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

MODÈLES QSPR DE DIFFERENTES CARACTERISTIQUES DE POLLUANTS ENVIRONNEMENTAUX

Présenté par : Madame Mounia ZINE

Devant le jury composé de :

Présidente : SAIFI Hayette	(MCA)	Université d'Annaba
Directeur de thèse : MESSADI Djelloul	(Prof)	Université d'Annaba
Examinateur : DJALLAL Ahmed	(Prof)	Université d'Annaba
Examinatrice : HEZIL Naouel	(MCA)	Université de Khenchela
Examinateur : ZENATI Noureddine	(MCA)	Université de Souk-Ahras

Dédicaces

Merci ; Dieu de m'avoir

Donné la capacité d'écrire et de réfléchir, la force d'y croire, la patience d'aller jusqu'au bout du rêve, le bonheur de lever mes mains vers le ciel et de dire « YARAB ».

Je dédie cet humble travail à

Mes parents et mes frères

Toute ma famille

Mes amis

Et toute l'équipe du laboratoire « LASEA ».

Remerciements

Ce mémoire n'aurait pas vu le jour sans la confiance, la patience et la générosité du responsable de thèse, Monsieur le Professeur ***MESSADI Djelloul*** que je remercie vivement. Je voudrais aussi le remercier pour le temps et la patience qu'il m'a accordés tout au long de ces années.

Je tiens également à remercier Madame ***SAIFI Hayette***, Maître de conférences à l'université d'Annaba, qui m'a fait l'honneur d'accepter de présider le jury de cette thèse.

J'adresse l'expression de mes sincères remerciements à Monsieur ***AHMED Djellal,*** Professeur à l'université d'Annaba, Mes vifs remerciements vont également à Madame ***HEZIL Naouel*** Maître de conférences à l'université de Khenchela, Je remercie Monsieur ***ZENATI Noureddine*** Maître de conférences à l'université de Souk Ahras, pour l'honneur qu'ils nous font en acceptant d'examiner notre travail.

Enfin, je ne saurais oublier mes collègues de laboratoire et également tous ceux qui par leur présence ou par leur aide m'ont permis de mener à bien ce travail.

Je tiens à remercier également tous ceux qui ont contribué de près ou de loin à la réalisation de ce travail.

Abstract

The work presented aims to develop and establish reliable and predictive QSRR/QSPR models for three properties: "relative retention time", "n-octanol / water partition coefficient" and "vapor pressure" for certain volatile organic compounds using a reduced number of relevant molecular descriptors and respecting the entire protocol of the QSAR methodology.

The work was conducted using statistical approaches, namely multiple linear regressions (MLR), Artificial Neural Networks (ANN) and Supports Vector Machines (SVM). Calculations were made with different programs using the PM3 method.

Key words: Relative Retention Time -n-octanol / Water Partition Coefficient - Vapor pressure -QSPR/ QSRR- VOCs-Molecular descriptors.

الملخص

يهدف العمل المقدم في هذه الأطروحة إلى تطوير وإنشاء نماذج موثوقة وقابلة للتنبؤ من QSAR لثلاثة خصائص: "وقت الاحتفاظ النسبي" ، "معامل فصل الماء بالأوكتانول / ن" و "ضغط البخار" لبعض المركبات العضوية المتطايرة باستخدام عدد أقل من الواصفات الجزيئية ذات الصلة والتمسك بالبروتوكول الكامل لمنهجية QSPR/QSRR

تم إجراء العمل باستخدام الأساليب الإحصائية ، في هذه الحالة الانحدار الخطي المتعدد (MLR) ، الشبكات العصبية الاصطناعية (ANN) وآلات المتجهات الحاملة (SVM). تم إجراء العمليات الحسابية باستخدام برامج مختلفة باستخدام طريقة PM3

الكلمات المفتاحية : وقت الاحتفاظ النسبي - معامل فصل الماء بالأوكتانول / ن - ضغط البخار - QSPR - الواصفات الجزيئية .

Résumé

Le travail présenté dans cette thèse a pour objectifs de développer et d'établir des modèles QSAR fiables et prédictifs pour trois propriétés : « le temps de rétention relatif », « le coefficient de partage n-octanol/eau » et « la pression de vapeur » pour certains composés organiques volatils (COVs) en utilisant un nombre réduit de descripteurs moléculaires pertinents et en respectant tout le protocole de la méthodologie QSRR/QSPR.

Le travail a été mené à l'aide des approches statistiques, en l'occurrence la régression linéaire multiple (RLM), les réseaux de neurones artificiels (RNA) et les machines à vecteurs supports (SVM). Les calculs ont été effectués avec différents programmes en utilisant la méthode PM3.

Mots clés : *Temps de rétention relatif -Coefficient de partage n-octanol/eau-Pression de vapeur -QSPR/QSRR- COVs-Descripteurs moléculaires.*

Liste des illustrations

Sommaire

Sommaire

Chapitre II : Outils et techniques utilisés

Sommaire

Symboles et Abréviations

AG	Algorithme génétique
BLUE	Best linear unbiased estimators.(Meilleurs estimateurs linéaires sans biais)
COV	Composé organique volatil
COTV	Composé organique très volatil
CPG	Chromatographie en phase gazeuse
D	Descripteur
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
FIV	Facteur d'inflation de la variance
LMO	Cross-validation by leave-many-out: Validation croisée par omission d'un ensemble d'observations
LOO	Cross-validation by leave-one-out: Validation croisée par omission d'une observation
OMS	Organisation Mondiale de la Santé
PCOP	Potentiel de Création d'Ozone Photochimique
PM3	Parametrization Method 3
PRESS	Somme des carrés des erreurs de prédiction
Pv	Pression de vapeur
Q^2_{BOOT}	Coefficient de prédiction par la technique du bootstrap
Q^2_{LOO}	Coefficient de prédiction par leave one out
QSAR	Quantitative Structure/ Activity Relationship (Relation quantitative structure / activité)
R^2	Coefficient de détermination
REACH	Un règlement européen (règlement n°1907/2006) entré en vigueur en 2007 pour sécuriser la fabrication et l'utilisation des substances chimiques dans l'industrie européenne.
RLM	Régression linéaire multiple

RMSE	Racine de l'écart quadratique moyen (Root Mean Squared Error)
RNA	Réseaux de neurones artificiels
SCAQD	South Coast Air Quality Management District (District de gestion de la qualité de l'air de la côte sud)
SCE	Somme des carrés des écarts
SCT	Somme des carrés totale
SVM	Support vector machine (machine à vecteur support)
d_R	distance de rétention
$e_{i\ std}$	Résidu standardisé
h_{ii}	Eléments diagonaux de la matrice chapeau
n	Dimension de la population (échantillon)
n-p	Nombre de degrés de liberté
r_i	Résidu studentisé interne
s	Erreur standard
t	t de Student
t'_r	Temps de rétention réduit
t_i	Résidu studentisé externe
t_m	Temps mort (temps de rétention nulle)
t_r	Temps relatif
y_i	Valeur observée
$\hat{y}_{(i)}$	Valeur prédite
$\hat{y}_i$	Valeur estimée

Liste des Tableaux

Partie.1 : Etude bibliographique

Partie.2 Application/Résultats et discutions

Liste des figures

Partie.1 : Etude bibliographique

Partie.2 : Application/Résultats et discutions

Liste des figures

INTRODUCTION GÉNÉRALE

INTRODUCTION GENERALE

La pollution est une réalité très présente dans notre environnement, les polluants environnementaux de tous genres contaminent l'eau, l'air et la terre mettant en péril les humains et les écosystèmes. De plus, ils sont souvent sources de conflit entre populations et industries. En adoptant une approche éco-systémique globale pour examiner les intérêts divergents et leur conséquence, les approches éco-santé s'efforcent de protéger la santé tout en assurant l'équilibre des besoins des divers intervenants et la préservation de l'écosystème. Partout dans le monde, des milliers de produits chimiques constituent un risque pour les populations et les écosystèmes.

Les composés organiques volatils (COVs) sont des gaz organiques qui s'évaporent plus ou moins rapidement à température ambiante et se retrouvent dans l'air. On en compte plus de 300 types. Les sources sont naturelles (forêts, prairies) ou anthropiques (transports, industrie, etc.). Ces COVs ont une action directe et indirecte dans l'atmosphère, ce qui conduit à des impacts du point de vue environnemental :

- d'une part, ils génèrent les plus importants des polluants dont résulte le smog.

- d'autre part, ils ont un impact sur le climat ainsi que la nature, de même qu'ils affectent l'espèce humaine et animale : certains de ces composés organiques volatils s'avèrent cancérogènes pour l'homme [1].

En Algérie où l'urbanisation et la motorisation se développent rapidement, la dégradation de la qualité de l'air et les nuisances sont déjà perçues. Des études ont montré que le Grand Alger (3,5 millions d'habitants, 800.000 véhicules) est, comme toute grande agglomération urbaine, confronté à une intense pollution atmosphérique [2].

Les relations quantitatives structure-activité/propriété (QSAR/QSPR) sont de plus en plus utilisées, du fait de la croissance des moyens de calculs. Très récemment, la mise en place du nouveau règlement européen (REACH), qui recommande leur utilisation pour limiter le recours à l'expérience, donne un nouvel essor au développement de tels modèles prédictifs. Dans les dernières années, l'utilisation des méthodes QSAR n'a cessé de progresser. Elle est même devenue indispensable en chimie pharmaceutique et pour la conception de médicaments. Les relations entre les structures des molécules et leurs activités ou propriétés sont généralement établies à l'aide des méthodes de modélisation moléculaire et des méthodes statistiques. Les techniques usuelles reposent sur la caractérisation des molécules par un ensemble de descripteurs ; nombres réels mesurés ou calculés à partir des structures moléculaires. Il est alors possible d'établir une relation entre ces descripteurs et la grandeur modélisée [3].

Nous projetons dans ce travail d'appliquer les méthodologies QSRR/QSPR pour la modélisation du temps de rétention relatif (t_{rr}) et quelques propriétés physico-chimiques telles que :

-Le coefficient de partage octanol/eau (K_{ow}) ; l'étude de la répartition est fondamentale puisque le risque dans un compartiment donné est lié à la concentration de la substance dans ce compartiment.

-La prédiction de la pression de vapeur d'un ensemble de 51 COVs.

Nous avons utilisé des approches QSRR/QSPR hybrides associant algorithme génétique pour la sélection de sous-ensembles de variables significatives parmi quelques 2000 calculées théoriquement, et soit une régression linéaire (RLM), soit une régression non linéaire (RNA, SVM).

Le manuscrit de cette thèse est articulé en plus de la bibliographie, d'une introduction et d'une conclusion générale, sur deux grandes parties :

Dans la partie Généralités, le premier chapitre a pour but de prendre connaissance des origines des COVs, leur définition et classification ainsi quelques propriétés et toxicités.

Le deuxième chapitre est consacré à une étude bibliographique sur les approches de base, les méthodologies, le développement, les techniques de validation et les applications des méthodes QSRR/QSPR. Une description des différents outils nécessaires à la mise en œuvre de ces méthodes sera ainsi détaillée (descripteurs, méthodes statistiques, principes de la validation des modèles...).

Dans la deuxième partie, nous présenterons et nous discuterons dans trois chapitres les résultats obtenus pour les études QSRR/QSPR que nous avons effectuées. La stratégie générale adoptée pour élaborer les modèles QSRR/QSPR dans ces études est celle qui utilise un petit nombre de descripteurs en se basant sur la nature des molécules de nos bases de données et les différents mécanismes possibles pour expliquer les propriétés étudiées et en respectant tous les critères concernant un modèle QSRR/QSPR : fiabilité, robustesse et prédictivité.

Partie. 1
Etude bibliographique

Les Composés Organiques Volatils (COV)

I-GENERALITES

I-1-Composé organique

Un composé organique est un composé chimique dont la molécule contient au moins un atome de carbone lié directement à un atome d'hydrogène. Cet hydrogène peut être substitué par d'autres atomes tels que l'oxygène, l'azote, le soufre, le phosphore, le silicium ou encore des halogènes (chlore, fluor, brome, iode...) à l'origine des différentes familles de composés organiques volatils.

I-2-Composé organique volatil (COV)

Un composé organique volatil peut être défini selon différentes caractéristiques physicochimiques. Ces critères peuvent varier en fonction du pays et/ou encore de la législation appliquée. En général, un COV est une substance chimique très volatile c'est-à-dire capable de s'évaporer à température ambiante et à la pression atmosphérique. Cependant, l'Europe restreint cette définition par une réglementation basée sur des propriétés physico-chimiques précises. En effet, la Directive Européenne IED (Industrial Emission Directive) a adopté, en 2010, la définition suivante d'un COV : un composé organique volatil est "un composé organique ayant une pression de vapeur de 0,01 kPa ou plus, à une température de 293,15 K (20°C)" [4].

La volatilité des COVs peut être définie en fonction de la température d'ébullition ou encore de la pression de vapeur saturante selon les conditions d'utilisation. La température d'ébullition, par exemple, correspond à la température du changement d'état du composé de la phase liquide à la phase gaz. Au delà de cette température, la phase liquide est complètement évaporée et le composé n'existe que sous sa forme gazeuse. Bien que la volatilité des liquides ne soit pas strictement proportionnelle à la température d'ébullition, il apparaît clairement qu'un liquide à point d'ébullition bas s'évapore plus rapidement que celui dont le point d'ébullition est plus élevé. Il est donc possible de déterminer la volatilité d'un composé à partir de sa température d'ébullition. Dès lors, un composé organique est d'autant plus volatil que sa température d'ébullition est basse. Implicitement, la mesure de la température d'ébullition renvoie à la tension de vapeur saturante du composé. Cette pression correspond à la pression de vapeur saturante du composé et caractérise sa volatilité. Il semblerait donc que le composé est d'autant plus volatil que sa pression de vapeur saturante est élevée. Une définition différente de la précédente, inscrite dans la directive européenne 2004/42/EC [5]

et la norme ISO 16000-6 [6], définit un COV en fonction de sa température d'ébullition. Il apparaît dans la directive européenne 2004/42/EC qu'un COV est "toute substance organique dont le point d'ébullition est inférieur à 250°C à une pression atmosphérique standard de 101,3 kPa". Celle utilisée par L'OMS (Organisation Mondiale de la Santé), dans la norme ISO 16000-6, stipule qu'un composé organique est considéré volatil lorsque sa température d'ébullition se situe entre (50°C-100°C) et (240 °C-260°C). Cela correspond à une pression de vapeur saturante supérieure à 100 kPa (25°C).

Aux Etats-Unis et en Allemagne, la règlementation est plus restrictive et la définition est décrite en termes de substances volatiles [7]. En effet, la règlementation allemande définit un COV comme substance organique volatile "dont le point d'ébullition est inférieur à 200°C à une pression atmosphérique standard de 101,3 kPa". La règlementation californienne du SCAQMD (South Coast Air Quality Management District), quant à elle, considère comme COV toute substance volatile dont la température d'ébullition est inférieure à 280°C. Cette définition exclut certains composés tels que l'acétone, l'éthane, l'acétate de méthyle..., dont la réactivité photochimique est faible.

