fait varié le C pour obtenir des valeurs petites et proches de RMSE et $RMSE_{ext}$.

Les valeurs optimales obtenues pour les paramètres SVM et les résultats de la régression sont présentés dans le tableau I. 7.

Tableau I. 7 : Paramètres et résultats du modèle SVM.

C	γ	ε	R^2	Q^2_{loo}	Q^2_{ext}	EQMC	EQMP	$EQMP_{ext}$
11,12	0,1111	0,2222	93,93%	92,02%	96, 86%	0,0354	0,0376	0,0316

La matrice de corrélation reproduite dans le tableau I. 8 montre que le logarithme de l'indice de rétention pour les 27 composés utilisés pour le calibrage est bien corrélé avec les quatre descripteurs d'où la grande valeur du coefficient de détermination R^2.

Tableau I. 8 : Matrice de corrélation.

	log Ir	piPC02	ChiA_B(p)	SM2_B(s)
piPC02	0,7400			
ChiA_B(p)	0,0070	0,2180		
SM2_B(s)	0,4170	0,1290	0,0190	
Mor15u	0,2970	0,0840	0,0170	0,0270

Les log Ir observés et prédits de l'ensemble de calibrage et de l'ensemble de validation sont présentés dans la figure I. 6. Les valeurs calculées sont, en général assez dispersés autour de la première bissectrice.

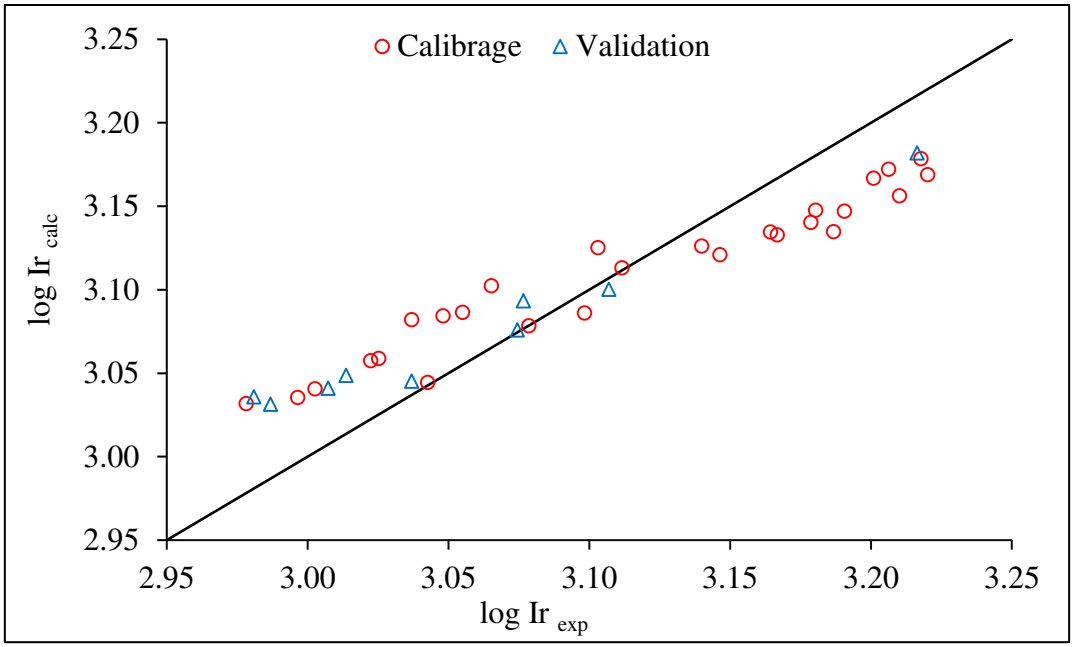


Figure I. 6 : Graphe des valeurs calculées, prédites en fonction des valeurs expérimentales.

La comparaison entre les résultats des méthodes MLR, RNA et SVM ont été utilisées pour prédire le logarithme de l'indice de rétention (log Ir) des composants d'huile essentielle sont présentés dans le tableau I. 9 :

Tableau I. 9 : Comparaison entre les résultats des modèles MLR, RNA et SVM.

	Méthodes			
		MLR	RNA	SVM
Calibrage Set= 27	R^2	97,81%	96,52%	93,93%
	Q^2_{loo}	96,91%	92,04%	92,02%
	EQMC	0,0110	0,0147	0,0354
	EQMP	0,0130	0,0218	0,0376
Validation Set = 9	Q^2_{ext}	95,46%	94,08%	96,86%
	$EQMP_{ext}$	0,0160	0,0176	0,0310

I. 4. Corrélation valeurs calculées – valeurs mesurées et intervalles de confiance :

D'après les valeurs des X et Y (tableau I. 10) on calcule les paramètres des méthodes MLR, RNA et SVM qui sont reproduits dans le tableau I. 11 pour les 27 composés de l'ensemble de calibrage.

Tableau I. 10 : Logarithmes des Ir mesurés X= $(\log Ir)_{Exp}$ et calculés Y= $(\log Ir)_{Calc.}$

		MLR	RNA	SVM
Composés	**X=log Ir $_{Exp}$**	**Y=log Ir $_{Calc}$**	**Y=log Ir $_{Calc}$**	**Y=log Ir $_{Calc}$**
1	2,9782	2,9944	2,9994	3,0320
2	2,9965	2,9896	2,9992	3,0356
3	3,0026	2,9958	2,9939	3,0407
4	3,0224	3,0409	3,0406	3,0576
5	3,0253	3,0355	3,0446	3,0590
6	3,0370	3,0539	3,0540	3,0822
7	3,0426	3,0247	3,0265	3,0446
8	3,0481	3,0578	3,0549	3,0845
9	3,0550	3,0610	3,0696	3,0865
10	3,0652	3,0746	3,0785	3,1025
11	3,0785	3,0490	3,0460	3,0785
12	3,0983	3,1038	3,0822	3,0862
13	3,1031	3,1069	3,1273	3,1253
14	3,1116	3,0935	3,1142	3,1132
15	3,1399	3,1350	3,1469	3,1263
16	3,1464	3,1500	3,1264	3,1210
17	3,1667	3,1623	3,1710	3,1330
18	3,1644	3,1701	3,1762	3,1348
19	3,1787	3,1785	3,1838	3,1404
20	3,1804	3,1769	3,1818	3,1478
21	3,1867	3,1735	3,1768	3,1349
22	3,1906	3,1908	3,1912	3,1472
23	3,2009	3,2100	3,1988	3,1669
24	3,2063	3,2103	3,1950	3,1724

Tableau I. 10 : suite et fin

25	3,2101	3,2001	3,1936	3,1563
26	3,2177	3,2280	3,2047	3,1786
27	3,2201	3,2233	3,1997	3,1690

Tableau I. 11 : Les valeurs des grandeurs T_X, T_Y, $\overline{X}$,$\overline{Y}$, S_{xx}, S_{yy}, S_{xy}.

	T_x	**T_y**	**$\overline{x}$**	**$\overline{y}$**	**S_{xx}**	**S_{yy}**	**S_{xy}**
MLR	84,0732	84,0902	3,1138	3,1145	0,1579	0,1548	0,1546
RNA	84,0732	84,0765	3,1138	3,1139	0.1579	0.1381	0.1451
SVM	84,0732	83,9570	3,1138	3.1095	0,1579	0,0532	0.0889

On utilise ces quantités pour trouver les valeurs de a et b, soit :

MLR :

$$b = \frac{S_{xy}}{S_{xx}} = 0,9793$$

$$a = \overline{y} - b\overline{x} = 0,0652$$

RNA :

$$b = \frac{S_{xy}}{S_{xx}} = 0,9189$$

$$a = \overline{y} - b\overline{x} = 0,2526$$

SVM :

$$b = \frac{S_{xy}}{S_{xx}} = 0,5629$$

$$a = \overline{y} - b\overline{x} = 1,3569$$

On en déduit les droites des moindres carrés :

MLR : Y= 0,0652 + 0,9793x

RNA : Y= 0,2526 + 0,9189x

SVM : Y= 1,3569 + 0,5629x

Tableau I. 12 : L'analyse de variance.

Source	SQ		∅	DQ
Due à la régression	MLR	0,1514		0,1514
	RNA	0,1333		0,1333
	SVM	0,0500		0,0500
Autour de la régression	MLR	0,0034	25	0,0001
	RNA	0,0048		0,0002
	SVM	0,0032		0,0001
Totale	MLR	0,1548	26	0,0060
	RNA	0,1381		0,0053
	SVM	0,0532		0,0020

On calcule la variance V(y) et l'erreur $\sqrt{V(y)}$:

MLR :

$V(y) = S_2^2 = 0{,}0001$

$\sqrt{V(y)} = SE(y) = 0{,}0116$

$SE(a) = S_2(a) = 0{,}0912$

$SE(b) = S_2(b) = 0{,}0293$

RNA :

$V(y) = S_2^2 = 0{,}0002$

$\sqrt{V(y)} = SE(y) = 0{,}0139$

$SE(a) = S_2(a) = 0{,}1088$

$SE(b) = S_2(b) = 0{,}0349$

SVM :

$V(y) = S_2^2 = 0{,}0001$

$\sqrt{V(y)} = SE(y) = 0{,}0114$

$SE(a) = S_2(a) = 0{,}0891$

$SE(b) = S_2(b) = 0{,}0286$

Ce qui permet le calcul de a_{sup} et a_{inf} ainsi que b_{sup} et b_{inf} :

Sachant que pour P=0,95, $t_{0,95\ ;\ 25}$= 1,708, il vient :

MLR:

a_{sup} = 0,2210

a_{inf} = -0,0904

b_{sup} = 1,0292

b_{inf} = 0,9292

RNA:

a_{sup} = 0,2210

a_{inf} = -0,0904

b_{sup} = 1,0292

b_{inf} = 0,9292

SVM:

a_{sup} = 0,2210

a_{inf} = -0,0904

b_{sup} = 1,0292

b_{inf} = 0,9292

Tableau I. 13 : Logarithmes des Ir mesurés X= log Ir_{Exp} et calculés Y= log Ir_{Calc} et les limites de confiance.