Il existe également d'autres définitions des COVs desquelles sont exclus des composés organiques volatils à cause de l'origine de leur source d'émission et leur influence sur l'environnement ou encore de leur degré de volatilité. Parmi ces définitions, la littérature propose les composés organiques non méthaniques (COVNM), les composés organiques très volatils (COTV) ou encore les composés organiques semi volatils (COSV) [8,9]. Pour les COVNM, le méthane est exclu de la famille des COV car son influence sur l'environnement (effet de serre) est différente de celle des COVs conventionnels dont l'effet néfaste est essentiellement axé sur la pollution photochimique.

En général, les COVs ont des températures d'ébullition comprises dans les intervalles $50°C < T < 100°C$ et $240°C < T < 260°C$. Les COTV sont les composés organiques très volatils c'est-à-dire que leur température d'ébullition est comprise dans les intervalles 0°C et 50°C. Les COSV sont des composés qui possèdent une faible volatilité. Ils sont dits composés organiques semi-volatils car leur température d'ébullition se situe dans les intervalles:
$240°C < T < 260°C$ et $380°C < T < 400°C$.

I-3-Classification des composés organiques volatils :

Les composés organiques volatils sont classés ci-après (Tableau I.1) selon leur température d'ébullition:

Tableau I.1**:** Classification des COVs selon leur température d'ébullition [10].

Volatilité	**Température d'ébullition**
Très volatils	< [50 - 100 °C]
Volatils	[50 - 100 °C] à [240 - 260 °C]
Semi-volatils	[240 - 260 °C] à [380 - 400 °C]

Une liste exhaustive des différents COVs est difficilement réalisable étant donnée la multiplicité des composés chimiques. Il est donc préférable de classifier les COVs en familles chimiques. Ces diverses familles concernent [11]:

– Les hydrocarbures (alcanes, alcènes, alcynes, hydrocarbures aromatiques…) ;

– Les alcools (méthanol, éthanol, propanol…) ;

– Les aldéhydes (formaldéhyde, acétaldéhyde, benzaldéhyde…) ;

– Les cétones (acétone, butanone, cyclohexanone…) ;

– Les acides carboxyliques (acide formique, acide acétique, acide butanoïque…) ;

– Les esters (formiate de méthyle, acétate de méthyle, propanoate de méthyle) ;

– Les éthers (éther éthylique, éthylène glycol, propylène glycol ;

– Les dérivés chlorés (dichlorométhane…), nitrés (β-nitropropane…) et aminés (éthylamine…).

Les COVs sont des précurseurs photochimiques c'est-à-dire qu'ils réagissent sous l'effet d'un rayonnement solaire. Cette réaction chimique contribue à la formation de l'ozone troposphérique (à basse altitude) [8]. La réactivité des COVs est différente d'un composé à un autre. Par exemple, les hydrocarbures sont des composés moins réactifs que certains composés aromatiques tels que le benzène.

La contribution à la formation d'ozone troposphérique a été définie sous forme d'échelle relative correspondant au potentiel de création d'ozone photochimique (PCOP). En

effet, cette grandeur permet d'évaluer la participation d'un COV dans les réactions photochimiques responsables de l'augmentation de l'ozone dans l'atmosphère. Il s'exprime en kg d'équivalent éthylène dont le potentiel de réactivité PCOP est égal à 1. Cette référence est due au fait que les alcènes contribuent le plus à la formation de l'ozone. Dans le tableau I.2, sont répertoriés les PCOP de quelques COVs [12]. Dans ce tableau, les aldéhydes et les aromatiques tels que l'acétaldéhyde et le toluène présentent des PCOP très élevés. Les autres familles chimiques composées d'acétylène, acétone, acide acétique et formique ont des PCOP faibles. L'acide acétique dont le PCOP est de 0,097 contribue faiblement à la production d'ozone à basse altitude contrairement au formaldéhyde lequel possède un PCOP de 0,519. A partir des valeurs de PCOP, une autre classification a été réalisée sur les COVs en fonction de leur rôle dans la production d'ozone. Les COVs ont été classés, dans le tableau I.2, par famille chimique.

Tableau I.2: Potentiel de Création d'Ozone Photochimique (PCOP) de quelques COVs [11] (Source : INERIS – DRC- 07 – 85842 – 12011A).

Famille chimique	COV	PCOP
Acides organiques	Acide acétique	0,097
	Acide formique	0,032
Alcynes	Acétylène	0,085
Aromatiques	Benzène	0,218
	Toluène	0,637
Aldéhydes	Formaldèhyde	0,519
	Acétaldéhyde	0,641
Cétones	acétone	0,094

Le tableau I.3 présente les COVs en fonction du rôle peu important, assez important et très important dans la production d'ozone [11].

Les COVs peuvent également être classés en fonction de leur odeur. Il existe, en effet, des COVs inodores et d'autres qui possèdent une odeur plus ou moins caractéristique. Parmi les plus odorants, il y a les amines, les composés soufrés, les dérivés oxygénés (aldéhydes et cétones) et les composés aromatiques.

Tableau I.3: Classification des COVs selon leur rôle dans la production d'ozone troposphérique [11].

Rôle	Famille COV
Très important	Alcanes (méthane), Alcynes (acétylène), Aromatiques (benzène), Aldéhydes (benzaldéhyde), Cétones (acétone), Alcools (méthanol), Esters (acétate de méthyle), Hydrocarbures chlorés, (méthyl chloroforme…).
Assez important	Alcène, Aromatique, Alcanes > C6, COV naturels (Isoprène)
Peu important	Alcanes, Cétones, Alcools (éthanol), Esters

Le tableau I.4 répertorie quelques COVs classés en fonction de leur famille chimique ainsi que de leur caractère odorant [8,13].

Tableau I.4 : Classement de quelques COVs en fonction de leur caractère odorant et leur pression de vapeur saturante [8,13].

Famille chimique	Noms usuels	Pression de vapeur saturante (Pa à 20 C)	Apparence
Acides organiques	Acide acétique	$1{,}57.10^3$	Liquide incolore, odeur piquante
	Acide formique	$4{,}4.10^3$	Liquide incolore, odeur piquante
Alcynes	Acétylène	$4{,}5.10^6$	Gaz incolore, inodore
Aromatiques	Benzène	$9{,}97.10^3$	Liquide incolore, odeur aromatique
	Toluène	$3{,}8.10^3$	Liquide incolore, odeur caractéristique
Aldéhydes	Formaldéhyde	$4{,}4.10^5$	Gaz incolore, odeur piquante
	Acétaldéhyde	$1{,}01.10^5$	Liquide incolore, odeur fruitée
Cétones	Acétone	$4{,}0.10^5$	Liquide incolore, odeur suave

I-4-Quelques familles de composés organiques :

I-4-1-Les alcools :

Un alcool est un composé organique dans lequel un hydroxyde « -OH » est fixé sur un atome de carbone saturé (hybridé sp3). Ces composés sont abondants dans la nature notamment dans la structure des sucres. On les retrouve aussi dans l'industrie pétrochimique.

I-4-1-1-Classification des alcools

On distingue trois catégories d'alcools (primaire, secondaire et tertiaire). Elles sont définies en fonction du nombre d'atomes de carbone auquel est lié celui porteur du groupement hydroxyde [14].

I-4-1-2-Oxydation des alcools :

A-Oxydation des alcools secondaires :

La réaction d'oxydation de l'alcool conduit à la formation d'une cétone ou d'un acide carboxylique. La structure de la chaîne carbonée ne peut permettre l'obtention d'un acide carboxylique lors d'une oxydation ménagée. On déduit donc que l'oxydation d'un alcool secondaire conduit à la formation d'une cétone.

$$CH_3—CH_2—CO—CH_3 \,/\, CH_3—CH_2—CHOH—CH_3$$

La réaction qui a eu lieu met en jeu le couple $MnO4^-/Mn^{2+}$. On en déduit les deux demi réactions :

$$MnO_4^- + 8H^+ + 5e^- \rightarrow Mn^{2+} + 4H_2O$$

$$CH_3—CH_2—CHOH—CH_3 \rightarrow CH_3—CH_2—CO—CH_3 + 2H^+ + 2e^-$$

La réaction d'oxydation du butan-2-ol s'écrit alors :

$$5CH_3—CH_2—CHOH—CH_3 + 2MnO_4^- + 6H^+ \longrightarrow 5CH_3—CH_2—CO—CH_3 + 2Mn^{2+} + 8H_2O$$

B- *O*xydation des alcools primaires :

Pour l'alcool primaire, la structure de la chaîne carbonée élimine la possibilité d'avoir une cétone. Il s'agit donc d'un acide carboxylique qui s'est formé. Le carbone passe donc d'un degré d'oxydation -I à +I. En réalité la réaction passe par la formation d'un aldéhyde (C au degré 0). Cependant les ions permanganates sont introduits en très large excès et vont donc oxyder l'aldéhyde en acide carboxylique. Si on avait réalisé l'expérience avec l'oxydant dans des proportions proches de la stœchiométrie, la réaction aurait conduit à la formation d'aldéhyde en grande partie mais aussi d'acide.

Le couple oxydant/réducteur dont fait partie l'alcool est le suivant :

$$CH_3—CH_2—CH_2—CH_2—OH \,/\, CH_3—CH_2—CH_2—COOH$$

pour la demi-réaction :

$$CH_3—CH_2—CH_2—CH_2—OH+H_2O \rightarrow CH_3—CH_2—CH_2—COOH+4H^+ +4e^-$$

La réaction d'oxydation du butan-1-ol s'écrit:

$$5CH_3—CH_2—CH_2—CH_2—OH+4MnO_4^- +12H^+ \longrightarrow 5CH_3—CH_2—CH_2—COOH+4Mn^{2+}+11H_2O$$

C-Oxydation des alcools tertiaires :

Dans le cas des alcools tertiaires, il n'est pas constaté de réaction. En effet, dans les cas précédents, chaque oxydation s'est soldée par la perte d'un atome d'hydrogène pour le carbone en alpha du groupement hydroxyde. Ceci n'est pas possible pour un alcool tertiaire.

I-4-2-Les aldéhydes :

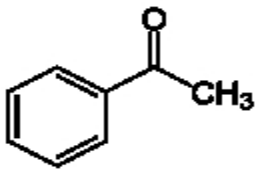
Un aldéhyde (contraction d'alcool déshydrogéné) est un composé carbonylé dont le carbone fonctionnel est lié à un hydrogène. Sa formule générale est : RCHO

I-4-3-Les cétones :

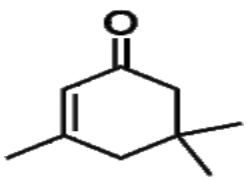
Une cétone est un composé carbonylé dont le carbone fonctionnel est lié à deux autres atomes de carbone. Sa formule générale est : $R_1CO\,R_2$ avec R_1 et $R_2 \neq H$

Les cétones sont synthétisées industriellement, le plus souvent par déshydrogénation d'alcools. Les cétones sont caractérisées par la présence, sur une chaine hydrocarbonée, d'un groupement carbonyle (-C=O). Pour faciliter la lecture des formules qui suivent, ce groupement sera noté (-CO-).

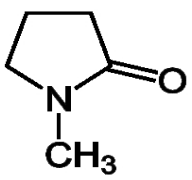
Elles peuvent être à chaine linéaire. Par exemple propanone ou acétone (CH_3-CO-CH_3), butanone ou méthyléthylcétone ou MEK (CH_3-CH_2-CO-CH_3), 4 méthyl-2-pentanone ou méthylisobutylcétone ou MIBK (CH_3-CO-CH_2-$(CH_3)_2$). Mais elles peuvent aussi être cycliques. Par exemple acétophénone ou phénylméthylcétone, isophorone, N-méthylpyrrolidone :



,



,



I-4-4-Les esters :

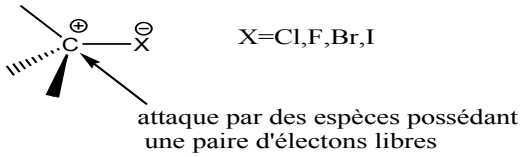
La formule générale d'un ester est : R_1COOR_2

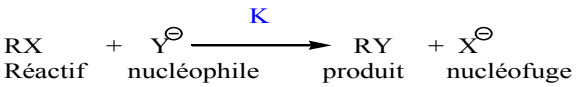
I-4-5-Dérivés halogénés :

Leur formule générique est RX où X est un halogène F, Cl, Br et I. Ils sont également appelés halogénures d'alkyles ou d'aryles selon la nature de R.

I-3-5-1-Réactivité des dérivés halogénés :

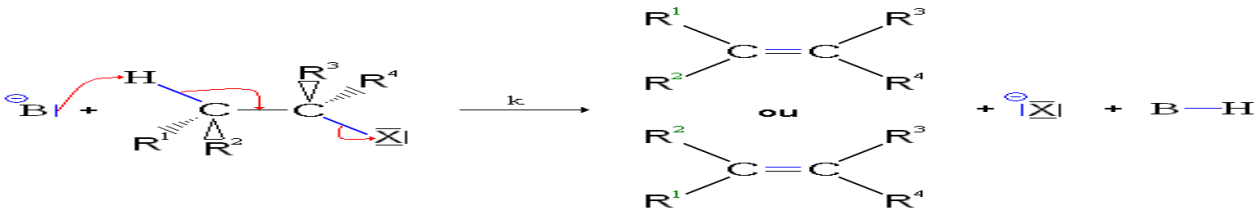
A-Réaction de substitution :





B-Réactions d'élimination :

ß-déshydrohalogénation



I-4-6-Les alcanes :

Les hydrocarbures non cycliques saturés, de formule générale C_nH_{2n+2}, sont des alcanes [15]. Du méthane au butane les alcanes se présentent à l'état gazeux dans les conditions normales de température et de pression. De C5 à C8, ils sont à l'état liquide, puis à l'état de cire et au delà de C13, ils sont sous forme solide. Le nom d'un alcane se forme à partir d'un préfixe, indiquant le nombre d'atomes de carbone, et d'une terminaison –ane.

I-4-7-Les alcènes :

On appelle alcènes, les hydrocarbures insaturés non cycliques de formule C_nH_{2n}. La chaîne principale est celle qui comporte le plus grand nombre de liaisons et, le cas échéant, le plus grand nombre d'atomes de carbone [16]

I-4-8-Les alcynes :

Un alcyne est un hydrocarbure dont la chaîne carbonée renferme une triple liaison.

La molécule est donc insaturée.