		MLR					RNA					SVM				
N°	**X**	**Y**	***SEY***	Y_{sup}	Y_{inf}	Y_{sup}-Y_{inf}	**Y**	***SEY***	Y_{sup}	Y_{inf}	Y_{sup}-Y_{inf}	**Y**	***SEY***	Y_{sup}	Y_{inf}	Y_{sup}-Y_{inf}
1	2,9782	2,9944	0,0125	3,0030	2,9603	0,0427	2,9994	0,0149	3,0148	2,9639	0,0509	3,0320	0,0122	3,0541	3,0123	0,0417
2	2,9965	2,9896	0,0123	3,0206	2,9785	0,0421	2,9992	0,0147	3,0313	2,9810	0,0502	3,0356	0,0121	3,0641	3,0229	0,0412
3	3,0026	2,9958	0,0123	3,0265	2,9846	0,0420	2,9939	0,0147	3,0368	2,9867	0,0501	3,0407	0,0120	3,0674	3,0264	0,0410
4	3,0224	3,0409	0,0121	3,0457	3,0042	0,0415	3,0406	0,0145	3,0547	3,0052	0,0495	3,0576	0,0119	3,0784	3,0378	0,0406
5	3,0253	3,0355	0,0121	3,0485	3,0071	0,0414	3,0446	0,0145	3,0573	3,0079	0,0494	3,0590	0,0119	3,0800	3,0395	0,0405
6	3,0370	3,0539	0,0121	3,0598	3,0186	0,0412	3,0540	0,0144	3,0679	3,0188	0,0491	3,0822	0,0118	3,0864	3,0462	0,0403
7	3,0426	3,0247	0,0120	3,0652	3,0242	0,0411	3,0265	0,0143	3,0730	3,0240	0,0490	3,0446	0,0118	3,0895	3,0494	0,0402
8	3,0481	3,0578	0,0120	3,0706	3,0296	0,0410	3,0549	0,0143	3,0780	3,0291	0,0489	3,0845	0,0117	3,0926	3,0525	0,0401
9	3,0550	3,0610	0,0120	3,0773	3,0364	0,0409	3,0696	0,0143	3,0843	3,0355	0,0488	3,0865	0,0117	3,0964	3,0564	0,0400
10	3,0652	3,0746	0,0119	3,0872	3,0465	0,0407	3,0785	0,0142	3,0936	3,0450	0,0486	3,1025	0,0117	3,1021	3,0622	0,0398
11	3,0785	3,0490	0,0119	3,1002	3,0596	0,0406	3,0460	0,0142	3,1057	3,0573	0,0484	3,0785	0,0116	3,1095	3,0698	0,0397
12	3,0983	3,1038	0,0119	3,1195	3,0790	0,0405	3,0822	0,0141	3,1238	3,0755	0,0483	3,0862	0,0116	3,1206	3,0810	0,0396
13	3,1031	3,1069	0,0118	3,1242	3,0837	0,0405	3,1273	0,0141	3,1282	3,0800	0,0483	3,1253	0,0116	3,1233	3,0837	0,0396
14	3,1116	3,0935	0,0118	3,1325	3,0920	0,0405	3,1142	0,0141	3,1360	3,0878	0,0483	3,1132	0,0116	3,1280	3,0885	0,0396
15	3,1399	3,1350	0,0119	3,1603	3,1197	0,0405	3,1469	0,0142	3,1621	3,1137	0,0484	3,1263	0,0116	3,1440	3,1044	0,0396
16	3,1464	3,1500	0,0119	3,1666	3,1261	0,0406	3,1264	0,0142	3,1681	3,1197	0,0484	3,1210	0,0116	3,1477	3,1080	0,0397
17	3,1667	3,1623	0,0119	3,1866	3,1458	0,0408	3,1710	0,0142	3,1869	3,1382	0,0487	3,1330	0,0117	3,1592	3,1193	0,0399

Tableau I. 13 : Suite et fin

18	3,1644	3,1701	0,0119	3,1844	3,1436	0,0408	3,1762	0,0142	3,1847	3,1361	0,0486	3,1348	0,0117	3,1579	3,1180	0,0399
19	3,1787	3,1785	0,0120	3,1985	3,1575	0,0410	3,1838	0,0143	3,1980	3,1491	0,0489	3,1404	0,0117	3,1661	3,1260	0,0401
20	3,1804	3,1769	0,0120	3,2001	3,1591	0,0410	3,1818	0,0143	3,1996	3,1507	0,0489	3,1478	0,0117	3,1670	3,1269	0,0401
21	3,1867	3,1735	0,0120	3,2064	3,1653	0,0411	3,1768	0,0144	3,2054	3,1564	0,0490	3,1349	0,0118	3,1706	3,1304	0,0402
22	3,1906	3,1908	0,0121	3,2102	3,1690	0,0412	3,1912	0,0144	3,2091	3,1599	0,0491	3,1472	0,0118	3,1729	3,1326	0,0403
23	3,2009	3,2100	0,0121	3,2204	3,1790	0,0414	3,1988	0,0145	3,2186	3,1693	0,0494	3,1669	0,0118	3,1787	3,1383	0,0405
24	3,2063	3,2103	0,0121	3,2258	3,1843	0,0415	3,1950	0,0145	3,2237	3,1742	0,0495	3,1724	0,0119	3,1818	3,1413	0,0406
25	3,2101	3,2001	0,0122	3,2295	3,1879	0,0416	3,1936	0,0145	3,2272	3,1776	0,0496	3,1563	0,0119	3,1840	3,1434	0,0407
26	3,2177	3,2280	0,0122	3,2371	3,1953	0,0418	3,2047	0,0146	3,2343	3,1845	0,0498	3,1786	0,0120	3,1884	3,1475	0,0408
27	3,2201	3,2233	0,0122	3,2394	3,1976	0,0418	3,1997	0,0146	3,2366	3,1867	0,0499	3,1690	0,0120	3,1898	3,1489	0,0409
28*	2,9809	2,9843	0,0125	3,0056	2,9630	0,0426	2,9918	0,0149	3,0172	2,9664	0,0508	3,0347	0,0122	3,0555	3,0139	0,0416
29*	2,9868	2,9901	0,0124	3,0113	2,9689	0,0424	2,9972	0,0148	3,0225	2,9719	0,0506	3,0380	0,0121	3,0588	3,0173	0,0415
30*	3,0073	3,0101	0,0122	3,0311	2,9892	0,0418	3,0161	0,0146	3,0410	2,9911	0,0499	3,0496	0,0120	3,0700	3,0291	0,0409
31*	3,0137	3,0164	0,0122	3,0372	2,9956	0,0417	3,0220	0,0146	3,0468	2,9971	0,0497	3,0532	0,0119	3,0736	3,0328	0,0408
32*	3,0370	3,0392	0,0121	3,0598	3,0186	0,0412	3,0434	0,0144	3,0679	3,0188	0,0491	3,0663	0,0118	3,0864	3,0462	0,0403
33*	3,0745	3,0759	0,0119	3,0963	3,0556	0,0406	3,0778	0,0142	3,1021	3,0536	0,0485	3,0874	0,0116	3,1073	3,0675	0,0397
34*	3,0766	3,0780	0,0119	3,0983	3,0577	0,0406	3,0797	0,0142	3,1040	3,0555	0,0485	3,0886	0,0116	3,1084	3,0687	0,0397
35*	3,1069	3,1077	0,0118	3,1279	3,0874	0,0405	3,1076	0,0141	3,1317	3,0835	0,0483	3,1056	0,0116	3,1254	3,0858	0,0396
36*	3,2164	3,2149	0,0122	3,2358	3,1940	0,0417	3,2082	0,0146	3,2331	3,1833	0,0498	3,1672	0,0119	3,1876	3,1468	0,0408

* Composés de validation

Les log Ir calculés reproduisent assez bien les log Ir mesurés (tableau I. 10) et les limites de confiance pour chaque point sont plutôt étroites pour le niveau de confiance P = 0,95 et égales à 0,0412 pour MLR, 0,0492 pour RNA et 0,0403 pour SVM en moyenne.