I-5-Les propriétés physico-chimiques des COVs :

I-5-1-Propriétés physico-chimiques des alcools :

La grande majorité des alcools utilisés industriellement sont liquides à température ambiante. Ils sont incolores et ont une odeur qui peut être agréable (éthanol), sucrée (cas des diols), âcre ou amère (propanol ou alcool furfurylique), ou encore piquante (alcool isoamylique). Les alcools communément utilisés sont miscibles dans l'eau, totalement pour les molécules les plus courtes (méthanol, éthanol...), partiellement pour les autres. Les alcools sont inflammables ou facilement inflammables. Le point d'éclair des plus utilisés se situe entre 12 et 40 °C. Leurs vapeurs peuvent former des mélanges explosifs avec l'air. Les diols, eux, ne sont pas considérés comme inflammables, leurs points d'éclair se situant à des températures supérieures à 100 °C.

Remarque : Le point d'éclair (Tec) est défini comme la température la plus basse, corrigée à 101,325 kPa, à laquelle l'application d'une source d'inflammation provoque l'inflammation des vapeurs dans les conditions spécifiques du test.

Ce paramètre fournit la connaissance nécessaire à la compréhension des processus physiques et chimiques fondamentaux de la combustion. De plus, il est important pratiquement pour les conditions de sécurité dans le stockage, le traitement et la manipulation, d'un composé donné. Et c'est l'une des principales caractéristiques d'inflammabilité utilisées pour évaluer les risques d'incendie et d'explosion des composés organiques.

Les alcools sont très volatils, leur diffusion dans le milieu ambiant ou dans l'atmosphère sera très importante. Ils dissolvent les graisses et certaines matières plastiques. Tous les alcools sont des liquides déshydratants.

- Les alcools sont associés par des liaisons hydrogène.
- Les alcools sont à la fois donneurs et accepteurs de liaisons hydrogène.

-points d'ébullition élevés [16].

I-5-2-Propriétés physico-chimiques des cétones :

À température ambiante, les cétones sont des liquides incolores. Leur odeur suave et sucrée est détectable par l'homme à des seuils très bas. Les cétones non cycliques sont très

volatiles et leur diffusion dans l'atmosphère sera importante et rapide. Elles sont toutes inflammables, le point d'éclair des cétones les plus utilisées étant souvent inférieur à 21 °C et parfois même inférieur à 0 °C (cas de l'acétone). Les vapeurs peuvent former des mélanges explosifs avec l'air. Les solutions aqueuses peuvent aussi s'enflammer aisément. Elles se dissolvent dans l'eau et dans un grand nombre de solvants organiques. Les cétones ne sont pas corrosives pour les métaux mais attaquent ou ramollissent les matières plastiques et les caoutchoucs [18].

I-5-3-Propriétés physico-chimiques des esters :

Les esters sont des liquides incolores. Les acétates sont volatils à température ambiante alors que les esters d'acides dicarboxyliques ou les agro solvants possèdent des tensions de vapeur relativement faibles. Tous les esters ont une odeur agréable et légère, souvent caractérisée de fruitée. Les acétates sont perceptibles à l'odorat à des valeurs de concentration dans l'atmosphère très faibles. Ils sont tous solubles dans de nombreux solvants organiques mais peu ou pas solubles dans l'eau (exception faite de l'acétate de méthyle). Leurs caractéristiques d'inflammabilité dépendent des substances. En effet, les acétates les plus légers (acétate de méthyle, acétate d'éthyle, acétate de propyle et acétate d'isobutyle par exemple) sont facilement inflammables (point d'éclair <21°C). Les autres esters, tout en restant combustibles, ne sont pas considérés comme inflammables [19].

I-5-4-Propriétés physico-chimiques des dérivés halogénés :

À part les plus petites molécules qui sont gazeuses, tous les dérivés halogénés couramment utilisés sont des liquides incolores.

Les dérivés halogénés ont des points d'ébullition plus élevés que ceux des hydrocarbures correspondants. Ils sont moins volatils. Les solvants chlorés ne sont pas ou peu inflammables, de même que les dérivés fluorés.

Certains dérivés bromés, quant à eux, sont considérés comme inflammables même si leur point d'éclair est difficilement mesurable (la méthode de mesure du point d'éclair peut être influencée par les atomes d'halogène).

Ils sont pratiquement insolubles dans l'eau mais sont de bons solvants pour de nombreux composés organiques, entre autres les corps gras.

Ils sont généralement plus fortement odorants que les hydrocarbures dont ils dérivent. Leur odeur, souvent agréable et éthérée, est détectable à des concentrations assez faibles pour certains composés.

Les solvants commerciaux sont généralement stabilisés par de petites quantités d'additifs antioxydants qui évitent la dégradation des produits en présence d'air, de lumière ou d'humidité, ou encore lors du contact avec l'aluminium ou les métaux légers [20].

I-5-5-Les propriétés physico-chimiques des alcanes :

Les propriétés physico-chimiques des alcanes (et des hydrocarbures en général) sont liées au nombre d'atomes de carbone de la chaîne.

A-Solubilité :

Les alcanes sont miscibles entre eux et insolubles dans l'eau. Ils sont solubles dans la plupart des solvants organiques [21].

Remarque : En terme général, un solvant est une substance qui sert à dissoudre une autre substance. Dans le contexte industriel, on se limite aux solvants organiques, c'est -à-dire ceux qui contiennent au moins un atome de carbone dans leurs structure moléculaires.

Un solvant organique est un composé chimique ou mélange qui est liquide entre 0°C et 250°C approximativement, qui est volatil et relativement inerte chimiquement.

Les solvants sont utilisés industriellement pour extraire, dissoudre ou suspendre des substances généralement insolubles dans l'eau (l'eau n'est donc pas un solvant organique) ou pour modifier les propriétés physiques d'un matériau.

Le concept de solvant organique ne doit pas être confondu avec celui des corps organiques volatils (COVs) que l'on retrouve dans la règlementation environnementale à protéger la qualité de l'atmosphère. Le terme solvant a une dimension utilitaire alors que les COVs sont définis en termes de réactivité photochimique dans l'atmosphère et de tension de vapeur minimale, généralement autour de 13,3Pa (soit 0,1mmHg) à 25°C.

B-Fusion et ébullition :

Les températures de fusion et d'ébullition des alcanes augmentent avec leur nombre d'atomes de carbone.

La température d'ébullition d'un alcane ramifié est inférieure à celle de son isomère linéaire.

De la même manière, un alcane ramifié s'auto-enflamme moins facilement que son isomère linéaire. On préfère donc utiliser des alcanes ramifiés dans les essences.

– Les températures d'ébullition étant différentes, les alcanes peuvent être séparés par distillation lors du raffinage du pétrole.

C-Densité :

La densité des alcanes augmente avec le nombre d'atomes de carbone. Elle est toujours inférieure à 1.

I-5-6-Propriétés physiques et chimiques des alcènes :

Les propriétés physiques et chimiques des alcènes sont très voisines de celles des alcanes; les 3 premiers membres de la série sont gazeux, ceux de C_5 à C_{15} liquides et les plus lourds solides à la température ambiante. L'introduction d'une insaturation dans une molécule accroît sa lipophilie.

I-5-7-Propriétés physicochimiques des alcynes :

A-Températures de changement d'état :

Sous une pression donnée, les températures de fusion et d'ébullition des espèces chimiques à chaînes linéaires d'une même famille augmentent lorsque la longueur de la chaîne carbonée augmente par exemple. $T_{éb}\ C_2H_4 = -102°C$ et $T_{éb}\ C_2H_2 = -83°C$

B-Importance industrielle :

L'éthyne (ou acétylène) est obtenu industriellement par réduction du carbonate de calcium par le coke à haute température (vers 1500°C) :

$$CaCO_3 + 3C \rightarrow CO + CaC_2$$

(Carbure de calcium, structure de type Na Cl),

l'anion étant : $(C^- \equiv C^-)$

Celui-ci s'hydrolyse très facilement (lampes à acétylène) en éthyne :

$$CaC_2 + 2H_2O \rightarrow Ca(OH)_2 + HC \equiv CH$$

L'enthalpie de combustion de C_2H_2 est très élevée, la flamme oxyacétylénique pouvant atteindre une température de 2900°C. Aussi la manipulation de l'éthyne est-elle hasardeuse à l'état sec. L'odeur caractéristique de l'acétylène obtenu par hydrolyse provient

de l'hydrolyse concomitante du phosphure de calcium, produit parasite de la réaction de $CaCO_3$ avec le carbone contenant du phosphore :

$$Ca_3 + H_2O \rightarrow Ca(OH)_2 + PH_3 \quad \text{(Phosphine nauséabonde)}$$

C-Acidité :

La forte électronégativité du carbone hybridé sp stabilise les charges négatives (plus que ne le fait l'azote par exemple). Le pK_A du couple acide / base / $RC{\equiv}CH$ amidure $RC{\equiv}C^-$ sodium transformera donc totalement l'alcyne en alcynure. L'hydroxyde de sodium ne donne lieu qu'à un équilibre partiel :

$$CaCO_3 + 3C \square \qquad aC_2$$

Cet équilibre peut être déplacé vers la droite en précipitant l'anion alcynure avec des cations tels que Cu^+ ou Ag^+. On obtient des composés insolubles dans l'eau et de diverses couleurs $CuC \equiv CCu$ est rouge, $RC \equiv CCu$ est jaune, $RC \equiv CAg$ est blanc.

Tous ces composés sont explosifs à sec [22].

I-6-Toxicité des composés organiques volatils chez l'Homme:

Les effets à court terme sont les plus faciles à mettre en liaison avec la pollution atmosphérique, du fait de la relation chronologique étroite entre les épisodes de pollution, et l'apparition des symptômes au sein de la population exposée. Ils se localisent essentiellement au niveau de l'appareil respiratoire, et affectent davantage les populations sensibles comme les jeunes enfants ou les sujets asthmatiques.

Les effets à long terme sur la fonction respiratoire sont plus difficiles à évaluer, du fait du délai d'apparition et de l'origine multifactorielle des maladies respiratoires dégénératives et les cancers de l'appareil respiratoire. Cependant, à risque tabagique égal, il apparaît que les pathologies respiratoires et les cancers soient plus fréquents dans les zones à pollution élevée.

La toxicité des benzènes, toluène, ethylbenzène, xylènes (BTEX) chez l'Homme a été particulièrement étudiée en raison de cas d'intoxiciations professionnelles révélées, ce qui a entraîné des études toxicologiques approfondies [14].

Le tableau I.5 synthétise les principaux effets induits chez l'Homme par des

expositions aiguës ou chroniques au benzène, toluène, ethylbenzène, xylènes **(**BTEX).

Tableau I.5: Toxicité des composés organiques volatils majeurs: benzène, toluène, ethylbenzène, xylènes [14].

Composé	Effets toxiques à court terme	Effets toxiques à long terme
Benzène	Mort à 20 000 ppm pendant 5 à 10 minutes. Atteinte de la moëlle osseuse; irritation des yeux; troubles cardiaques et digestifs; troubles du système nerveux; céphalées.	Effets hémotoxiques et immunotoxiques Cancérogène (leucémie) classe 1 Mutagène classe 2.
Toluène	Irritation des yeux; troubles cardiaques et digestifs; troubles du système nerveux; céphalées; vertiges; somnolence; muqueuses irritées.	Effets neurologiques Tératogène (classe.3 Union Européenne).
Ethylbenzène	Irritation des yeux, nez, muqueuses; atteintes neurologiques, hépatiques et rénales.	Atteintes hépatiques, rénales et du système hématologique.
Xylènes	Mort à 43500 $mg.m^{-3}$ irritation du nez, de la gorge; tachycardie ; perte de mémoire; déséquilibre.	Atteintes des voies respiratoires; troubles hématologiques mais difficulté de diagnostic du fait d'expositions toujours à des polluants

Compte tenu de la dangerosité de ces composés organiques, il apparaît nécessaire d'approfondir l'étude de leurs mécanismes d'action et d'optimiser des paramètres indicateurs d'exposition pouvant servir d'indicateurs d'effets précoces, pour une mesure en amont du développement d'une pathologie cliniquement décelable.

I-6-1-Toxicité des alcools :

L'alcool est une substance toxique liée à plus de 60 troubles différents. Pour certaines maladies chroniques dans lesquelles il est impliqué, comme le cancer du sein chez la femme, le risque augmente avec l'importance de la consommation alcoolique, sans qu'il y ait apparemment d'effet de seuil. Pour certaines autres maladies, comme la cirrhose du foie,

le risque est curvilinéaire et augmente avec l'augmentation de la consommation. L'alcool est un tératogène puissant. La conséquence la plus grave d'un alcoolisme pendant la grossesse est le syndrome d'alcoolisme foetal, un trouble du développement caractérisé par des anomalies craniofaciales, un retard de croissance et des lésions du système nerveux pouvant entraîner un handicap mental.

De tous les alcools, le plus toxique est le méthanol dans la mesure où il exerce une action sélective au niveau du nerf optique, pouvant provoquer la cécité ou la mort. Les effets néfastes de l'absorption d'éthanol sont aussi bien connus, l'alcoolémie entraînant notamment des incoordinations motrices ou une excitation intellectuelle. De manière générale, les manifestations d'une intoxication modérée se traduiront par des maux de tête, des troubles digestifs et un syndrome ébrieux. Les alcools liquides et leurs vapeurs sont irritants pour la peau, les yeux et les muqueuses en cas de contact prolongé ou répété.

L'inhalation accidentelle d'une grande quantité de vapeurs d'alcool peut conduire à des syndromes ébrieux ou narcotiques avec nausées, malaises, vomissements et maux de tête. L'exposition des salariés aux alcools, dans le cadre de leur activité professionnelle, peut provoquer des maladies reconnues et indemnisées par le régime général d'assurance maladie [23].

I-6-2-Toxicité des cétones :

La plupart des cétones simples ont des effets sur l'homme communs aux autres solvants : elles sont irritantes pour les voies respiratoires, la peau et les yeux et agissent sur le système nerveux central. Les premiers symptômes d'une exposition seront la toux, un larmoiement, des irritations cutanées mais aussi une diminution de la vigilance, des maux de tête etc….

À plus fortes concentrations ou lors d'expositions répétées peuvent apparaître des dermatoses ou des problèmes digestifs. De nombreuses cétones sont très facilement absorbées à travers la peau mais sont aussi rapidement évacuées par l'organisme humain. Une des cétones les plus toxiques est la méthylbutylcétone, utilisée parfois en laboratoire, qui, à doses répétées, peut induire des pertes de sensibilité pouvant dégénérer en déficit moteur [18].