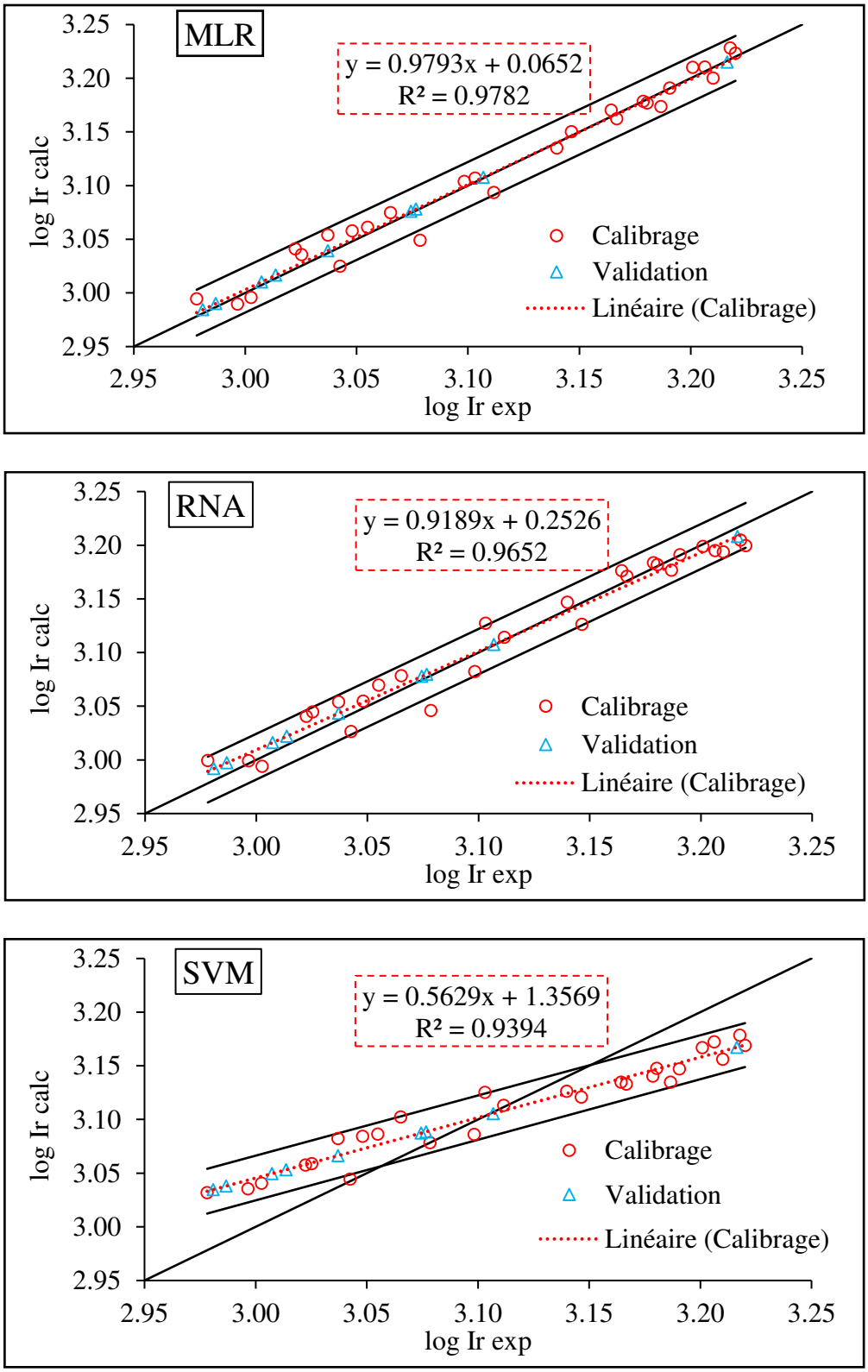


Figure I. 7 : Vérification des limites de confiance.

I. 4. 1. Comparaison des droites de régression :

Tableau I. 14 : Comparaison des droites de régression : droite I : MLR ; droite II : RNA ; droite III : SVM – Calculs intermédiaires.

	MLR	**RNA**	**SVM**
	Droite I	**Droite II**	**Droite III**
N	27	27	27
$\sum x_i$	84,073181319	84,073181319	84,073181319
$\sum y_i$	84,090200000	84,076530000	83,956980000
$\sum x_i^2$	7068,299817062	7068,299817062	7068,299817062
$\sum y_i^2$	7071,161736040	7068,862896841	7048,774490720
$\sum x_i y_i$	261,996478117	261,944381616	261,515913960
$\overline{x}$	3,113821530	3,113821530	3,113821530
$\overline{y}$	3,114451852	3,113945556	3,109517778
S_{xx}	0,157874995	0,157874995	0,157874995
S_{xy}	0,154602868	0,145072307	0,088862015
S_{yy}	0,154774747	0,138115364	0,053244880
a	0,065167586	0,252635735	1,356862414
b	0,979273936	0,918906171	0,562863140
S_2^2	0,000135048	0,000192301	0,000129109
S_1^2	0,151398559	0,133307838	0,050017153
S_b^2	0,000855408	0,001218059	0,000817793
x_0	3,1116	3,1116	3,1116
$x_0 - \overline{x}$	-0,0022221530	-0,002221530	-0,002221530
Y	3,112276365	3,111904178	3,108267360
S_Y^2	3,10145E-06	4,42E-06	2,96507E-06

I. 4. 1. 1. Comparaison des droites (I) et (II) – [cf. tableau I. 14] :

*** Comparaison des ordonnées des deux droites au point moyen :**

$$t = \frac{|Y_I - Y_{II}|}{\sqrt{S_{Y_I}^2 + S_{Y_{II}}^2}} = \frac{0,000372187}{0,002741852} = 0,135743096$$

D'après la table de Student pour $N_I+N_{II}-4 = 50$ degrés de liberté, la limite de t est : $t_{0,975}=2,009$ au niveau de confiance 95% et $t_{0,995}=2,678$ au niveau de confiance 99%.

La différence des valeurs des ordonnées au point moyen n'est donc pas significative.

*** Comparaison des coefficients angulaires :**

$$t = \frac{|b_I - b_{II}|}{\sqrt{S_{b_I}^2 + S_{b_{II}}^2}} = \frac{0,060367765}{0,045535338} = 1,325734404$$

Cette valeur étant inférieure aux limites données par la table de Student, aux niveaux de confiance 95 % et 99 %, pour 50 degrés de liberté, le test n'est pas significatif.

*** Comparaison des variances résiduelles :**

Pour le risque $\alpha = 0,05$ partagé en α_1 et α_2 tels que $\alpha_1 + \alpha_2 = 0,025$ et les nombres de degrés de liberté $\nu_1 = \nu_2 = 27$, la table de Snedecor donne :

$$F_{1-\alpha_2} = F_{0,975} = 2,19$$

$$F_{\alpha_1} = \frac{1}{F_{1-\alpha_2}} = \frac{1}{2,19} = 0,4566$$

$$F = \frac{S_{2\,I}^2}{S_{2\,II}^2} = \frac{0,000135048}{0,000192301} = 0,70227152$$

La valeur expérimentale F=0,7002 étant comprise entre les valeurs $F_{1-\alpha_2}$ et F_{α_1}, l'égalité des variances ($\sigma_1^2 = \sigma_2^2$) ne sera pas refusée dans ce cas.

Ainsi, les tests font ressortir l'absence de translation ou de rotation de l'une des droites de régression par rapport à l'autre.

I. 4. 1. 2. Comparaison des droites (I) et (III) – [cf. tableau I. 14] :

*** Comparaison des ordonnées des deux droites au point moyen :**

$$t = \frac{|Y_I - Y_{III}|}{\sqrt{S_{Y_I}^2 + S_{Y_{III}}^2}} = \frac{0,004009005}{0,002463029} = 1.627672633$$

D'après la table de Student pour $N_I+N_{II}-4 = 50$ degrés de liberté, la limite de t est : $t_{0,975}=2,009$ au niveau de confiance 95% et $t_{0,995}=2,678$ au niveau de confiance 99%.

La différence des valeurs des ordonnées au point moyen n'est donc pas significative.

*** Comparaison des coefficients angulaires :**

$$t = \frac{|b_I - b_{III}|}{\sqrt{S_{b_I}^2 + S_{b_{III}}^2}} = \frac{0,060367765}{0,045535338} = 10.18000275$$

Cette valeur étant supérieur aux limites données par la table de Student, aux niveaux de confiance 95 % et 99 %, pour 50 degrés de liberté, le test est significatif.

*** Comparaison des variances résiduelles :**

Pour le risque $\alpha = 0,05$ partagé en α_1 et α_2 tels que $\alpha_1 + \alpha_3 = 0,025$ et les nombres de degrés de liberté $\nu_1 = \nu_3 = 27$, la table de Snedecor donne :

$$F_{1-\alpha_3} = F_{0,975} = 2,19$$

$$F_{\alpha_1} = \frac{1}{F_{1-\alpha_3}} = \frac{1}{2,19} = 0,4566$$

$$F = \frac{S_{2\,I}^2}{S_{2\,III}^2} = \frac{0,000135048}{0,000129109} = 1,045995731$$

La valeur expérimentale F=1,0460 étant comprise entre les valeurs $F_{1-\alpha_3}$ et F_{α_1}, l'égalité des variances ($\sigma_1^2 = \sigma_3^2$) ne sera pas refusée dans ce cas.

Ainsi, les tests font ressortir l'absence de translation et il y a d'une rotation de l'une des droites de régression par rapport à l'autre.

I. 4. 1. 3. Comparaison des droites (II) et (III) – [cf. tableau I. 14] :

*** Comparaison des ordonnées des deux droites au point moyen :**

$$t = \frac{|Y_{II} - Y_{III}|}{\sqrt{S^2_{Y_{II}} + S^2_{Y_{III}}}} = \frac{0,003636817}{0,002716868} = 1,338606568$$

D'après la table de Student pour $N_I+N_{II}-4 = 50$ degrés de liberté, la limite de t est : $t_{0,975}=2,009$ au niveau de confiance 95% et $t_{0,995}=2,678$ au niveau de confiance 99%.

La différence des valeurs des ordonnées au point moyen n'est donc pas significative.

*** Comparaison des coefficients angulaires :**

$$t = \frac{|b_{II} - b_{III}|}{\sqrt{S^2_{b_{II}} + S^2_{b_{III}}}} = \frac{0,356043031}{0,045120417} = 7,890951666$$

Cette valeur étant supérieur aux limites données par la table de Student, aux niveaux de confiance 95 % et 99 %, pour 50 degrés de liberté, le test est significatif.

*** Comparaison des variances résiduelles :**

Pour le risque $\alpha = 0,05$ partagé en α_1 et α_2 tels que $\alpha_2 + \alpha_3 = 0,025$ et les nombres de degrés de liberté $\nu_2 = \nu_3 = 27$, la table de Snedecor donne :

$$F_{1-\alpha_3} = F_{0,975} = 2,19$$

$$F_{\alpha_2} = \frac{1}{F_{1-\alpha_3}} = \frac{1}{2,19} = 0,4566$$

$$F = \frac{S^2_{2\,II}}{S^2_{2\,III}} = \frac{0,000192301}{0,000129109} = 1,489446321$$

La valeur expérimentale F=1,4894 étant comprise entre les valeurs $F_{1-\alpha_3}$ et F_{α_2}, l'égalité des variances ($\sigma_2^2 = \sigma_3^2$) ne sera pas refusée dans ce cas.

Ainsi, les tests font ressortir l'absence de translation et il y a d'une rotation de l'une des droites de régression par rapport à l'autre.

Conclusion Générale

Conclusion Générale

Le présent travail s'inscrit dans le cadre de l'étude de l'indice de rétention des composés des huiles essentielles en utilisant la méthode de modélisation QSRR.

Nous avons utilisé la méthodologie QSRR pour relier l'indice de rétention d'une série des composés des huiles essentielles, à des descripteurs moléculaires théoriques caractéristiques de la molécule entière ou de ses fragments, calculés à l'aide de logiciels spécialisés du commerce.

L'objectif est de développer des modèles QSRR stables, fiables et prédictifs pour la prédiction de l'indice de rétention d'une série constituée de 36 composés des huiles essentielles.

La méthodologie QSRR standard a été utilisée dans ce travail :

- ✓ Elaboration du modèle pour une série d'apprentissage et analyse des paramètres statistiques de modèle obtenu.
- ✓ Vérification de la stabilité interne du modèle obtenu avec la validation croisée (Q^2_{LOO}).
- ✓ Test la stabilité du modèle avec une série externe.
- ✓ La Y-randomisation pour vérifier que le modèle QSRR obtenu n'est pas dû au hasard.

Les 36 composés ont été divisés selon l'algorithme de Kennard et Stone (CADEX) en deux ensembles disjoints :

- ♣ Un ensemble principal de 27 composés utilisés pour la construction de modèle (ensemble de calibrage).
- ♣ Un ensemble de 9 composés pour la prédiction externe (ensemble de validation).

Nous avons recherché des corrélations linéaires entre variables dépendantes et variables explicatives sélectionnées par algorithme génétique « dans la version MOBYDIGS de TODESCHINI, en maximisant Q^2_{L00} », en utilisant la régression linéaire multiple.

Enfin, nous avons recherché des corrélations non linéaires en utilisant les réseaux de neurones standards à 3 couches (les entrées, une couche cachée et une couche de sortie), et les Machines à vecteurs supports.

La qualité de l'ajustement de modèle développé a été vérifié en procédant à la représentation des valeurs calculées en fonction du celles observées.

Le domaine d'application de modèle a été étudié à l'aide du diagramme de Williams, ce dernier fait ressortir parmi les composés de l'ensemble de calibrage et de validation les composés influents et aberrants.

Le test de randomisation est un outil puissant pour vérifier et s'assurer que le modèle obtenu n'est pas dû au hasard.

Le modèle obtenu montre que la relation entre l'indice de rétention et les descripteurs moléculaires est linéaire.

La comparaison de la qualité des modèles MLR, SVM et RNA montrent qu'il n'y a pas une grande différence tant par la qualité de l'ajustement, la robustesse ou la capacité prédictive.

Étant donnée cela et vu la simplicité de la régression linéaire multiple la modélisation de l'indice de rétention par approche MLR est la meilleure.

Références Bibliographiques

[1] Pierron C., 2014. Thèse de Doctorat, Université de Lorraine. France. "Les huiles essentielles et leurs expérimentations dans les services hospitaliers de France : exemples d'applications en gériatrie-gérontologie et soins palliatifs".

[2] Frank J., 2007. "Introduction to Computational Chemistry". Wiley, 2nd edition. 599 pages.

[3] Charton M. Charton B I., 1999. "Advances in Quantitative Structure-Property Relationships". Vol 2. Stamford, Conn.: JAI Press.

[4] Guendouzi A., 2015. Thèse de Doctorat, Université Abou Bekr Belkaïd de Tlemcen. "Élaboration des modèles QSPR prédictifs des propriétés physico-chimiques à l'aide des descripteurs moléculaires".

[5] Lardry J M. et Haberkorn V., 2007. "L'aromathérapie et les huiles essentielles". Revue de Kinésithérapie. 7(61), pp.14-17.

[6] Lucchesi M E., 2005. Thèse de Doctorat, Université de la Réunion. "Extraction sans solvant assistée par micro-ondes conception et application à l'extraction des huiles essentielles".

[7] Tabet Zatla A, 2017. Thèse de Doctorat, Universite Abou-Bekr Belkaid. Tlemcen, Algérie. "Caractérisations chimiques et études biologiques d'extraits de quatre plantes aromatiques "*Daucus. carota ssp. sativus*, *Marrubium vulgare*, *Ballota nigra et Cynoglossum cheirifolium*" de la région de Tlemcen".

[8] Fouché J G., Marquet A. et Hambuckers A., 2000. "Les plantes médicinales, de la plante au médicament". Observatoire du monde des plantes Sart-Tilman.

[9] Bruneton J., 1999. "Pharmacognosie: phytochimie, plantes médicinales". 3ème édition, Technique et documentation Lavoisier, Paris. 1120 pages.

[10] Balz R., 1986. "Les huiles essentielles et comment les utiliser". Ed Lavoisier. Paris. 155 pages.

[11] AFNOR NF T 75-006., 1998. Matières premières aromatiques d'origine naturelle - Vocabulaire.

[12] Bruneton J., 2009. "Pharmacognosie: phytochimie, plantes médicinales". 4ème édition, Technique et documentation Lavoisier, Paris. 1269 pages.

[13] Pellecuer J., Jacob M., Simeon B M De., Dusart G., Attisso M., Barthez M., Gourgas L., Pascal B. et Tomei B., 1980. "Essais d'utilisation d'huiles essentielles de plantes aromatiques méditerranéennes en odontologie conservatrice". *Plantes Médicinales et Phytothérapie*. 14, pp. 83-98.

[14] Viollon C., Chaumont J P. et Leger D., 1993. "Activités antagonistes in vitro de certains composés volatils naturel vis-à-vis de germes de la flore vaginale". *Plantes Médicinales et Phytothérapie*. 26 (4), pp. 17-22.

[15] Chaumont J P. et Leger D., 1989. "Propriétés antifongiques de quelques phénols et composés chimiquement très voisins. Relations structure-activité". *Plantes médicinales et phytothérapie*, 23(2). pp. 124.

[16] Sivropoulou A., Papanikolaou E., Nikolaou C., Kokkini S., Lanaras T. et Arsenakis M., 1996. "Antimicrobial and Cytotoxic Activities of Origanum Essential Oils". *Journal of Agricultural and Food Chemistry*. 44 (5), pp. 1202-1205.

[17] Zambonelli A., D'Aurelio A Z., Severi A., Benvenuti E., Maggi L. et Bianchi A., 2004. "Chemical Composition and Fungicidal Activity of Commercial Essential Oils of *Thymus vulgaris L*". *Journal of Essential Oil Research*. 16 (1), pp. 69-74.

[18] Mangena T. et Muyima N Y O., 1999. "Comparative evaluation of the antimicrobial activities of essential oils of Artemisia afra, Pteronia incana and Rosmarinus officinalis on selected bacteria and yeast strains". *Letters in Applied Microbiology*. 28(4), pp. 291-294.

[19] Robert H., Waterman K M. et Peter G., 1993. "Volatile oil crops: their biology, biochemistry and production". Harlow: Longman Scientific and Technical.

[20] Torssell Kurt B G., 1983. "Natural Products Chemistry". John Wiley & Sons Limited. 401 pages.

[21] Croteau R., 1987. "Biosynthesis and catabolism of monoterpenoids". *Chemical Reviews*. 87(5), pp. 929-954.

[22] Bakkali F., Averbecka S., Averbeck D. et Idaomar M., 2008. "Biological effects of essential oils – A review". *Food and Chemical Toxicology*. 46 (2), pp. 446-475**.**

[23] Teuscher E., Anton R. et Lobstein A., 2005. "Plantes aromatiques-Epices, aromates, condiments et huiles essentielles". Éditions Tec et Doc, Lavoisier Paris. 521 pages.

[24] Bruneton J., 1987. "Élément de phytochimie et pharmacognosie". Technique et documentation Lavoisier, Paris. 585 pages.

[25] Kurkin V A., 2003. "Phenylpropanoids from Medicinal Plants: Distribution, Classification, Structural Analysis, and Biological Activity". *Chemistry of Natural Compounds*. 39 (2), pp. 123-153.

[26] Padua L S De., Bunyapraphatsara N. et Lemmens R H M J., 1999. "Plant resources of South-East Asia No.12(1). Medicinal and poisonous plants 1". 711 pages.

[27] Bruneton J., 1993. "Pharmacognosie: phytochimie, plantes médicinales". 2ème édition, Technique et documentation Lavoisier, Paris. 915 pages.

[28] Pavia D L., Lampman G M. et Kriz G S., 1976. "Introduction to organic laboratory techniques: a contemporary approach". Philadelphia : Saunders. 699 pages.

[29] El Haib A., 2011. Thèse de Doctorat, Université de Toulouse, France. "Valorisation de terpènes naturels issus de plantes marocaines par transformations catalytiques".

[30] Meyer-Warnod B., 1984. "Natural essential oils: extraction processes and applications to some major oils". *Perfumer & Flavorist*. 9, pp. 93-103.