I-6-3-Toxicité des esters :

On dispose de peu de données toxicologiques spécifiques pour ces substances de natures extrêmement diverses. La plupart de ces données sont jugées comme sans effet

particulier. Il faut toutefois noter que les acétates sont rapidement transformés au niveau de l'organisme, par des enzymes spéciales, en acide acétique et alcool correspondant, il conviendra donc de vérifier l'effet de ce dernier pour évaluer les dangers de l'ester. Un solvant doit être ici mentionné, la γ-butyrolactone. En milieu professionnel, elle pénètre dans l'organisme par voie respiratoire (vapeurs) et par voie cutanée. Mais aucune étude sur les effets d'une intoxication aiguë par voies dermale et inhalatoire n'est disponible. Les effets aigus liés à la γ-butyrolactone ont été décrits suite à des ingestions accidentelles surtout chez l'enfant et / ou des volontaires (dissolvant pour colle ou vernis à ongles à base de γ butyrolactone, suppléments alimentaires, toxicomanies). Ont été notamment observés des troubles de conscience avec des mouvements anormaux pouvant aller jusqu'au coma dans les cas les plus graves [19].

I-6-4-Toxicité des dérivés halogénés :

Les solvants présentent des caractéristiques communes plus ou moins marquées selon la substance et en même temps des propriétés toxicologiques propres à chaque produit.

Les effets communs incluent une irritation principalement de la peau et des muqueuses (oculaire et respiratoire) en cas d'exposition unique ou répétée, des troubles neurologiques aigus (somnolence, ébriété, céphalée, vertige, coma…) en cas d'exposition à des concentrations élevées, et surtout une atteinte neurologique plus progressive en relation avec des expositions répétées. Cette encéphalopathie se traduit notamment par des troubles de la mémoire et du comportement d'aggravation progressive tant que l'exposition persiste.

Les effets spécifiques peuvent être très différents selon la substance. Toutefois, la plupart peuvent provoquer des troubles d'excitabilité cardiaque ; de nombreux dérivés chlorés et bromés entraînent des atteintes hépatiques ou rénales.

Il faut signaler l'action particulièrement sévère du chlorure de méthyle sur le système nerveux (polynévrite et atteinte du système nerveux central).

Des solvants comme le tétrachlorure de carbone, le trichloroéthylène et de façon moins certaine le chloroforme, sont susceptibles d'induire des cancers. Le dichlorométhane peut être à l'origine d'intoxications au monoxyde de carbone par sa transformation dans l'organisme [20].

.I-6-5-Toxicité des alcanes :

Les alcanes sont peu toxiques. A forte dose, le n-hexane agit sur le système nerveux central (euphorie puis somnolence et vertiges). Par intoxication chronique, il a des effets neurotoxiques (dus à un métabolite, l'hexane-2,5-dione) mais n'est ni mutagène, ni cancérogène [24].

I-7- Emissions de COVs

I-7-1-Sources d'émissions

Les émissions de COVs sont classées en deux catégories : les émissions primaires et les émissions secondaires. Les émissions primaires correspondent aux COVs libérés directement dans l'atmosphère par évaporation de matériaux (solvants, plastifiants, biocides…). Les émissions secondaires des COVs sont généralement issues des différents processus de transformation chimiques des matériaux (combustion, décomposition…). Ces émissions proviennent de différents secteurs d'activité tels que le transport, l'industrie ou encore de sources biotiques comme la forêt, l'agriculture... La commission européenne a déterminé la formation de l'ozone troposphérique par secteur d'activité au cours des années 2000 et 2012 [25].

En 2000, les secteurs contribuant le plus à la formation de l'ozone à basse altitude sont les ménages, la fabrication et dans une moindre mesure, le transport. Il a été observé en 2012, une augmentation de 9% des émissions d'ozone issues du secteur des transports au détriment des ménages. Tous les autres secteurs ont enregistré des variations relativement faibles. Il semblerait que cette évolution soit en rapport avec la croissance économique des pays et le développement du parc automobile en Europe [26]. Les émissions des composés organiques volatils peuvent être d'origine anthropique ou naturelle. Les émissions naturelles sont celles qui proviennent des forêts, des cultures et des prairies. Les COVs émis dans l'atmosphère sont essentiellement issus de la famille des hydrocarbures. Les plus abondants sont l'isoprène et les terpènes. Les émissions anthropiques de type industrielles concernent les secteurs de l'imprimerie, traitement de surface, pharmacie, ... Les COVs émis sont issus des familles des esters (acétate d'éthyle…), alcools (méthanol…) ou encore des aromatiques et des dérivés chlorés (dichlorométhane). Il apparaît dans la littérature que les secteurs de l'industrie manufacturière et du résidentiel/tertiaire contribuent fortement à l'émission des COVs dans l'atmosphère [25,27].

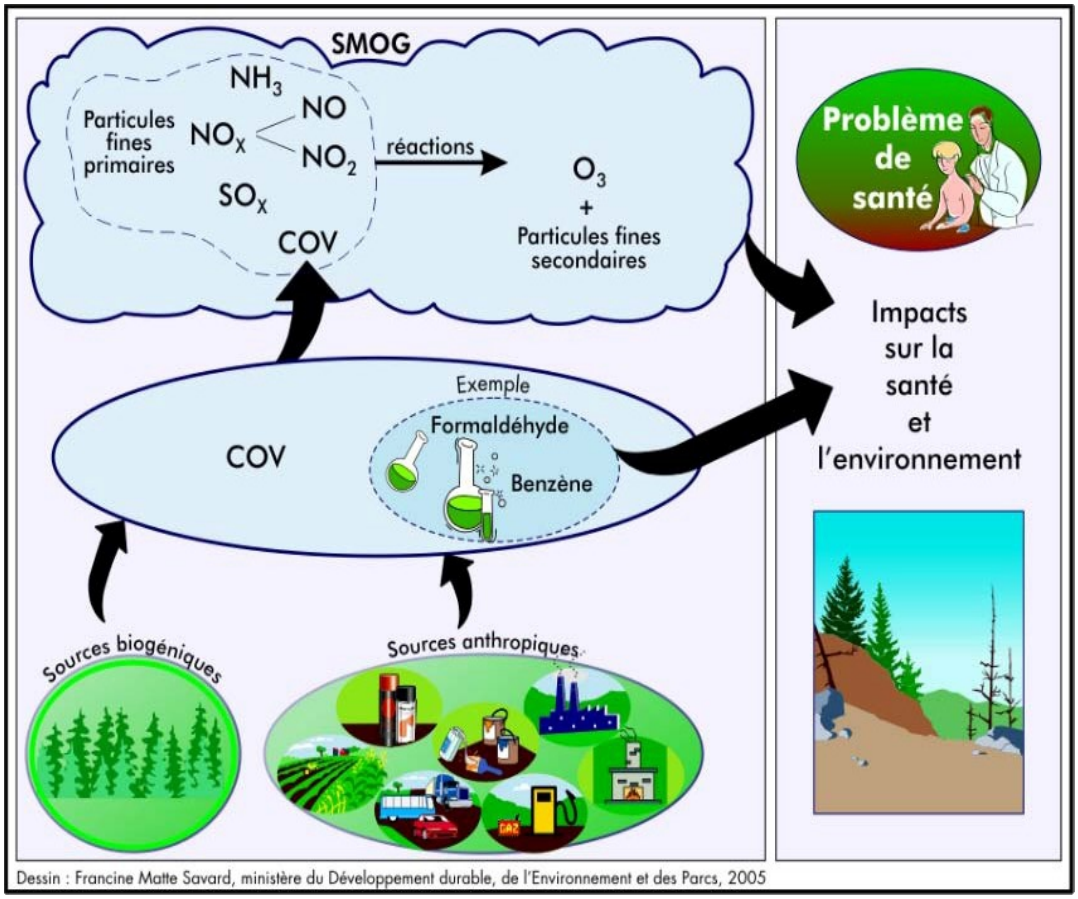


Figure I.1: Sources des COVs participant à la formation du smog et occasionnant certains problèmes de santé [28].

I-7-2-Quantification des émissions

❖ **Par secteur d'activité**

Dans le cadre du protocole de Göteborg [29], 26 pays se sont engagés le 1er décembre 1999 devant la Commission Economique pour l'Europe des Nations Unies (CEE-NU) dans une politique de réductions des émissions de COVs. Afin de respecter les objectifs fixés, les quantifications des émissions de COVs sont actualisées chaque année. Les règles de comptabilisation des émissions de COVs ont été réalisées conformément aux recommandations de la CEE-NU et de la CCNUCC (Convention Cadre des Nations Unies pour les Changements Climatiques). Ces règles ont été généralisées en France et adoptées par le CITEPA lequel est en charge, au niveau national français, de la réalisation des inventaires d'émissions à la demande du Ministère en charge de l'écologie. En France, conformément à l'arrêté du 24 août 2011 et à l'aide du SNIEBA (Système National d'Inventaires d'Emissions et de Bilans dans l'Atmosphère), le CITEPA élabore les inventaires au format SECTEN

(Secteur Economique et Energie) c'est-à-dire sous forme d'image reformulée des données de base. D'autres critères de mesure sont pris en compte notamment en termes d'émissions totales ou hors-total.

En effet, les émissions de COVs comptabilisées à l'échelle nationale française sont dites "totales". Celles appelées " hors-total " exclues du total national sont les émissions provenant des secteurs d'activités tels que le secteur maritime (international), les trafics aériens (domestiques et internationaux) et les sources biotiques (agriculture et forêts). En 2011, le CITEPA a comptabilisé les émissions totales de COVs de 7 secteurs d'activités (résidentiel/tertiaire, industrie manufacturière…) dont 43 sous-secteurs [30].

Par risque sanitaire

La quantification des émissions de COVs permet de mesurer leur seuil de toxicité sur l'organisme.

Deux critères d'évaluations des limites de toxicité des COVs ont été définis par des valeurs correspondant aux concentrations de COVs dans l'atmosphère sans risques d'altération de la santé. Le premier critère est la Valeur Limite d'Exposition (VLE) laquelle correspond au seuil de concentration limite d'exposition d'un individu, durant 15 minutes, sans altération physiologique immédiate.

Le deuxième critère appelé La Valeur limite Moyenne d'Exposition (VME), correspond à la concentration limite d'exposition d'un individu durant une période de 8h par jour et de 40h par semaine sans altérations à long terme sur l'organisme. Ces deux valeurs sont spécifiques à chaque COV et s'expriment en partie par million (ppm) ou en milligramme par mètre cube (mg/m3).

I-8-Impacts environnementaux

Les émissions de COVs dans l'atmosphère engendrent des impacts directs et indirects sur la santé humaine et l'environnement. Les impacts directs affectent immédiatement la santé tandis que les impacts indirects sont dus à la production de l'ozone dans des conditions particulières de température et de rayonnement solaire.

I-8-1-Impacts directs

Les effets des COVs sur l'Homme varient en fonction de la nature chimique de chaque polluant. Généralement, une gêne des voies respiratoires (nez, gorge) ainsi qu'une irritation des yeux et de la peau sont constatées après une exposition longue. La gêne olfactive créée par certains composés organiques volatils (benzène, formaldéhyde…) a des effets mutagènes et cancérigènes [31]. Il est également constaté une dépression du système nerveux central traduit par des maux de têtes, nausées, vomissements...etc

Certains COVs pourraient également nuire à la reproduction. De fortes concentrations de COVs dans l'organisme limiteraient le développement prénatal et postnatal. Les composés les plus préoccupants à cet effet seraient essentiellement le toluène et les éthers de glycol.

I-8-2-Impacts indirects

Les COVs contribuent indirectement à la pollution atmosphérique. Lorsqu'ils sont libérés dans l'atmosphère, ils participent aux réactions photochimiques à l'origine de la formation de l'ozone (O_3). L'ozone troposphérique généralement mis en cause à une altitude comprise entre 0 et 10 km. Il est vrai que l'ozone existe naturellement dans l'air à des concentrations comprises entre 0,005 et 0,05 ppm [32].

Cependant, à concentration élevée, la molécule provoque des lésions pulmonaires sévères (œdèmes pulmonaires…) à cause de son action irritante mais également des complications rénales, neurologiques... Les COVs sont impliqués dans les réactions photochimiques en tant que précurseurs de composés oxydants. Lorsqu'ils sont libérés dans l'atmosphère, ils se dégradent de manières à perturber les équilibres chimiques dans l'air. En effet, l'ozone est naturellement formé dans l'air par recombinaison d'un atome d'oxygène, issu de la photodissociation du dioxyde d'azote (NO_2), avec l'oxygène moléculaire O_2. Lorsque ce cycle est perturbé par la présence de COV dans l'atmosphère, les mécanismes d'oxydation de ces composés, via le radical OH• et en présence des molécules d'azote (NOx), augmentent la production d'ozone [8].

Parallèlement à ce phénomène, la présence de composés organiques volatils dans l'atmosphère engendre d'une autre manière, un effet de serre au niveau de la troposphère en réfléchissant les rayons infrarouges émis par le rayonnement solaire à la surface de la Terre.

I-9-Méthodes de traitement des COVs

Les COVs sont traités par différents procédés dont le processus est soit destructif, soit de récupération. Certains traitements utilisés sont, pour la plupart, spécifique à l'élimination d'un COV ou de familles de COV. D'autres, de manière plus globale, sont utilisés pour le traitement des odeurs. Parmi les procédés de récupération, on retrouve : la condensation, l'absorption et l'adsorption [8]. Les procédés destructifs sont essentiellement les procédés d'oxydation thermique, la voie biologique ou l'oxydation catalytique [33].

I-9-1-Techniques de récupération

❖ **La condensation**

Le traitement par condensation est réservé aux fortes concentrations de COVs. Le procédé privilégie les faibles débits (inférieurs à 1000 m^3/h) afin de pouvoir récupérer les molécules sans modification de la composition. La technique de traitement consiste à récupérer le COV sous forme liquide en abaissant, à l'aide de la température, sa pression de vapeur saturante. On distingue deux types de techniques :

– La condensation mécanique : limitée aux températures -30°C et -40°C et élaborée à l'aide de compresseurs ou d'échangeurs;

– La condensation cryogénique : effectuée à l'aide d'azote liquide car les températures peuvent atteindre -180°C.

L'avantage d'utiliser cette méthode réside dans la récupération quasi-totale du polluant. Par contre, un inconvénient apparaît au niveau du traitement des faibles débits de COV lesquels limitent les performances d'analyse.

❖ **L'absorption**

La technique d'absorption des COVs consiste en un transfert des molécules gazeuses vers un liquide de lavage potentiellement réactif. La régénération du liquide se fait généralement par distillation sous vide et permet la récupération du polluant. La solution utilisée est soit de l'eau, soit une solution aqueuse acide, basique ou oxydante. Il est à noter que l'utilisation de cette méthode est relativement simple et rapide. Cependant, les limites de ce procédé sont liées à l'usage de solutions adaptées au COV et au risque de transfert de pollution par régénération du liquide.

❖ **L'adsorption**

La technique d'adsorption met en jeu deux étapes dans laquelle le COV est d'abord retenu sur un support sélectif (charbon actif…), puis désorbé sous l'effet d'une diminution de la pression (sous vide) ou d'une augmentation de la température (gaz neutre chaud…) pour être récupéré. Cette technique peut être utilisée pour les COVs difficilement traitables tels que les COVs chlorés. Cependant, la régénération est indispensable afin de récupérer le COV, ce qui augmenterait le coût d'exploitation.