[31] Martini M C. et Seiller M., 1999. "Actifs et additifs en cosmétologie". 2ème édition, Editions Tec et Doc, Editions médicales internationales Paris. Lavoisier, 656 pages.

[32] Wang Z., Ding L., Li. T., Zhou X., Wang L., Zhang H., Liu L., Li Y., Liu Z., Wang H., Zeng H. et He H., 2006. "Improved solvent-free microwave extraction of essential oil from dried *Cuminum cyminum* L. and *Zanthoxylum bungeanum* Maxim". *Journal of Chromatography A*. 1102 (1-2), pp. 11-17.

[33] Kim N S. et Lee D S., 2002. "Comparison of different extraction methods for the analysis of fragrances from Lavandula species by gas chromatography–mass spectrometry". *Journal of Chromatography A*. 982 (1), pp. 31-47.

[34] Hubert R., 1992. "Epices et aromates". Tec et Doc – Lavoisier, APRIA., Paris. 340 pages.

[35] Joulain D., 1994. "Methods for analyzing essential oils. Modern analysis methodologies: use and abuse". *Perfumer and Flavorist*. 19 (5), pp. 5-17.

[36] Julien P., 2005. Thèse de Doctorat, Université de Corse Pascal Paoli. "Caractérisation des huiles essentielles par CPG/IR, CPG/SM-(IE et IC) et RMN du carbone-13 de *cistus*

albidus et de deux asteraceae endémiques de Corse : *eupatorium cannabinum* subsp. *corsicum* et *doronicum corsicum*".

[37] Adams R P., 2001. Identification of essential oil Components by gas chromatography/ quadrupole mass spectroscopy, 3rd ed, Allured, Carol Stream Ill: USA. 804 pages.

[38] Souici M L., Lourici L. et Messadi D., 2007. "Relation structure/retention chromatographique de treize alkylnaphtalenes". *Lebanese Science Journal*. 8 (1), pp. 64.

[39] Lourici L. et Messadi D., 2009. "Methodes non lineaires pour le calcul des indices de retention en chromatographie gazeuse à programmation linéaire de température". *Lebanese Science Journal*. 10 (1), pp. 77-86.

[40] Bouchonnet S. et Libong D., 2002. "Le couplage chromatographie en phase gazeuse-spectrométrie de masse [Gas chromatography-mass spectrometry coupling]". *L'Actualité Chimique*, Société chimique de France. pp.7-14.

[41] Besombes C., 2008. Thèse de Doctorat, Université de la Rochelle. France. "Contribution à l'étude des phénomènes d'extraction hydro-thermomécanique d'herbes aromatiques. Applications généralisées". pp. 98.

[42] Bouchonnet S., 2001. "Comparaison des performances des analyseurs quadrupolaires en spectrométrie de masse: trappes ioniques versus quadripôles". *Spectra Analyse*, Paris : PCI, 222, pp. 11-18.

[43] March R E., 1997. "An introduction to quadrupole ion trap mass spectrometry". *Journal of Mass Spectrometry*. 32(4), pp. 351-369.

[44] Bouchonnet S., Hoppilliard Y. et Kargar-Grisel T., 1999. "Les différents types de spectromètres de masse utilisées pour l'analyse des composés organiques et bio-organiques". *Spectra Analyse*, 28 (207), pp. 11-25.

[45] Bouderdara N., 2013. Thèse de Doctorat en Sciences. Université Mentouri-Constantine, "Séparation et détermination de structures des métabolites secondaires de *Cachrys libanotis* L". p. 103.

[46] Van den Dool H. et Kratz P D., 1963. "A generalization of the retention index system including linear temperature programmed gas-liquid partition chromatography". *Journal of Chromatography A*. 11, pp. 463-471.

[47] Messadi D., Souici M L. et Lourici L., 2013. "Interpolation par fonction B-splines pour le calcul des indices de rétention en programmation de température : application à un mélange de phénols tests séparés sur des colonnes garnies de Tenax-GC modifié". *Revue des Sciences et de la Technologie., Synthèse*. 27, pp. 89-98.

[48] Cu J Q., 1990. Thèse de Doctorat, Institut National Polytechnique. Toulouse, France. "Extraction de compositions odorantes végétales par divers solvants organiques".

[49] Paris M. et Hurabielle M., 1981. "Abrégé de Matière Médicale (Pharmacognosie)". Tome 1, Généralités, monographies, Paris. 339 pages.

[50] Cavalli J F., 2002. Thèse de Doctorat, Université de Corse Pascal Paoli, "Caractérisation par CPG/IK, CPG/SM et RMN du carbone-13 d'huiles essentielles de Madagascar".

[51] Purchon N., 2001. "La bible de l'aromathérapie Edition Marabout". 414 pages.

[52] Willem J P., 2002. "Le guide des huiles essentielles pour vaincre vos problèmes de santé aromathérapie: médecine d'avenir". Editions LMV. Paris. 318 pages.

[53] Florence M., 2012. Thèse de Doctorat, Université de Lorraine. France. " Utilisations thérapeutiques des huiles essentielles : Etude de cas en maison de retraite ".

[54] Baudoux D., Blanchard J-M, et Malotaux A F., 2006. "Les cahiers pratiques d'aromathérapie selon l'école française". Volume 4 Soins palliatifs, Editions Inspir, Collection L'aromathérapie professionnellement. 318 pages.

[55] Cuba R., 2001. "Toxicity myths essential oils and their carcinogenic potential". *International Journal of Aromatherapy*. 11 (2), pp. 76-83.

[56] Raynaud J., 2006. "Prescription et conseil en aromathérapie". Éditions Tec et Doc, Lavoisier. 247 pages.

[57] AFNOR NF T 75-001., 1996. Huiles essentielles - Règles générales d'emballage, de conditionnement et de stockage.

[58] AFNOR NF T 75-002., 1996. Huiles essentielles - Règles générales d'étiquetage et de marquage des récipients.

[59] Larive C., 1997. Thèse de Doctorat, Ecole nationale des ponts et chaussées. France. "Apports combinés de l'expérimentation et de la modélisation à la compréhension de l'alcali-réaction et de ses effets mécaniques".

[60] Depret M H. et Hamdouch A., 2009. "Quelles politiques de l'innovation et de l'environnement pour quelle dynamique d'innovation environnementale? ". *Innovations*. 29 (1), pp. 127-147.

[61] Lerbet-Sereni F., 2004. "Expériences de la modélisation, modélisation de l'expérience". Editions L'Harmattan, Paris.

[62] Cancès E., Le Bris C. et Maday Y., 2006. "Méthodes mathématiques en chimie quantique. Une introduction, Mathématiques et Applications". Tome 53. Springer Science & Business.

[63] Hladik J., Chrysos M., Hladik P E. et Ancarani L. U., 1997. "Mécanique quantique, atomes et noyaux applications technologiques". 3ème édition, Dundo, Paris.

[64] Born M. et Oppenheimer J R., 1927. "On the quantum theory of molecules". *Annalen der Physik*. 84 (20), pp.457-484.

[65] Slater J C.,1929. "The Theory of Complex Spectra". *Physical Review.* 34 (10), pp. 1293-1322.

[66] Slater J C., 1974. "Quantum Theory of Molecules and Solids Vol. 4: The Self-Consistent Field for Molecules and Solids". New York, McGraw-Hill. pp. 583.

[67] Pauli W., 1925. "Über den Zusammenhang des Abschlusses der Elektronengruppen im Atom mit der Komplexstruktur der Spektren". *Zeitschrift für Physik*. 31 (1), pp. 765-783.

[68] Roothaan C C J., 1951. "New Developments in Molecular Orbital Theory". *Reviews of Modern Physics.* 23 (2), pp. 69-89.

[69] Hartree D R., 1928. "The Wave Mechanics of an Atom with a Non-Coulomb Central Field. Part I. Theory and Methods". Mathematical Proceedings of the Cambridge Philosophical Society. 24, pp. 89-110.

[70] Fock V., 1930. "Näherungsmethode zur Lösung des quantenmechanischen Mehrkörperproblems". *Zeitschrift für Physik*. 61, pp. 126-148.

[71] Møller C. et Plesset M S., 1934. "Note on an Approximation Treatment for Many-Electron Systems". *Physical Review*. 46 (7), pp. 618–622.

[72] Piron C., 1998. "Mécanique quantique : Bases et applications". 1ére Édition. Presses polytechniques et universitaires romandes (PPUR).

[73] Domingo L R., Chamorro E. et Pérez P., 2008. "Understanding the reactivity of captodative ethylenes in polar cycloaddition reactions. A theoretical study". *The Journal of Organic Chemistry.* 73 (12), pp. 4615-4624.

[74] Becke A D., 1993. "Density-functional thermochemistry. III. The role of exact exchange". *The Journal of Chemical Physics.* 98 (7), pp. 5648-5652.

[75] Lee C., Yang W. et Parr R., 1988. "Development of the Colle-Salvetti correlation-energy formula into a functional of the electron density". *Physical Review*. B.37 (2), pp. 785-789.

[76] Hohenberg P. et Kohn W., 1964. "Inhomogeneous Electron Gas". *Physical Review.* 136 (3B), pp. B864-B871.

[77] Kohn W., Becke A. D. et Parr R G., 1996. "Density Functional Theory of Electronic Structure ". *The Journal of Physical Chemistry*. 100 (31), pp. 12974–12980.

[78] Dugas H., 1996. "Principes de base en modélisation moléculaire, Aspects théoriques et pratiques, Chapitre 3 introduction aux méthodes de minimisation d'énergie". Quatrième édition, Librairie de l'Université de Montréal.

[79] Pople J A., Santry D P. et Segal G A., 1965. "Approximate self-consistent molecular orbital theory. i. invariant procedures". *Journal of Chemical Physics.* 43 (10), pp. S129-S135.

[80] Pople J A., Beveridge D L. et Dobosh P A., 1967. "Approximate self-consistent molecular-orbital theory. v. intermediate neglect of differential overlap". *Journal of Chemical Physics.* 47 (6), pp. 2026-2033.

[81) Bingham R C., Dewar M J S. et Lo D H., 1975. "Ground states of molecules. XXV. MINDO/3. Improved version of the MINDO semiempirical SCF-MO method". *Journal of the American Chemical Society,* 97 (6), pp. 1285–1293.

[82] Dewar M J S. et Thiel W., 1977. "Ground States of Molecules, 38. The MNDO Method. Approximations and Parameters". *Journal of American Chemical Society*, 99 (15), pp. 4899-4907.

[83] Dewar M J S., Zoebisch E G., Healy E F. et Stewart J J P., 1985. "Development and use of quantum mechanical molecular models. 76. AM1: A new general purpose quantum mechanical molecular model". *Journal of the American Chemical Society*. 107 (13), pp. 3902-3909.

[84] Stewart J J P., 1989. "Optimization of parameters for semiempirical methods I. Method". *Journal of Computational Chemistry.* 10 (2), pp. 209 220.

[85] Dewar M J S., Jie C. et Yu J., 1993. "SAM1; The first of a new series of general purpose quantum mechanical molecular models". *Tetrahedron.* 49 (23), pp. 5003-5038.

[86] Stewart J J P., 2007. "Optimization of parameters for semiempirical methods V: Modification of NDDO approximations and application to 70 elements". *Journal of Molecular Modeling*. 13 (12), pp. 1173-1213.

[87] Voityuk A A. et Rösch N., (2000). "AM1/d Parameters for Molybdenum". *The Journal of Physical Chemistry A*. 104 (17), pp. 4089-4094.

[88] Řezáč J., Fanfrlík J., Salahub D. et Hobza P., 2009. "Semiempirical Quantum Chemical PM6 Method Augmented by Dispersion and H-Bonding Correction Terms Reliably Describes Various Types of Noncovalent Complexes". *Journal of Chemical Theory and Computation*. 5 (7), pp. 1749-1760.

[89] Andrews D H., 1930. "The Relation Between the Raman Spectra and the Structure of Organic Molecules". *Physical Review.* 36 (3), pp. 544-554.

[90] Hetényi C., Maran U. et Karelson M., 2003. "A comprehensive docking study on the selectivity of binding of aromatic compounds to proteins". *Journal of chemical information and computer sciences*. 43 (5), pp. 1576-1583.

[91] Keseru G M. et Menyhárd D K., 1999. "Role of proximal His93 in nitric oxide binding to metmyoglobin. Application of continuum solvation in Monte Carlo protein simulations". *Biochemistry.* 38 (20), pp. 6614-6622.

[92] Holm R H., Kennepohl P. et Solomon E I., 1996. "Structural and Functional Aspects of Metal Sites in Biology". *Chemical Reviews.* 96 (7), pp. 2239-2314.

[93] Bremner I., 1998. "Manifestations of copper excess". *The American journal of clinical nutrition.* 67 (5), pp. 1069S-1073S.

[94] Magali T., 2004. Thèse de Doctorat, Université de Montpellier 1. France. "Analyse du transcriptome rénal murin dans des conditions d'exposition aiguë et chronique à l'uranium".

[95] Gold V. et Bethell D., 1976. "Advances in Physical Organic Chemistry". Volume 12. Academic Press, pp. 318.

[96] Young D C., 2001. "Computational chemistry: A Practical Guide for Applying Techniques to Real World Problems". John Wiley & Sons, Inc. New York. pp. 408.

[97] Burkert U. et Allinger N L., 1982. "Molecular Mechanics". ACS Monograph Series (Book 177). American Chemical Society, Washington DC. pp. 340.

[98] Allinger N L., 1977. "Conformational analysis. 130. MM2. A hydrocarbon force field utilizing V1 and V2 torsional terms". *Journal of the American Chemical Society.* 99 (25), pp. 8127-8134.

[99] Allinger N L., Yuh Y H. et Lii J H., 1989. "Molecular Mechanics. The MM3 Force Field for Hydrocarbons". *Journal of the American Chemical Society*. 111 (23), pp. 8551-8566.

[100] Allinger N L., Chen K. et Lii J H., 1996. "An improved force field (MM4) for saturated hydrocarbons". *Journal of Computational Chemistry.* 17 (5-6), pp. 642-668.

[101] Jones G B. et Chapman B J., 1995. "π Stacking Effects in Asymmetric Synthesis". *Synthesis*, 1995 (5), pp. 475–497.

[102] Hocquet A. et Langgård M., 1998. "An evaluation of the MM+ force field". *Molecular modeling annual.* 4 (3), pp. 94-112.

[103] Weiner S J., Kollman P A., Nguyen D T. et Case D A., 1986. "An all atom force field for simulations of proteins and nucleic acids". *Journal of Computational Chemistry.* 7 (2), pp. 230-252.

[104] Weiner P K. et Kollman P A., 1981. "AMBER: Assisted model building with energy refinement. A general program for modeling molecules and their interactions". *Journal of Computational Chemistry.* 2 (3), pp. 287-303.

[105] Brooks B R., Bruccoleri R E., Olafson B D., States D J., Swaminathan S. et Karplus M., 1983. "CHARMM: A program for macromolecular energy, minimization, and dynamics calculations". *Journal of Computational Chemistry*. 4 (2), pp. 187-.217.

[106] Errahoui B K., 2015. Thèse de Doctorat, Université Abou Bekr-Belkaïd de Tlemcen. "Etude des relations quantitatives structure–toxicité des composés chimiques à l'aide des descripteurs moléculaires « Modélisation QSAR » ".

[107] Margossian N., 2007. "Le règlement REACH-La règlementation européenne sur les produits chimiques". Dunod / L'usine Nouvelle, Paris.

[108] Hansch C. et Fujita T., 1964. "*p*-σ-π Analysis. A method for the correlation of biological activity and chemical structure", *Journal of the American Chemical Society*. 86 (8), pp. 1616- 1626.

[109] Hansch C. et Lien E J., 1971. "Structure- activity relationships in antifungal agents. A survey", *Journal of Medicinal Chemistry*. 14 (8), pp. 653- 670.

[110] Tham S Y. et Agatonovic- Kustrin S., 2002. "Application of the artificial neural network in quantitative structure- gradient elution retention relationship of phenylthiocarbamyl amino acids derivaties", *Journal of Pharmaceutical and Biomedical Analysis*. 28 (3-4), pp. 581-590.

[111] Ghasemi J., Saaidpour S. et Brown S D., 2007. "QSPR study for estimation of acidity constants of some aromatic acids derivatives using multiple linear regression (MLR) analysis". *Journal of Molecular Structure: THEOCHEM*. 805 (1-3), pp. 27-32.

[112] Geladi P. et Kowalski B R., 1986. "Partial least- squares regression: a tutorial". *Analytica Chimica Acta*. 185, pp. 1-17.

[113] Myles A J., Feudale R N., Liu Y., Woody N A. et Brown S D., 2004. "An introduction to decision tree modeling". *Journal of Chemometrics*. 18 (6), pp. 275-285.

[114] Duprat A F., Huynh T. et Dreyfus G., 1998. "Toward a principled methodology for neural network design and performance evaluation in QSAR. Application to the prediction of logP". *Journal of Chemical Information and Computer Sciences*. 38, pp. 586-594.

[115] Tetko I V., Villa A E P. et Livingstone D J., 1996. "Neural network studies. 2. Variable selection". *Journal of Chemical Information and Computer Sciences*. 36, pp. 794-803.

[116] Gasteiger J. et Zupan J., 1993. "Neural Networks in Chemistry". *Angewandte Chemie International Edition in English*. 32 (4), pp. 503-527.

[117] Leardi R., 2001. "Genetic algorithms in chemometrics and chemistry: a review". *Journal of Chemometrics*. 15 (7), pp. 559-569.

[118] Kennard R W. et Stone L A., 1969. "Computer Aided Design of Experiments", *technometrics*. 11 (1), pp 137-148.

[119] Karelson M., 2000. "Molecular Descriptors in QSAR/QSPR". Wiley, New York. 448 pages.

[120] Todeschini R. et Consonni V., 2000. "Handbook of Molecular Descriptors". Wiley, Weinheim.

[121] Dudek A Z., Arodz T. et Gàlvez J., 2006. "Computational methods in developing quantitative structure-activity relationships (qsar): A review". *Combinatorial Chemistry & High Throughput Screening*. 9 (3), pp. 213-228.

[122] Kier L B. et Hall L H., 1976. "Molecular connectivity in chemistry and drug research". New York: Academic Press. 257 pages.

[123] Randić M., 1975. "On characterization of molecular branching". *Journal of the American Chemical Society*. 97 (23), pp. 6609-6614.

[124] Leszczynski J. Shukla M., 2009. "Practical Aspects of Computational Chemistry: Methods, Concepts and Applications". Springer.

[125] Touhami I., 2017. Thèse de Doctorat, Université Badji Mokhtar Annaba. "Calcul des équations de prédiction : des indices de rétention en chromatographie gazeuse [HAP comportant un hétéroatome (O, S et N)], pyrazine, et des propriétés physicochimiques pour une série de pesticides".

[126] Bouveresse D J R., Maalouly J. et Jaillais B., 2004. "Sélection d'échantillons représentatifs par des méthodes chimiométriques : Application à des modèles d'étalonnage". *Spectra Analyse*. 33 (237), pp. 23-27.