I-9-2-Techniques destructives

❖ **Oxydation thermique**

L'oxydation thermique est une technique énergivore car elle consiste à brûler les molécules polluantes aux températures supérieures à 750°C, afin de les transformer en dioxyde de carbone (CO_2) et en eau (H_2O). Elle représente 60% des traitements mis en place dans l'industrie [8]. Cependant, ce procédé est limité par son incapacité à traiter les composés organiques volatils possédant un hétéroatome (soufre, azote…) dans la mesure où l'oxydation favoriserait la formation de gaz toxiques tels le SO_2, les NOx, ou encore Cl_2 gazeux...

❖ **Oxydation par voie biologique**

L'oxydation des COVs par traitement biologique représente actuellement 5% des méthodes utilisées pour leur élimination [8]. La technique consiste à utiliser des bio filtres de telle manière qu'au contact des COVs, les microorganismes fixés sur le support biologique (tourbes, copeaux de bois…) dégradent le polluant en présence d'oxygène. Les propriétés physico-chimiques du polluant à dégrader sont directement liées à l'efficacité d'un tel système. En effet, la solubilité du COV et sa biodégradabilité sont des caractéristiques non négligeables. Les COVs capables d'être dégradés par cette méthode sont des familles de composés oxygénés tels que des alcools, cétones, esters.... Les composés organiques ayant un hétéroatome (chlore, soufre…) sont difficilement biodégradables.

❖ **Oxydation catalytique**

Le principe de la méthode est le même qu'en oxydation thermique. Les composés organiques volatils sont transformés en molécules inorganiques CO_2 et H_2O. Cependant, l'oxydation catalytique est une technique de choix car elle nécessite moins d'énergie que l'oxydation thermique [34]. En effet, les conditions de mise en œuvre (débit, concentration, température...) sont adaptées aux traitements d'un COV où d'un mélange de COV. Les réactions sont réalisées à basse température ($T<500°C$) en présence d'un catalyseur ce qui

diminue l'apport énergétique d'une part, et limite, d'autre part, la formation d'oxydes d'azote NOx. Le catalyseur peut être utilisé sous forme de bille poreuse, de nids d'abeilles monolithiques, de poudre... Il peut être à base de métaux précieux (platine, rhodium, palladium...) ou à base d'oxydes métalliques (cuivre, cobalt, nickel...). Les concentrations des polluants peuvent être très faibles ainsi que les débits de gaz. L'oxydation catalytique des COVs est une technique destructive dont la réaction d'oxydation par l'oxygène correspond [33] à la réaction:

$$C_aH_b+(a+\frac{b}{4})O_2 \rightarrow aCO_2+\frac{b}{2}H_2O$$

Où a et b sont respectivement le nombre de carbone et d'hydrogène de la molécule organique. La vitesse de réaction dépend de la concentration du polluant et de l'oxygène ainsi que de la température. Le mélange polluant/oxygène (O_2 en excès) doit être homogène et la température doit être suffisamment élevée pour initier la réaction. La présence d'un catalyseur permet de diminuer l'énergie d'activation de la réaction ce qui favorise l'activité catalytique à basse température. Malgré tout, l'inconvénient de ce système réside dans l'éventualité d'un empoisonnement du catalyseur (masquage de sites actifs, effets thermiques, perte de matière…) à l'origine de sa désactivation.

I-9-3-Techniques émergentes

L'élimination des composés organiques volatils dans l'atmosphère est une problématique majeure pour les laboratoires de recherche et développement (R&D) spécialisés dans l'environnement. A cet effet, l'élaboration de nouvelles techniques de traitement à petite échelle présente un intérêt technologique en termes de coût et de faisabilité en industrie. Parmi ces techniques, on retrouve : la photocatalyse, les procédés à membranes tels que la perméation gazeuse, ou encore l'oxydation par plasma froid.

- **Photocatalyse**

La photocatalyse est une technique innovante qui consiste à oxyder le polluant sur un catalyseur semi-conducteur à l'aide d'un rayonnement photonique apporté par la lumière. Son principe s'apparente à de la catalyse hétérogène où le réactif est décomposé à la surface du catalyseur. Les catalyseurs utilisés sont généralement des oxydes métalliques (oxyde de titane…) supportés sur des supports inertes (alumine, silice…). L'inconvénient de cette technique repose sur l'incapacité à maîtriser les sous-produits d'oxydation ainsi que l'empoisonnement du catalyseur.

❖ **Plasma froid**

L'oxydation des COVs par plasma est une technique prometteuse en termes de gain énergétique.

En effet, le plasma est formé à partir d'un gaz ionisé à faible énergie lequel est "froid" car sa température reste proche de la température ambiante (25°C) [35]. La décomposition du COV est réalisée à pression atmosphérique. Les effets induits par le plasma fragilisent les liaisons C-H des COVs jusqu'à la formation de composés inorganiques CO_2 et H_2O. La difficulté de cette technique est relative à sa faisabilité à grande échelle et à la maîtrise des sous-produits d'oxydation.

❖ **Perméation gazeuse**

La perméation gazeuse est une technique de séparation et de purification des gaz [36]. C'est une technique de récupération des constituants gazeux, développée et appliquée à l'échelle industrielle. Son principe est essentiellement basé sur les vitesses de perméation à travers une membrane des composants gazeux. Son efficacité dépend d'un gradient de pression et une force motrice suffisamment élevés pour permettre le transfert à travers la membrane. Elle nécessite une faible consommation d'énergie mais les puretés des produits ne sont pas toujours très élevées. En définitive, les applications industrielles des différents traitements des composés organiques volatils dépendent fortement du débit de flux et de la concentration en COV à traiter. Ces caractéristiques mettent en jeu la faisabilité économique et l'efficacité du procédé [37].

I-10-Normes sur la qualité de l'air en Algérie :

L'Algérie se dote de Normes sur la qualité de l'air à partir de la publication au journal officiel du Décret exécutif n° 06-02 du 7 janvier 2006 définissant les valeurs limites, les seuils d'alerte et les objectifs de qualité de l'air en cas de pollution atmosphérique qui sont présentés sur le tableau I.6 (Journal officiel de la république algérienne n° 01 du 8 janvier 2006. http//www.joradp.dz). Ces normes algériennes sur la qualité de l'air sont surtout destinées aux industriels [38].

Tableau I.6: Normes algériennes sur la qualité de l'air [38].

Les polluants atmosphériques				
	Dioxyde d'azote (NO_2)en(µg/m³)	Dioxyde de Soufre (SO_2) en (µg/m³)	Ozone (O_3) en (µg/m³)	Particules fines en suspension en(µg/m³)
Objectif de qualité	135	150	110	50
Valeur limite	200	350	200	80
Seuil d'information	400	350	180	-
Seuil d'alerte	600	600	360	-

CHAPITRE II

Outils et techniques utilisés

I-a-LA MODÉLISATION MOLÉCULAIRE

La modélisation moléculaire peut être considérée comme un ensemble de techniques informatiques basées sur des méthodes de chimie théorique et les données expérimentales qui peuvent être utilisées pour la description et la prédiction des propriétés des systèmes moléculaires complexes dans différents environnements.

Cette approche procède de l'hypothèse d'une correspondance univoque entre n'importe quelle propriété physique, affinité chimique, ou activité biologique d'un composé chimique et sa structure moléculaire.

La stabilité de la structure tridimenensionnelle d'une molécule est déterminée par les interactions intramoléculaires et les interactions avec le milieu extérieur (solvant). La recherche des conformations stables d'une molécule consiste à déterminer les minima de l'énergie globale d'interaction. Cette énergie peut être calculée par des méthodes quantiques *ab initio* ou semi-empiriques généralement longues et onéreuses. Pour faciliter les calculs, on considère habituellement que le terme variable de cette énergie dépend de la construction de la molécule et de l'arrangement de ses atomes : c'est le principe des méthodes empiriques (mécanique moléculaire, dynamique moléculaire). Dans la plupart de ces méthodes, il n'est pas tenu compte des interactions avec le solvant, mais uniquement des interactions entre les atomes constitutifs de la molécule. La recherche d'une conformation consiste alors à faire une minimisation de le l'énergie intramoléculaire. Cette énergie potentielle est fractionnée en un certain nombre de termes additifs indépendants. Chacun de ces termes est représenté par une fonction analytique simple justifiée par des calculs quantiques et incluant des paramètres empiriques.

II-b-OPTIMISATION DE LA GEOMETRIE DES MOLECULES

II-b-1- La Méthode de HARTREE-FOCK-ROOTHAAN (Méthode de HFR)

II-b-1-1-Energie d'un micro système représenté par un déterminant de Slater

Les calculs quanto-mécaniques courants sont basés sur le modèle de l'électron indépendant où l'on suppose les orbitales soit vides soit garnies de deux électrons au plus.

Dans le cadre de ce modèle, la fonction d'onde polyélectronique peut s'écrire sous la forme d'un produit anti-symétrisé de spin-orbitales :

$$\psi(1,2,\dots,n)=\frac{1}{\sqrt{n!}}\begin{vmatrix}\psi_1(1) & \bar{\psi}_1(1)\dots\dots\dots\bar{\psi}_n(1)\\ \psi_1(2) & \bar{\psi}_1(2)\dots\dots\dots\bar{\psi}_n(2)\\ \vdots & \vdots \quad\quad \vdots\\ \vdots & \vdots \quad\quad \vdots\\ \psi_1(n) & \bar{\psi}_1(n)\dots\dots\dots\bar{\psi}_n(n)\end{vmatrix} \quad (1.II.1)$$

Les spin-orbitales sont obtenues en multipliant chaque orbitale par l'une des deux fonctions de spin possibles :

$$\psi_m(n)=\varphi_m(n)\alpha(n) \quad (1.II.2)$$
$$\bar{\psi}_m(n)=\varphi_m(n)\beta(n)$$

Nous considérerons le cas des systèmes à couches complètes (gaz inertes, molécules courantes dans l'état fondamental…..) pour lesquels $n=2m$.

La fonction déterminantale $\psi(1,2,3,\dots,n)$ est appelée **déterminant de Slater**.

L'hamiltonien du système est l'hamiltonien résultant, à l'approximation de Born-Oppenheimer.

$$H(1,2,\dots,n)=\sum_{i=1}^{n}h_{(i)}^{c}+\sum_{i<j}\frac{e^2}{r_{ij}} \quad (1.II.3)$$

$h_{(i)}^{c}$: est **l'hamiltonien monoélectronique de cœur** ; le symbole $\sum_{i<j}\frac{e^2}{r_{ij}}$désigne une sommation sur couples ordonnés.

Comme ψ est normé à l'unité (constante de normalisation$1/\sqrt{n!}$), l'énergie du système est donnée par :

$$E=<\psi|H|\psi> \quad (1.II.4)$$

Lorsqu'on développe cette intégrale on arrive [39] au résultat :

$$E=\sum_{i=1}^{m}2h_{ii}^{c}+\sum_{i=1}^{m}\sum_{j=1}^{m}(2J_{ij}-K_{ij}) \quad (1.II.5)$$

L'écriture$\sum_{i=1}^{m}$, signifie que l'on somme sur toutes les orbitales occupées.

$$\boldsymbol{h_{ii}^{c}=<\psi_i(\mu)\left|h_{(\mu)}^{c}\right|\psi_i(\mu)>} \quad (1.II.6)$$

est **l'intégrale monoélectronique moléculaire de cœur**, intégrale triple qui porte sur les coordonnées d'un seul électron : le $\mu^{\text{ème}}$ dans ce cas.

$$\boldsymbol{J_{ij}=\iint\psi_i^*(\mu)\psi_i(\mu)\frac{e^2}{r_{\mu\nu}}\psi_j^*(\nu)\psi_j(\nu)d\tau_\mu d\tau_\nu} \quad (1.II.7)$$

est **l'intégrale monoélectronique moléculaire coulombienne**, parce qu'elle représente une somme de termes d'interactions coulombiennes, intégrale sextuple qui porte sur les coordonnées de deux électrons.

$$\boldsymbol{K_{ij}=\iint\psi_i^*(\mu)\psi_i^*(\nu)\frac{e^2}{r_{\mu\nu}}\psi_j(\mu)\psi_j(\nu)d\tau_\mu d\tau_\nu} \quad (1.II.8)$$

est **l'intégrale biélectronique moléculaire d'échange** ; elle représente également une somme de répulsions entre charges élémentaires, l'électron occupant deux orbitales moléculaires ψ_i et ψ_j. $r_{\mu\nu}$ représente la distance entre les deux électron μ et ν.

Remarques :

1)- Dans l'expression de l'énergie *E*, nous trouvons deux termes :

*- E^c, qui est l'énergie de l'ensemble des électrons évoluant dans le champ des noyaux sans interactions les uns avec les autres.

*-E^{RE}, qui est l'énergie de répulsion électronique.

$$E=E^c+E^{RE} \qquad (1.II.9)$$

Evidemment si l'on suppose qu'il n'existe pas d'interactions entre électrons, le second terme disparaît complètement.

2)- Si on a à traiter une molécule, il faut ajouter un terme supplémentaire de répulsion nucléaire.

$$E_T = E + \sum_{N<L} \frac{Z_K E_L e^2}{R_{KL}} \qquad (1.II.10)$$

Z_K et Z_L sont les charges des noyaux *K* et *L* et R_{KL} la distance entre ces noyaux.

La relation (1.II.5) est équivalente à :

$$E = \sum_{i=1}^{m}\{h_{ii}^c + [h_{ii}^c + \sum_{j=1}^{m}(2J_{ij} + K_{ij}]\} \qquad (1.II.5)$$

Le terme :

$$e_i = h_{ii}^c + \sum_{j=1}^{m}(2J_{ij} + K_{ij}) \qquad (1.II.11)$$

correspond à ce qu'on appelle l'énergie des orbitales moléculaires.

E se réduit donc à :

$$E = \sum_{i=1}^{m}(h_{ii}^c + e_i) \qquad (1.II.12)$$

Remarque : Dans les méthodes approchées, comme la méthode de Slater par exemple, on prend :

$$E = \sum_{i=1}^{m} 2\, e_i \qquad (1.II.13)$$

Dans la méthode de Hatree-Fock-Roothaan ceci n'est plus vrai : l'énergie des micro-systèmes n'étant pas égale à la somme des énergies des orbitales moléculaires.

Pour qu'il en soit ainsi, il faudrait que $h_{ii}^c = e_i$ ce qui n'est pas vrai.

Les orbitales moléculaires ne sont pas connues. Le déterminant de Slater n'est connu que par rapport à un jeu de $\{\psi_i\}$ dont on ne sait rien, à part qu'elles sont orthogonales.