[127] Snee R D., 1977. "Validation of Regression Models: Methods and Examples". *Technometrics*. 19 (4), pp. 415-428.

[128] Despagne F. et Massart D L., 1998. "Neural networks in multi variate calibrage". *Analyst*. 123 (11), pp. 157-178.

[129] Sprevak D., Azuaje F. et Wang H., 2004. "A non-random data sampling method for classification model assessment". Proceedings of the 17th International Conference on Pattern Recognition. 3, pp. 406-409.

[130] Ren Y Y., Liu H X., Yao X J. et Liu M C., 2007. "Prediction of ozone tropospheric degradation rate constants by projection pursuit regression". *Analytica Chimica Acta*. 589 (1), pp, 150-158.

[131] Graybill F A., 1976. "Theory and Application of the Linear Model". Duxbury, North Scituate, Mass, pp. 231-236.

[132] Chang J., Lei B., Li J., Li S., Shen Y. et Yao X., 2008. "Accurate and Validated Quantitative Structure-Activity Relationship Model of Caspase-mediated Apoptosis-inducing Activity of Phenolic Compounds Using Density Functional Theory Calculation and Genetic Algorithm-Multiple Linear Regression". *QSAR & Combinatorial Science*, 27 (11-12), pp. 1318-1325.

[133] Knuth D E., 1997. "The art of computer programming", Vol, 2 (3rd ed): Seminumerical Algorithms. Addison-Wesley Longman Publishing Co., Inc. Boston, MA, USA.

[134] Tourassi G A, Frederick E D, Markey M K, Floyd Jr C E, (2001). "Application of the mutual information criterion for feature selection in computer-aided diagnosis". *Medical Physics*, 28(12), pp. 2394-2402.

[135] Todeschini R., Ballabio D., Consonni V., Mauri A. et Pavan M., 2004. MOBY DIGS Software for Multilinear Regression Analysis and Variable Subset Selection by Genetic Algorithm, Release 1.0 for Windows, Milano.

[136] Todeschini R., Consonni V., Pavan M., 2005. DRAGON, Software for the Calculation of Molecular Descriptors, Release 5.3 for windows, Milano.

[137] Kowalski B., Gerlach R. et Wold H., 1982. "Chemical Systems under Indirect Observation, in K. Joreskog and H. Wold (Eds.), Systems under Indirect Observation". North-Holland, Amsterdam, pp. 191-209.

[138] Erikson I., Johansson E., Kettaneh- Wold N. et Wold S., 2001. "Multi- and megavariate data analysis. principles and applications". Umetrics Academy, Umeå. 533 pages.

[139] Wold S., Ruhe A., Wold H. et Dunn W J., 1984. "The Collinearity Problem in Linear Regression. The Partial Least Squares (PLS) Approach to Generalized Inverses". *SIAM Journal on Scientific and Statistical Computing*, 5 (3), pp. 735-743.

[140] Wold S., Albano C., Dunn W J., Edlund U., Esbensen K., Geladi P., Hellberg S., Johansson E., Lindberg W. et Sjöström M. 1984. "Multivariate Data Analysis in Chemistry. In: Kowalski B.R. (eds) Chemometrics. NATO ASI Series (Series C: Mathematical and Physical Sciences)". 138. Springer, Dordrecht, pp. 17-95.

[141] Höskuldsson A., 1988. "PLS regression methods". *Journal of Chemometrics*. 2 (3), pp. 211-228.

[142] Burns J A. et Whitesides G M., 1993. "Feed-forward neural networks in chemistry: mathematical systems for classification and pattern recognition". *Chemical Reviews*. 93 (8), pp. 2583-2601.

[143] Anker L S. et Jurs P C., 1992. "Prediction of carbon-13 nuclear magnetic resonance chemical shifts by artificial neural networks". *Analytical Chemistry*. 64 (10), pp. 1157-1164.

[144] Aoyama T., Suzuki Y. et Ichikawa H., 1990. "Neural networks applied to quantitative structure-activity relationship analysis". *Journal of Medicinal Chemistry*. 33 (9), pp. 2583-2590.

[145] Andrea T A., 1991. "Applications of neural networks in quantitative structure-activity relationships of dihydrofolate reductase inhibitors". *Journal of Medicinal Chemistry*. 34 (9). pp, 2824-2836.

[146] Jurs P C., 1996. "Computer Software Applications in Chemistry", Second Edition. John Wiley & Sons, Inc.: New York. 291 pages.

[147] Chouquet C., 2010. "Modèles Linéaires". Laboratoire de Statistique et Probabilités-Université Paul Sabatier-Toulouse.

[148] Lejeune M., 2004. "Statistiques : la théorie et ses applications", 2ème Edition. Springer-Verlag, Paris.

[149] Fayet G., 2010. Thèse de Doctorat, Université Pierre et Marie Curie. "Développement de modèles QSPR pour la prédiction des propriétés d'explosibilité des composés nitroaromatiques".

[150] Jodouin J F., 1994. "Les réseaux de neurones, principes et définitions". Hermès, Paris.

[151] Davalo E. et Naim P., 1998. "Les réseaux de neurones". Edition Eyrolles.

[152] Souahi F., Hachemaoui A. et Chitour C E., 2007. "Caractérisation des mélanges complexes par une méthode utilisant les réseaux de neurones artificiels". *Journal de la Société Algérienne de Chimie*. 17 (1), pp. 19-26.

[153] Vladimir N V., 1995. "The nature of statistical learning theory". Springer-Verlag Berlin, Heidelberg.

[154] Smola A J. et Schölkop B., 2004. "A tutorial on support vector regression". *Statistics and Computing*. 14 (3), pp, 199-222.

[155] Cristianini N. et Shawe-Taylor J., 2000. "An Introduction to Support Vector Machines and Other Kernel-based Learning Methods". Cambridge: Cambridge University Press.

[156] Christopher M B., 1995. "Neural Networks for Pattern Recognition". Oxford University Press, Inc. New York, NY, USA.

[157] Lourici L., 2004. Thèse de Doctorat. Université Badji Mokhtar Annaba. "Contribution à l'étude des indices de rétention en chromatographie gazeuse".

[158] Rehailia M E H, 1995. "Modèles linéaires statistiques : modèles à effets fixes". Alger : Office des publications universitaires. 461 pages.

[159] Commissariat à l'Energie Atomique., 1969. Méthodes Statistiques en Chimie Analytique. Etalonnage. Volume IV, estimation de l'étalonnage d'une méthode d'analyse, CETAMA, Dunod, Paris.

[160] Commissariat à l'Energie Atomique., 1969. Méthodes Statistiques en Chimie Analytique. Volume I, généralités sur le calcul statistique. Fascicule 2, tables statistiques, CETAMA, Dunod, Paris.

[161] Draper N R. et Smith H., 1998. "Applied Regression Analysis". Third Edition. Wiley Series in Probability and Statistics, New York.

[162] Gramatica P., 2007. "Principles of QSAR models validation: internal and external". *Qsar & Combinatorial Science*. 26 (5), pp. 694–701.

[163] Eriksson L., Jaworska J., Worth A P., Cronin M T., McDowell R M. et Gramatica P., 2003. "Methods for reliability and uncertainty assessment and for applicability evaluations of classification- and regression-based QSARs". *Environmental health perspectives*. 111 (10), pp. 1361–1375.

[164] Golbraikh A. et Tropsha A., 2002. "Beware of q2!", *Journal of Molecular Graphics and Modelling*. 20 (4), pp. 269-276.

[165] Dearden J C., Cronin M T D. et Kaiser K L E., 2009. "How not to develop a quantitative structure-activity or structure-property relationship (QSAR/QSPR)". *SAR and QSAR in Environmental Research*, 20 (3-4), pp. 241-266.

[166] Lozano S., Halm-Lemeille M P., Lepailleur A., Rault S. et Bureau R., 2010. "Consensus QSAR Related to Global or MOA Models: Application to Acute Toxicity for Fish". *Molecular Informatics*. 29 (11), pp. 803–813.

[167] Roy P P., Kovarich S. et Gramatica P., 2011. "QSAR model reproducibility and applicability: A case study of rate constants of hydroxyl radical reaction models applied to polybrominated diphenyl ethers and (benzo-)triazoles". *Journal of Computational Chemistry*. 32 (11), pp. 2386-2396.

[168] Rumelhart D E, Mcclelland J, L, et PDP Research Group, (1988) Parallel Distributed processing, Vol, 1, Massachusetts: MIT press, pp, 547.

[169] Hopfield J J, (1982) "Neural networks and physical systems with emergent collective computational abilities". *Proceedings of the National Academy of Sciences of the United States of America*. 79 (8), pp. 2554-2558.

[170] Tropsha A., Gramatica P. et Gombar V K., 2003. "The Importance of Being Earnest: Validation is the Absolute Essential for Successful Application and Interpretation of QSPR Models". *Qsar & Combinatorial Science*. 22 (1), pp. 69-77.

[171] Shen M., Béguin C., Golbraikh A., Stables J P., Kohn H. et Tropsha A., 2004. " Application of Predictive QSAR Models to Database Mining: Identification and Experimental Validation of Novel Anticonvulsant Compounds", *Journal of Medicinal Chemistry*. 47 (9), pp. 2356-2364.

[172] Weisberg S., 2005. "Applied Linear Regression". 3rd ed, John Wiley and sons, Inc, New Jersey.

[173] ChemDraw Ultra chemical structure drawing standard. Version 7. Copyright Cambridge Soft Corporation. 2002.

[174] HyperChem™ Release 6.03 for windows., 2000. Molecular Modeling System.