Le problème est de déterminer le jeu d'orbitales qui permet de construire le système de Slater.

II-b-1-2- Détermination des Orbitales ou équations de Hartree-Fock

On construit le système de Slater à partir d'un jeu de $\{\psi_i\}$.

Quelles propriétés doivent posséder les ψ_i pour être acceptables au sens de la mécanique ondulatoire, et qu'elles puissent s'adapter au système particulier envisagé ?
Il faut que le déterminant de Slater soit une solution approchée de l'équation de Schrödinger totale :

$$H(1,2,\ldots.,n)\psi(1,2,\ldots.,n) = E\psi(1,2,\ldots.,n) \qquad (1.II.14)$$

La propriété la plus fondamentale des solutions de l'équation de Schrödinger est leur stabilité : c'est-à-dire que si on fait subir à la fonction d'onde déterminantale une perturbation du premier ordre, il s'ensuit une perturbation du premier ordre de l'énergie nulle.
Il faut donc réaliser absolument cette condition.

Comme la variation du déterminant de Slater s'exprime par la variation du jeu des $\{\psi_i\}$, il faudrait avoir, pour une variation première du jeu d'orbitales choisies, une variation première de l'énergie totale nulle, et pour cela il faut que les ψ_i soient solutions des équations de Hartree-Fock [40-41]:

$$\{\delta\psi_i\} \rightarrow \delta E^1 = 0 \qquad (1.II.15)$$

Ces deux conditions contiennent les équations de Hartree-Fock :

$$F_{(\mu)}\psi_i(\mu) = e_i\psi_i(\mu) \qquad (1.II.16)$$

L'équation de Hartree-Fock est une équation intégro-différentielle qui, contrairement à une équation de Schrödinger mono-électronique, fait intervenir un opérateur F qui dépend des fonctions inconnues ψ_i .
Opérateur de Hartree-Fock :

$$\boldsymbol{F_{(\mu)} = \left[h^c_{(\mu)} + \sum_{i=1}^{m} 2J_l(\mu) - K_l(\mu)\right]} \qquad (1.II.17)$$

J_l et K_l sont, respectivement, les opérateurs coulombien et d'échange relatifs à chaque orbitale doublement occupéeψ_i.

II-b-1-3- Equations de Roothaan et Hall

Découlent de la méthode de Hartree-Fock lorsqu'on introduit la condition CLOA (Combinaison Linéaire des Orbitales Atomiques).

Chaque orbitale moléculaire psi se présentera sous la forme :

$$\psi_i(\mu) = \sum_{p=1}^{N} C_{pi}\varphi_p(\mu) \qquad (1.II.18)$$

L'ensemble des orbitales atomiques $\{\varphi_p\}$ étant supposé connues, la détermination des ψ_i se ramène à la détermination des C_{pi}.

Les équations de Hartree-Fock prennent, en tenant compte de (1.II.18), une expression vectorielle assez simple :

$$\sum_{p=1}^{N} C_{pi}\left[F_{pq} - e_i S_{pq}\right] = 0 \qquad , \qquad q \in [1, N] \tag{1.II.19}$$

Les coefficients :

$$\boldsymbol{S_{pq} = \int \varphi_p^* \, \varphi_q d\tau} \tag{1.II.20}$$

$$\boldsymbol{F_{pq} = \int \varphi_p^* \, (F\varphi_q) d\tau}$$

sont les intégrales de recouvrement sur la base des fonctions φ_p et les éléments matriciels de l'opérateur de Hartree-Fock F, et les valeurs propres sont les énergies orbitalaires *ei*.

L'équation (1.II.19) est un système linéaire homogène (*N* équations à *N* inconnues) qu'on peut écrire sous la forme matricielle :

$$[\boldsymbol{F} - \boldsymbol{e_i S}]C_i = \boldsymbol{0} \tag{1.II.21}$$

où $\boldsymbol{F}$ est la matrice $[F_{pq}]$; $\boldsymbol{S}$ est la matrice $[S_{pq}]$; $\boldsymbol{C}$ est la matrice $[C_{pi}]$.

$$F_{pq} = h_{pq}^c + \sum_{l=1}^{N}\sum_{m=1}^{N} p_{lm}[< pq \,|lm> - \tfrac{1}{2} < pm|\, lq >] \tag{1.II.22}$$

-* h_{pq}^c = intégrale monoélectronique sur les orbitales atomiques de base.

$$h_{pq}^c = \langle \varphi_p(\mu) | h_{(\mu)}^c | \varphi_q(\mu) \rangle \tag{1.II.23}$$

$$-*\langle pq|lm\rangle = \iint \varphi_p(\mu)\varphi_q(\mu)\frac{e^2}{r_{\mu\nu}}\varphi_l(\nu)\varphi_m(\nu)d\tau_\mu d\tau_\nu \tag{1.II.24}$$

$$\langle pm|lq\rangle = \int \varphi_p(\mu)\varphi_m(\mu)\frac{e^2}{r_{\mu\nu}}\varphi_l(\nu)\varphi_q(\nu)d\tau_\mu d\tau_\nu \tag{1.II.25}$$

-* $p_{lm} = \sum_{i=1}^{N} 2C_{li}C_{mi}$=éléments de la matrice densité (1.II.26)

-* $\boldsymbol{P} = [p_{lm}]$ = matrice densité (1.II.27)

II-b-1-4- Quelques remarques sur les processus de résolution des équations de Hartree-Fock-Roothaan.

L'équation de Hartree-Fock-Roothaan sous forme matricielle est :

$$\boldsymbol{F\, C_i = e_i\, S\, C_i} \tag{1.II.21}$$

Löwdin [43] a proposé un procédé qui permet de se ramener dans tous les cas au calcul des valeurs propres et vecteurs propres d'une matrice moyennant une transformation de la base des orbitales atomiques (**orthogonalisation de Löwdin**).

Multiplions à gauche les deux membres de (1.II.21) par la matrice $\mathbf{S}^{-1/2}$, qui n'est jamais singulière puisque **S** ne l'est pas ; il vient successivement :

$$\boldsymbol{S^{-1/2}\, F\, C_i = e_i\, S^{-1/2} S\, C_i}$$

$$\boldsymbol{[S^{-1/2}\, F\, I\, S^{-1/2}]\, S^{-1/2}\, C_i = e_i\, S^{-1/2}\, C_i}$$

Soit en posant :

$$\boldsymbol{S^{-1/2}\, F\, S^{-1/2} = F} \quad \text{et} \quad \boldsymbol{S^{-1/2}\, C_i = C_i} \qquad (1.II.28)$$

$$\boldsymbol{F\, C_i = \bar{e}_i\, \bar{C}_i}\,, \quad \text{c'est-à-dire } \boldsymbol{[F - e_i\, I]\, C_i = 0} \qquad (1.II.29)$$

Les équations de Hatree-Fock-Roothaan sont résolues selon un procédé itératif qui se fait sur l'ensemble orthogonalisé.

$$\boldsymbol{\bar{F}\, \bar{C} = e_i\, \bar{C}_i} \qquad (1.II.29)$$

On peut toujours initialiser le problème en choisissant a priori une matrice densité, obtenue en négligeant la matrice des interactions électroniques (problème d'ordre zéro). Le nombre d'itérations dépend du problème à résoudre.

II-b-1-5- Détermination des intégrales de la méthode de Hartree-Fock-Roothaan (HFR)

Le très gros problème dans la méthode HFR est la détermination des intégrales.

***- Intégrales monoélectroniques atomiques de cœur :**

$$h_{pq}^{c} = < \varphi_p(\mu) | h_{\mu}^{c} | \varphi_q(\mu) > \qquad (1.II.30)$$

Il existe deux types d'intégrales de ce genre : **monocentres** lorsque $\varphi_p$et $\varphi_q$appartiennent au même atome R ; **bicentres**, lorsque $\varphi_p$et $\varphi_q$appartiennent à des atomes différents.

Les intégrales monoélectroniques de cœur monocentres comprennent : **les intégrales de cœur coulombiennes** (même orbitale atomique des deux côtés) et **les intégrales de cœur d'échange** (les deux orbitales atomiques sont différentes).

$$h_{pq}^{c} = \underbrace{-\frac{\hbar^2}{2m} \int \varphi_p(\mu) \Delta_{(\mu)} \varphi_q(\mu) d\tau_\mu}_{Intégrales\ cinétiques} \underbrace{- \sum_k Z_K \int \varphi_p(\mu) \frac{e^2}{r_{k\mu}} \varphi_q(\mu) d\tau_\mu}_{intégrales\ d'attractions\ nucléaires} \qquad (1.II.31)$$

Les intégrales d'attractions nucléaires peuvent être monocentres, bicentres ou tricentres

(très compliquées à calculer).

***- Intégrales bi-électroniques**

$$G_{pq} = \sum_l \sum_m p_{lm} \left[\langle pq|lm\rangle - \frac{1}{2} \langle pm|lq\rangle\right] \quad (1.II.32)$$

$\langle pq|lm\rangle$, $\langle pm|lq\rangle$ et p_{lm} sont respectivement définis par les relations (1.II.24), (1.II.25) et (1.II.26).

On a plusieurs types d'intégrales :

- **monocentres,** lorsque, φ_p, φ_q, φ_l, φ_m appartiennent au même atome.
- **bicentres**, lorsque parmi, φ_p, φ_q, φ_l, $\varphi_m$il y en a qui appartiennent à deux atomes différents.
- **tricentres,** parmi, , φ_p, φ_q, φ_l, φ_m il y en a qui appartiennent à trois atomes différents.
- **tétracentres,** chaque orbitale appartient à un atome différent.

Le calcul des intégrales biélectroniques prend le plus grand temps, et il n'est pas possible, en prenant des orbitales de Slater (33) d'en donner des expressions analytiques.

$$\varphi_{n,l,m}(k,\vec{r}) = Nr^{n-1}e^{-kr}y_{l,m}(\theta,\varphi) \quad (1.II.33)$$

$y_{l,m}(\theta,\varphi)$ étant les harmoniques sphériques.

On décompose alors chaque orbitale de Slater en orbitales gaussiennes dont la partie radiale est de la forme e^{-kr^2}, ce qui permet de ramener un problème d'analyse numérique à un problème d'algèbre.

II-b-2- Les méthodes semi-empiriques

Dans le précédent chapitre, nous avons exposé la théorie des orbitales moléculaires **d'un point de vue ab-initio**, déterminant une fonction d'onde qui nécessite le calcul d'un certain nombre d'intégrales et l'utilisation d'une procédure algébrique auto-cohérente.

Dans le cadre de cette théorie, une approche plus approximative est développée, ce qui permet d'éviter l'évaluation difficile de beaucoup d'intégrales et de sélectionner les valeurs de certaines autres en tenant compte des données expérimentales.

Les approches semi-empiriques, qui traitent des électrons de valence, sont désignées par des sigles dont les lettres correspondent aux approximations admises dans le recouvrement différentiel des orbitales.

II-b-2-1- Définition du semi-empirisme

Une méthode est semi-empirique si elle admet le cadre de Hatree-Fock-Roothan, en y incorporant un certain nombre de simplifications.

On arrive ainsi à réduire considérablement le nombre d'intégrales. En particulier on élimine les intégrales biéléctroniques à 3 et 4 centres, qui sont très faibles.

Une fois le cadre HFR simplifié, on évalue empiriquement les intégrales restantes en ajustant la méthode sur des molécules bien connues.

II-b-2-2- Quelques théories semi-empiriques

La première théorie semi-empirique, ou théorie de Pople-Pariser-Parr (PPP), introduite en 1953 par Pariser et Parr [44,45], et utilisée la même année par Pople [46], permet d'étudier les systèmes conjugués sans tenir compte du squelette σ.

La première théorie des orbitales moléculaires semi-empirique tri-dimensionnelle est **l'approximation au recouvrement différentiel nul (CNDO pour : Complete Neglect of Differential Overlap)**, introduite par Pople, Santry et Segal [47], pour être appliquée à tous les électrons de valence de molécules quelconques organiques ou minérales.

L'approximation utilisée dans CNDO, et dans de nombreuses approximations subséquentes, pour traiter des interactions électron-électron est connue comme :

- Approximation du champ moyen ;
- Théorie du champ auto-cohérent (SCF : Self Consistent Field)

et - Théorie de Hartree- Fock (HF).

De ces appellations, l'approximation du champ moyen est probablement la plus descriptive, mais c'est le terme SCF qui est le plus courant.

Comme le problème du calcul de l'énergie d'interaction électron- électron dans un système poly-électronique ne peut avoir de solution exacte, on doit utiliser des approximations. La théorie SCF traite chaque électron comme s'il interagissait (au cours du temps) avec le champ moyen de tous les autres électrons de la molécule. Ce qui signifie que les électrons restants de la molécule ne réagissent pas avec l'électron considéré dans sa position instantanée. Ainsi, le calcul de l'énergie de chaque électron individuellement devient un problème mono-électronique auquel nous avons à ajouter l'effet du champ causé par les électrons restants. Cette approximation néglige le fait que les mouvements des électrons sont corrélés de manière à réduire leurs répulsions mutuelles (c'est-à-dire que chaque électron réagit aux positions instantanées de tous les autres). Ainsi, la théorie SCF rend la tâche

computationnelle gérable au prix d'une surestimation de l'énergie de répulsion électron-électron.

Cependant, en 1965, les ressources computationnelles nécessaires pour l'approche SCF complète n'étaient pas encore disponibles. La pratique des théories des orbitales moléculaires nécessitaient donc encore des approximations. Le principal problème réside dans le calcul et le stockage des intégrales tétracentres notées $\langle \mu\nu|\lambda\sigma\rangle$, nécessaires pour le calcul des interactions électron-électron dans le cadre de l'approximation SCF. Les indices $\mu, \nu, \lambda \; et \; \sigma$ dénotent quatre centres d'orbitales atomiques de sorte que le nombre de telles orbitales à calculer croît proportionnellement à N^4, où N est le nombre d'orbitales atomiques. En fait, le nombre de telles intégrales n'est pas exactement égal à la puissance quatrième du nombre de fonctions de base parce que beaucoup d'entr'elles sont reliées par symétrie. Ce qui était une tâche très difficile en 1965 ; ainsi Pople, Santry et Segal ont introduit [47] l'approximation que seules les intégrales pour lesquelles $\mu = \nu$ et $\lambda = \sigma$ c'est-à-dire : $\langle \mu\mu|\nu\nu\rangle$ seront prises en compte et que, de plus, toutes les orbitales atomiques seront traitées de la même façon (comme si elles étaient des orbitales s), de sorte que l'équation (1.II.34) s'applique, où μ est centrée sur l'atome A et λ sur l'atome B et ainsi γ_{AB} ne dépend que des identités de A et B, et peut être traité comme paramètre.

$$\langle \mu\mu|\lambda\lambda\rangle = \gamma_{AB} \qquad (1.II.34)$$

Une première approximation, due à Pariser et Parr [44,45] consiste à traiter le terme mono-centre γ_{AA}comme différence entre le potentiel d'ionisation PI_A et l'affinité électronique AE_A de A [Eq.(1.II.35)] :

$$\gamma_{AA} = PI_A - AE_A \qquad (1.II.35)$$

Les termes di-centres sont alors données par l'éq.(1.II.36) :

$$\gamma_{AB} = \frac{\gamma_{AA}+\gamma_{BB}}{2+r_{AB}(\gamma_{AA}+\gamma_{BB})} \qquad (1.II.36)$$

Ce qui conduit à : $\gamma_{AB} = (\gamma_{AA} + \gamma_{BB})/2$ pour une distance interatomique, r_{AB}, nulle et $\gamma_{AB} \approx {}^{1}/_{r_{AB}}$ pour des distances interatomiques plus grandes. Ces expressions (Eqs. (1.II.34)–(1.II.36)) montrent la simplicité de la technique CNDO, qui a été utilisée pour calculer les propriétés électroniques comme les moments dipolaires ou les énergies d'excitation, généralement à partir des géométries expérimentales. Il ya eu beaucoup de modifications des eqs. (1.II.35) et (1.II.36), mais elles restent d'une simplicité comparable. Pareillement, des expressions simplifiées ont aussi été utilisées pour les intégrales mono-électroniques.