[175] Talete Srl, DRAGON (Software for Molecular Descriptor Calculation) Version 6.0 – 2011 - http://www.talete.mi.it/.

[176] Minitab, Release 13,31, Statistical software, (2000).

[177] Matlab, Version 7,0,0,19920 (Release 4), The Language of Technical Computing, The Math Works, Inc, May 06 (2004).

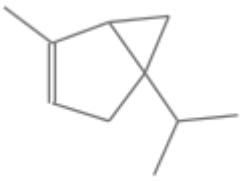
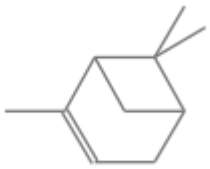
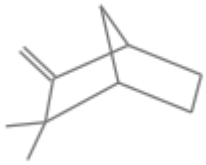
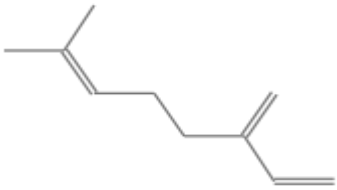
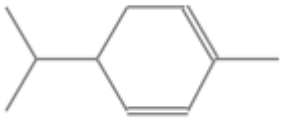
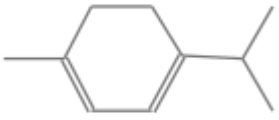
[178] Molegro, Data Modeller (MDM), v,2,1,0, Copyright Molegro (2009).

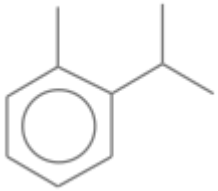
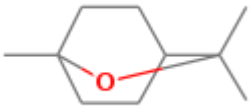
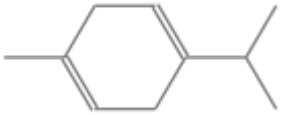
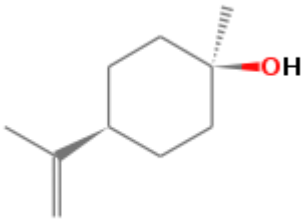
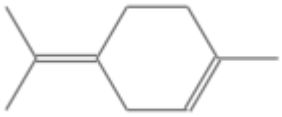
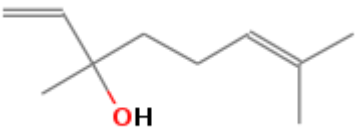
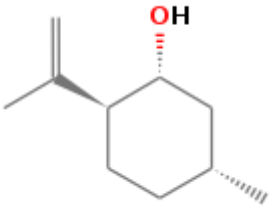
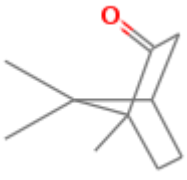
[179] Nezhadali A., Nabavi M., Rajabian M., Akbarpour M., Pourali P. et Amini F., 20 14. "Chemical variation of leaf essential oil at different stages of plant growth and in vitro antibacterial activity of *Thymus vulgaris Lamiaceae*, from Iran". *beni-suef university journal of basic and applied sciences*. 3, pp. 87-92.

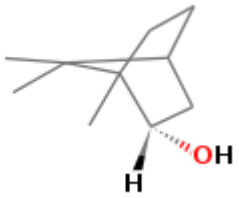
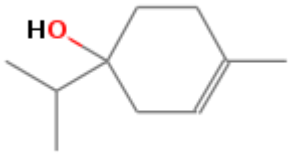
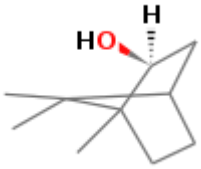
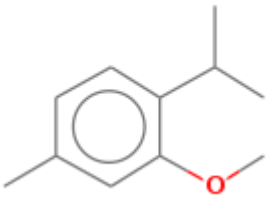
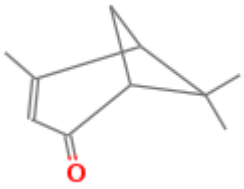
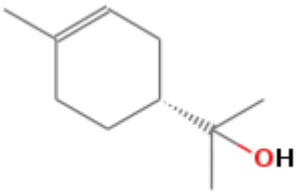
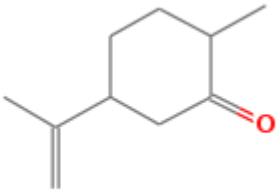
[180] Ramsey F L. et Schafer D W., 2002. "The statistical sleuth a course in methods of data analysis". Duxbury Press. Belmont, CA.

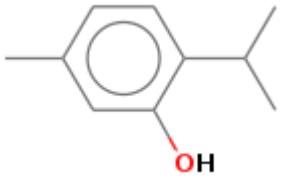
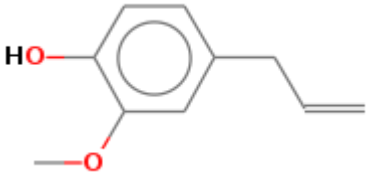
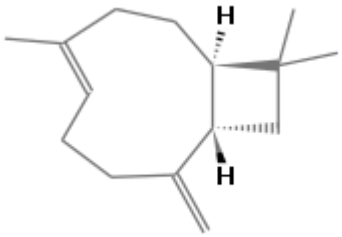
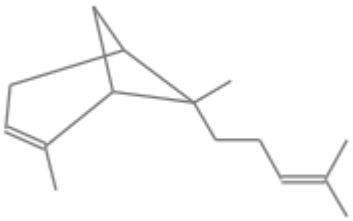
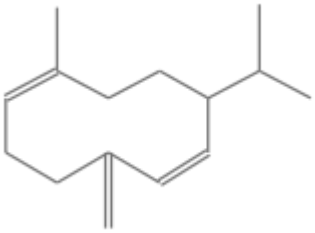
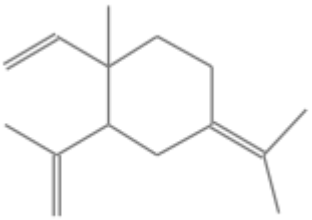
Annexe

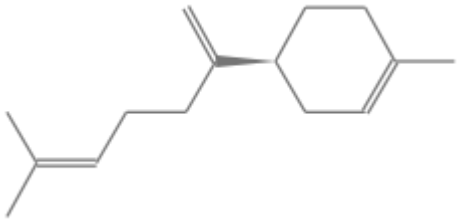
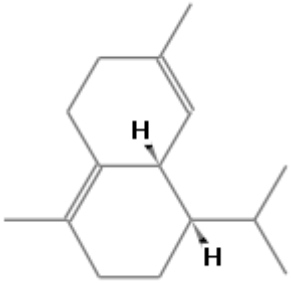
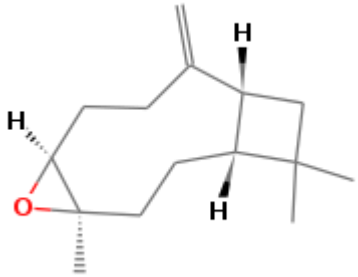
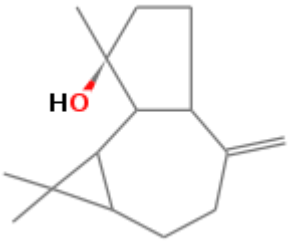
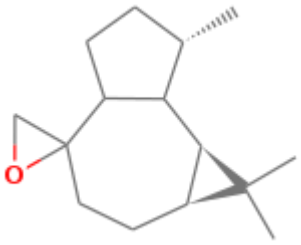
Tableau I : Liste des composés étudiés

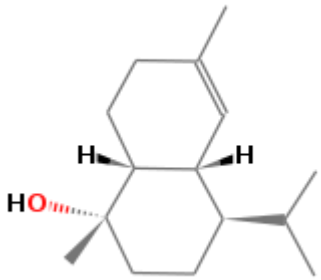
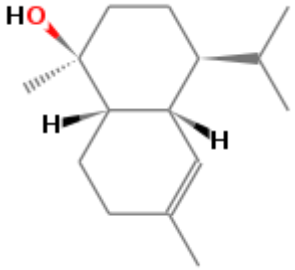
N°	Name	CAS N	Structure
1	alpha-Thujene	2867-05-2	
2	alpha-Pinene	80-56-8	
3	Camphene	79-92-5	
4	beta-Pinene	127-91-3	
5	beta-Myrcene	123-35-3	
6	alpha-Phellandrene	99-83-2	
7	alpha-Terpinene	99-86-5	

N°	Name	CAS N	Structure
8	o-Cymene	527-84-4	
9	1,8-cineole	470-82-6	
10	gamma-Terpinene	99-85-4	
11	cis-beta-Terpineol	7299-41-4	
12	Terpinolene	586-62-9	
13	Linalool	78-70-6	
14	Isopulegol	89-79-2	
15	Camphor	76-22-2	

N°	Name	CAS N	Structure
16	Borneol	507-70-0	
17	4-Terpineol	562-74-3	
18	Isoborneol	124-76-5	
19	Thymol methyl ether	1076-56-8	
20	Verbenone	80-57-9	
21	alpha-Terpineol	98-55-5	
22	Dihydrocarvone	7764-50-3	

N°	Name	CAS N	Structure
23	Thymol	89-83-8	
24	Eugenol	97-53-0	
25	beta-Caryophyllene	87-44-5	
26	alpha-Bergamotene	17699-05-7	
27	Germacrene D	23986-74-5	
28	gamma-Elemene	29873-99-2	

N°	Name	CAS N	Structure
29	beta-Bisabolene	495-61-4	
30	Delta-cadinene	483-76-1	
31	Caryophyllene oxide	1139-30-6	
32	Spathulenol	6750-60-3	
33	Aromadendrene oxide	85760-81-2	

N°	Name	CAS N	Structure
34	Cadinol	19435-97-3	
35	Muurolol	19912-62-0	
36	Bisabolol	515-69-5