Cependant, la méthode CNDO montra des insuffisances systématiques directement imputées aux simplifications ébauchées précédemment, aussi fut-elle remplacée par la méthode **INDO (Intermediale Neglect of Differential Overlap)**, introduite en 1967 par Pople, Beveridge et Dobosh [48]. L'approximation qui conduit à l'éq. (1.II.34) s'étant avérée très sévère, elle fut remplacée par des valeurs individuelles pour les différents types d'interactions entre deux orbitales atomiques. Ces valeurs individuelles, souvent désignées par G_{ss}, G_{sp}, G_{pp} et G^2_{pp} dans la littérature, peuvent être ajustées pour donner un accord avec l'expérience meilleur que celui obtenu avec la méthode CNDO. Cependant, en INDO les termes di-centres sont maintenus du même type que ceux apparaissant dans les éqs. (1.II.35) et (1.II.36). Cette approximation conduit à des affaiblissements systématiques, comme par exemple dans le traitement des interactions entre doublets isolés.

Pour surmonter ces carences, Pople et collaborateurs revinrent à une approche plus complète que celle qu'ils proposèrent initialement en 1965 [47] : **l'approximation au recouvrement différentiel diatomique nul (NDDO : Neglect of Diatomic Differential Overlap)**.

Dans la NDDO, toutes les intégrales tétracentres insuffisances $\langle \mu\nu|\lambda\sigma\rangle$ dans lesquelles μ et ν sont sur le même centre, comme le sont λ et σ (mais pas nécessairement sur le même comme le sont μ et ν) sont prises en compte. De plus, les intégrales pour lesquelles les deux centres atomiques sont différents sont traitées de manière analogue que les intégrales mono-centres en INDO, entrainant, une amélioration de la description des interactions (doublet isolé)-(doublet isolé) par rapport aux méthodes précédentes. La NDDO forme la base de presque toutes les autres méthodes semi-empiriques qui, à quelques exceptions ont été développées par MJS Dewar et son école.

Les premières techniques semi-empiriques développées par Dewar et son groupe ont été désignées par MINDO/1-3 et ont été basées sur INDO. Beaucoup d'approximations d'intégrales de l'INDO originale ont été remplacées et les méthodes paramétrées pour reproduire un large intervalle de données expérimentales, particulièrement les énergies et les géométries.

Les méthodes MINDO sont maintenant largement obsolètes.

La méthode avantageuse pour la plupart des techniques modernes d'orbitales moléculaires semi-empiriques est la MNDO, qui a été publiée par Dewar et Thiel en 1977 [49]. La MNDO est une méthode NDDO dans laquelle Dewar et Thiel ont introduit un formalisme basé sur les multipoles pour le calcul des intégrales bi-électroniques. Elle a été paramétrée pour reproduire les chaleurs de formation expérimentales, les géométries, les moments dipolaires et les

potentiels d'ionisation. Elle s'avéra très supérieure aux méthodes MINDO pour la plupart des grandeurs calculées. Cependant la MNDO présente une faiblesse qui limite sévèrement son utilité ; elle ne reproduit pas la liaison hydrogène. Cette faiblesse a été surmontée de façon pragmatique par Burstein et Isaev [50] qui modifièrent simplement le potentiel de répulsion cœur-cœur par addition de fonctions gaussiennes en vue d'obtenir des liaisons hydrogène. Ce « fixe » a été adopté par le groupe Dewar pour leur méthode suivante AM1 [51] qui est par ailleurs identique à la MNDO. AM1, en retour, s'avéra présenter une faiblesse dans le traitement des composés nitrosés et hypervalents. Ces faiblesses ont été abordées par Stewart dans une nouvelle paramérisation nommée PM3 [52], qui est par ailleurs identiques à AM1. Cependant, MNDO, MNDO/H, AM1 et PM3 sont pour l'essentiel identiques du point de vue quanto-mécanique. Leurs différences se limitent à la « correction » classique des potentiels entre atomes et pour laquelle les paramètres sont traités comme variables dans la procédure de paramétrisation.

II-b-2-3- Limites et avantages des méthodes semi-empiriques [53]

La négligence de toutes les intégrales bi-électroniques tri et tétracentres réduit la matrice de Fock d'un ordre formel M^4 à M^2. Toutefois, le temps requis pour la diagonalisation de la matrice $\boldsymbol{F}$ croît comme le cube de la dimension de la matrice. La diagonalisation d'une matrice devient importante lorsque la dimension dépasse ~10 000 × 10 000. De nombreuses itérations sont nécessaires pour la résolution des équations SCF, et habituellement la géométrie est également optimisée, nécessitant de nombreux calculs pour différentes géométries. Ce qui situe la limite actuelle des méthodes semi-empiriques à environ 1000 atomes. Il est à noter que la méthode classique de résolution des équations HF par diagonalisation de la matrice de Fock s'impose rapidement comme l'étape limitante réelle dans les méthodes semi-empiriques. Des développements récents se sont ainsi focalisés sur la formulation de méthodes alternatives pour l'obtention d'orbitales SCF sans passer par la diagonalisation [54,55]. De telles méthodes utilisent des ajustements (combinaisons) linéaires avec le nombre d'atomes, ce qui permet d'effectuer des calculs pour des systèmes comprenant plusieurs milliers d'atomes.

La paramétrisation de MNDO/AM1/PM3 est réalisée en ajustant les constantes impliquées dans les différentes méthodes de façon à ce que les résultats des calculs HF ajustent les données expérimentales aussi près que possible. Ce qui est faux dans un sens. On sait que la méthode HF ne peut conduire au résultat correct, même à la limite d'un ensemble de base infini et sans approximations. Les résultats HF ne reproduisent pas la corrélation

électronique, mais les données expérimentales impliquent naturellement de tels effets. Ceci peut être considéré comme un avantage, les effets de corrélation électronique sont implicitement pris en compte dans la paramétrisation, et il n'est pas besoin d'exécuter des calculs compliqués pour surmonter les déficiences de la procédure HF. Cependant, il y a réellement problème quand la fonction d'onde HF ne peut décrire le système correctement, même qualitativement, comme par exemple avec les bi-radicaux et les états excités.

Une flexibilité additionnelle peut être introduite dans la fonction d'onde d'essai en ajoutant davantage de déterminants de Slater, par exemple par l'intermédiaire d'une procédure d'interaction de configuration (CI : pour configuration Interaction). Seulement la corrélation électronique est prise en compte deux fois, une première fois lors de la paramétrisation au niveau HF, et une seconde fois explicitement par le calcul CI.

Remarque : l'interaction de configuration CI résoud le problème de la corrélation électronique en considérant plus d'un schéma d'occupation des orbitales moléculaires (OM) et en combinant les micro-états obtenus par permutation des positions électroniques sur toutes les OM disponibles. Dans sa forme la plus simple, un calcul CI consiste en un calcul SCF préliminaire qui fournit les OM qui seront utilisées telles quelles tout au long du reste du traitement. Des micro-états sont alors construits en déplaçant les électrons des orbitales occupées à celles vacantes selon des schémas pré-établis. La matrice CI est alors calculée, ses éléments diagonaux représentent les énergies des micro-états et les éléments non diagonaux leurs interactions. Cette matrice est diagonalisée en vue d'obtenir les énergies des différents états (fondamental et excités) de la molécule comme combinaisons linéaires des micro-états. De nouveau les énergies sont fournies par les valeurs propres et les coefficients de la combinaison linéaire par les vecteurs propres. Cette procédure conduit à la stabilisation de l'état fondamental, et fournit également les énergies et les fonctions d'onde des états excités. Le problème est que si l'on doit considérer chacun des arrangements possibles de tous les électrons dans toutes les OM (CI complète), les calculs deviennent par trop importants même pour des molécules de taille moyenne avec un ensemble de base pas trop important (parce qu'il y a de trop nombreuses orbitales virtuelles).

Aussi, deux types de restrictions sont habituellement utilisés ; seul un nombre limité d'OM autour de l'intervalle des orbitales frontières (HOMO-LUMO) est inclus dans CI, et seuls certains types de réarrangements (excitations) des électrons sont utilisés.

La forme la plus économique est celle pour laquelle seuls les micro-états dans lesquels un électron est promu de l'état fondamental à une orbitale virtuelle (excitations simples) sont utilisées. Ce qu'on désigne, dans une forme abrégée, par CIS. En ajoutant toutes les

excitations doubles (pour lesquels deux électrons sont promus) on est conduit à CISD, et ainsi de suite (Figure II.2).

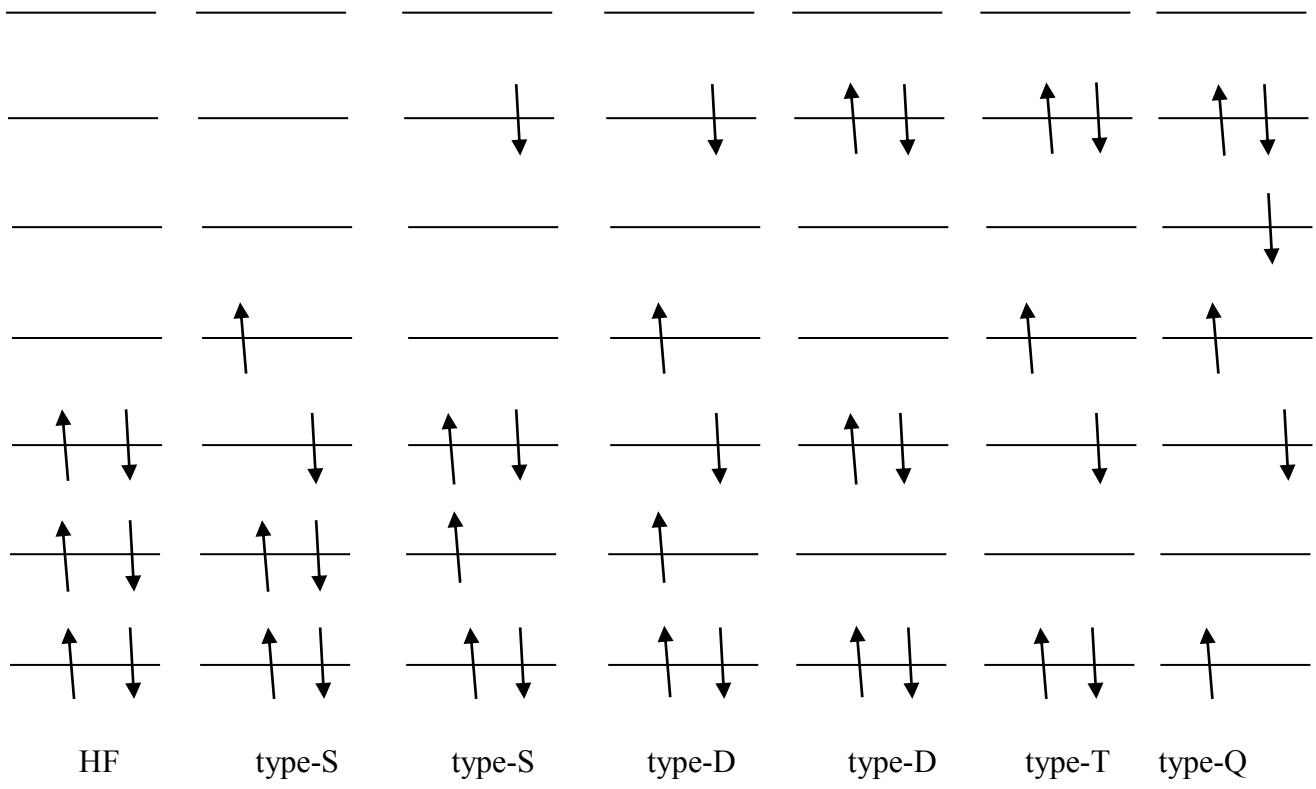


Figure II.2 : Déterminants de Slater excités générés à partir d'une référence HF

Les déterminants sont désignés par simples (S), doubles (D), Triples (T), quadruples (Q) etc…

La fonction d'onde avec interaction de configuration (ψ_{CI}) peut être représentée par l'équation suivante :

$$\psi_{CI} = a_0\phi_{SCF} + \sum_{Simples\ (S)} a_s\phi_s + \sum_{Doubles\ (D)} a_D\phi_D + \cdots = \sum_{i=0} a_i\phi_i \qquad (1.II.37)$$

La méthode des multiplicateurs indéterminés de Lagrange [56] est ensuite appliquée pour minimiser l'énergie :

$$E = (\langle\psi|\hat{H}|\psi\rangle/\langle\psi|\psi\rangle) \qquad (1.II.38)$$

Les méthodes semi-empiriques partagent les avantages/désavantages des méthodes de champ de force (cf : III), elles sont davantage performantes avec les systèmes pour lesquels on dispose de données expérimentales en quantités, mais il leur est impossible de faire des prédictions pour des types de composés totalement inconnus. La dépendance des données expérimentales n'est pas aussi sévère que pour la méthode du champ de force, à cause de la

forme complexe de la fonctionnelle du modèle. Les méthodes NDDO nécessitent uniquement des paramètres atomiques, et nullement des paramètres di-, tri- et tétra-atomiques comme dans les méthodes de champ de force. Une fois un atome donné paramétré, tous les types de composés possibles contenant cet élément peuvent être traités. Le plus petit nombre de paramètres et la forme plus complexe de la fonctionnelle ont l'inconvénient, par rapport aux méthodes de champ de force, qu'il est très difficile de « réparer » un problème spécifique par re-paramétrisation.

Les méthodes semi-empiriques sont de dimension nulle, tout comme les méthodes de champ de force. Il n'y a aucun moyen d'évaluer la fiabilité d'un résultat donné dans les limites de la méthode. Cela est dû à la sélection d'un ensemble de base fixe (minimum). La seule façon de juger les résultats est de comparer la précision d'autres calculs sur des systèmes similaires avec des données expérimentales.

Les méthodes semi-empiriques fournissent une méthode de calcul de la fonction d'onde électronique, qui peut être utilisée pour la prévision d'une variété de propriétés. Il n'y a rien qui entrave le calcul, par exemple, de la polarisabilité d'une molécule, bien qu'il soit connu des calculs *ab-initio* que l'obtention de bons résultats nécessite un grand ensemble de base polarisé incluant des fonctions diffuses. Les méthodes semi-empiriques comme AM1 ou PM3 n'ont qu'une base minimale (absence de polarisation et de fonctions diffuses), la corrélation électronique n'est qu'implicitement incluse par les paramètres et aucune donnée de polarisabilité n'a été utilisée pour dériver ces paramètres. Il est douteux que de tels calculs puissent conduire à des résultats comparables à ceux fournis par l'expérience, et ils nécessitent, pour le moins, un calibrage soigné [53]. Encore une fois, il convient de souligner que la capacité d'effectuer un calcul ne garantit pas la fiabilité des résultats obtenus.

II-b-3-Analyse des distributions de charges

Plutôt que de décrire la distribution électronique d'une molécule par des cartes d'isodensité, on préfère caractériser cette distribution, dans le voisinage d'un atome ou d'une liaison, par des nombres simples ou indices. Cette procédure, qui entraîne une perte d'information, est avantageuse dans les études comparatives.

La caractérisation d'une molécule par un tel ensemble d'indices est appelée son **analyse de population**.

Il existe une famille d'analyses de population, parmi lesquelles nous citerons celles de Coulson et Longuet-Higgins [57], exprimée en termes de charges (ou « densités de charge »)

et d'ordres de liaison, celle de Mulliken [58], que nous rappellerons brièvement, et qui fait intervenir les populations atomiques et de recouvrement.

II-b-3-1- Analyse de population de Mulliken

Mulliken introduit le concept important de **population de recouvrement**, c'est-à-dire de population électronique non localisée sur un atome mais répartie dans la liaison entre deux atomes. Ce concept permet une représentation très nuancée de la liaison chimique.

Dans l'analyse de population électronique qu'il propose, Mulliken définit les grandeurs :

$$P_v = \sum_k^{OM.occupées} N_k C_{kv}^* C_{kv} \qquad (1.II.39)$$

ou N_k est la population de l'O.M. ψ_k ; P_v est la population électronique localisée dans l'O.A. $\varphi_{\mu v}$, que l'on appelle la population nette de l'O.A. φ_v, dans la molécule.

$$R_{\mu v} = 2\sum_k^{OM.occupées} C_{k\mu}^* C_{kv} S_{\mu v} \qquad (1.II.40)$$

$R_{\mu v}$ est la population électronique localisée ni dans φ_μ, ni dans $\varphi_v$mais répartie entre ces deux O.A, que l'on appelle population de recouvrement entre les O.A φ_μe tφ_v.

En désignant par N le nombre total d'électrons, on a :

$$\sum_\mu R_{\mu v} = \sum_\mu \sum_v P_{\mu v} S_{\mu v} = N \quad \text{[Décomposition sur les OA]} \qquad (1.II.41)$$

$$\int \psi^* \psi \, d\tau = N \quad \text{[Décomposition sur les OM]} \qquad (1.II.42)$$

Posons :

$q_\mu = \sum_v P_{\mu v} S_{\mu v}$= Quantité d'électricité qui peut être attribuée à la $\mu^{ème}$orbitale atomique de base.

Alors, la quantité d'électricité qui peut être attribuée à l'atome M, dans la molécule, est la somme des $q_\mu(M)$ $(\mu \in M)$, soit :

$$Q_M = \sum_{\mu(M)} q_\mu (M) \qquad (1.II.43)$$

q_μ=densité électronique de l'orbitale μ ;

Q_M= densité électronique de l'atome M.

On peut ainsi déterminer la **charge (formelle) de l'atome M, dans la molécule, soit δ_M** :

$$\delta_M = Z_M - Q_M \qquad (1.II.44)$$

Z_M = nombre d'électrons de l'atome isolé ; Q_M = quantité d'électricité qu'il possède dans la molécule.

II-b-3-2-Calcul du moment dipolaire

Le moment dipolaire d'une molécule peut être décomposé, de façon unique, en trois composantes : une composante atomique ou d'hybridation, une composante de recouvrement,

et une composante de transfert de charge (qui permet de définir les charges atomiques nettes), chacune étant définie de façon univoque dans le cadre du schéma OM-CLOA.

Dans ce schéma, l'expression en u.a du moment dipolaire d'une molécule, dans la convention des chimistes, est [59]

$$\vec{\mu} = \sum_P \sum_Q \sum_{r\in P} \sum_{s\in Q} P_{rs}^{PQ} \int \varphi_r^* \vec{r}\, \varphi_s d_s d_r - \vec{\mu}_{nucl} \qquad (1.II.45)$$

Avec :

$$P_{rs}^{PQ} = \sum_i n_i C_{ir} C_{is} \qquad (1.II.46)$$

n_i = taux d'occupation de l'OM ψ_i , C_{ir} et C_{is}, coefficients des orbitales $\varphi_r$et $\varphi_s$appartenant respectivement, aux atomes *P* et *Q*, dans l'approximation CLOA des ψ_i.. Le vecteur position d'un électron en général et le vecteur position d'un atome *P* (mesurés en u. a par rapport à la même origine arbitraire) seront notée $\vec{\boldsymbol{r}}$ et $\vec{\boldsymbol{r}}_{\boldsymbol{p}}$, alors que *np* désignera le nombre d'électrons de l'atome *P* engagés dans la formation de la molécule.

On peut alors faire les substitutions suivantes :

$\vec{r}=\vec{r}_p+\vec{\xi}$, dans les termes tels que $\boldsymbol{P = Q}$ (1.II.47)

$\vec{r} = \frac{1}{2}\left(\vec{r}_p + \vec{r}_Q\right) + \vec{\chi}$, dans les termes tels que **P≠ Q**

Evidement $\vec{\xi}$ est le rayon vecteur qui a pour origine la position de l'atome *P*, $\vec{\chi}$ est le rayon vecteur dont l'origine coïncide avec le milieu du segment *PQ*. En tenant compte de l'orthogonalité des deux orbitales $\varphi_r$et $\varphi_{r'}$ centrées sur le même atome *P*, en appelant S_{rs}^{PQ} l'intégrale de recouvrement des orbitales centrées sur des atomes *P* et *Q* différents, et en posant :

$$\vec{\xi}_{rr'}^{P} = \int \varphi_r^* \vec{\xi}\, \varphi_{r'} d\tau \;\; ; \;\; \vec{\chi}_{rs}^{PQ} = \frac{\int \varphi_r^* \vec{\chi} \varphi_s d\tau}{S_{rs}^{PQ}} \qquad (1.II.48)$$

Le moment dipolaire (1.II.45) devient [58] :

$$\vec{\mu} = \sum_p \delta_p \vec{r}_p + \vec{\mu}_{hybrid} + \vec{\mu}_{recouvr} \qquad (1.II.49)$$

Avec :

$$\vec{\mu}_{hybrid} = \sum_p \sum_{r,r'\in P} P_{rr'}^{PP} \vec{\xi}_{rr'}^{P} \qquad (1.II.50)$$

Et :

$$\vec{\mu}_{recouvr} = \sum_p \sum_{r,r'\in P} \sum_Q \sum_{s\in Q} P_{rs}^{PQ} S_{rs}^{PQ} \vec{\chi}_{rs}^{PQ} \qquad (1.II.51)$$

II-b-3-3-Application

Nous avons réuni dans la figure 1.II.2 quelques applications [60] des indices électroniques de la méthode des orbitales moléculaires.

Sur la base des charges atomiques partielles on peut calculer des descripteurs électrostatiques simples qui peuvent servir pour le développement d'équations QSXR [Relations Quantitatives Structures –X ; où X= P (propriété) – A (activité) – R (rétention chromatographique) – T (toxicité)…].

-. Les charges partielles minimale (la plus négative) et maximale (la plus positive) dans la molécule (q_{min}, q_{max}).

-. Les charges partielles minimale et maximale pour les atomes particuliers (C, O etc…).

-. Un paramètre de polarité simple (q_{max}, q_{min}) ou pondéré par une fonction de la distance r_{max} entre les atomes portant les charges partielles minimale et maximale.

$$P_f = \frac{q_{max} - q_{min}}{F(r_{max})} \qquad (1.II.52)$$

II-c-LA MÉCANİQUE MOLÉCULAİRE

Si une molécule est trop grosse pour subir un traitement semi-empirique, il est toujours possible de modéliser son comportement en évitant complètement la mécanique quantique. Les méthodes désignées par mécanique moléculaire, établissent une expression algébrique simple de l'énergie d'un composé, sans avoir à calculer une fonction d'onde ou une densité électronique totale [61]. L'expression de l'énergie consiste en des équations classiques simples, comme l'équation de l'oscillateur harmonique, dans le but de décrire l'énergie associée à l'étirement de liaison, de flexion, de rotation, et aux forces intermoléculaires, telles que les interactions de Van der Waals et de liaison hydrogène. Toutes les constantes apparaissant dans ces équations doivent être obtenues à partir de données expérimentales ou d'un calcul *ab initio*.

Dans une méthode de mécanique moléculaire, la base de données des composés utilisés pour paramétrer la méthode (un ensemble de paramètres et de fonctions est appelé un champ de force) est cruciale pour son succès. La méthode de mécanique moléculaire peut être paramétrée à partir d'une classe spécifique de molécules, telles que des protéines, des molécules organiques, organo-métalliques, etc…

La mécanique moléculaire permet la modélisation de très grosses molécules, comme les protéines et des segments de DNA, la faisant le premier outil de la biochimie computationnelle. Le défaut de cette méthode est qu'il y a beaucoup de propriétés chimiques qui n'y sont pas définies, comme par exemple les états électroniques excités. De plus, pour

travailler avec des systèmes très grands et très compliquées, les logiciels doivent être très puissants et faciles dans l'utilisation des interfaces graphiques.

II-c-1- Pas de calculs de champ de force sans définition préalable des types d'atomes.

La géométrie de la molécule traitée (caractérisée par les coordonnées internes ou les coordonnées cartésiennes), le numéro atomique de chaque noyau, et l'état général de charge et de spin, constituent le nombre minimal d'entrées préalable à un calcul par mécanique moléculaires. Les informations concernant les distributions des électrons, en terme de densité électronique ou de fonction d'onde, ou les charges atomiques partielles, sont mieux interprétées sur la base de la géométrie moléculaire. Dans le contexte de la méthodologie du champ de force, l'entrée de la charge totale et du spin d'une molécule n'est pas obligatoire car ces types de calculs ne traitent pas des électrons. Pour représenter l'aspect électrostatique, il n'est même pas besoin des charges atomiques partielles si l'on utilise, par exemple, des dipôles de liaisons. Au contraire de la mécanique quantique, la mécanique moléculaire nécessite plus d'informations que le numéro atomique seul. En fait, chaque atome doit être décrit de manière plus détaillée.

Le concept de types d'atomes permet une différenciation en termes d'environnement local, d'état d'hybridation, ou de conditions spécifiques telles que la tension dans les systèmes comportant un petit anneau. Allinger et ses co-auteurs, qui ont développé les champs de force MM2, MM3, et MM4 pour les « petites molécules » [cf : III-3] ont défini dans la paramétrisation de MM3 plus de 15 types d'atomes différents pour le seul carbone. A savoir, alcanes sp^3, alcènes sp^2, cyclopropanes sp^2, carbonyles sp^2, alcynes sp etc…, tous nécessaires pour rendre MM3 applicable (ce qui signifie l'obtention de résultats raisonnables) pour un ensemble de molécules diverses. On peut constater immédiatement la difficulté de cette approche : le plus d'atomes définis, le plus de paramètres de contribution à la fonction énergie potentielle (liaisons, angles, dièdres…) doivent être développés. Des champs de force plus généraux affecteront donc, un seul type d'atome de carbone générique sp^2, sacrifiant en faveur d'une application générale. Une autre tendance consiste à utiliser pour les champs de force de classes spécifiques des types d'atomes plus importants en nombres, qu'on ne le ferait dans le cas de paramétrisations pour une application générale.

II-c-2- Forme fonctionnelle des champs de force courants

Un champ de force ne consiste pas uniquement en une expression mathématique qui décrit l'énergie d'une molécule en fonction des coordonnées atomiques. La deuxième partie

indispensable est le jeu de paramètres lui-même. Deux champs de force différents peuvent présenter la même forme fonctionnelle, mais utilisent un paramétrage complètement différent. D'un autre côté, différentes formes fonctionnelles peuvent conduire à des résultats presque identiques, en fonction des paramètres mis en jeu. Cette comparaison montre que les champs de force sont empiriques : il n'y a pas de forme « correcte ».

Parce que certaines formes fonctionnelles donnent de meilleurs résultats que d'autres, la plupart des implémentations dans les logiciels disponibles (académiques et commerciaux) sont très similaires.

Une hypothèse importante est que le champ de force déterminé à partir d'un ensemble de molécules est transférable à d'autres molécules.

Notons qu'un ensemble de paramètres développés et testés sur un nombre relativement petit de cas est encore applicable à une plus large gamme de problèmes. En outre, les paramètres développés à partir de données relatives à de petites molécules peuvent être utilisés pour étudier des molécules plus grandes telles que les polymères.

II-c-3- Quelques exemples

Parmi les champs disponibles nous citerons les plus répandus largement utilisés pour le traitement de petites molécules.

-MM2, MM3, et MM4 : (http://europa. chem. uga. edu/allinger/mm2mm3.html).
Introduit par Allinger *et al*.[62-65], largement utilisé pour le traitement de petites molécules.

-AMBER : (Assisted Method Building and Energy Refinement) (http://amber. scripps. edu)
Introduit par Cornell *et al.* [66] très largement utilisé dans le traitement des protéines et des acides nucléiques.

-CHARMM : (Chemistry at Harvard molecular Modeling) (http://yuri. harvard. edu)
Développé par Mackerall, Karplus *et al.,* [67-69] qui est largement utilisé pour la simulation de petites molécules et jusqu'aux complexes solvatés de grandes macromolécules biologiques. CHARMm est une version commerciale disponible de CHARMM qui est également applicable aux petits composés organiques [70].

-MMFF : (MerckMolecular Force Field)
Développé par Halgren [71,72], il est similaire à MM3 dans la forme, mais en diffère par son application focalisée sur les processus de condensation de phases (en dynamique

moléculaire). Il reproduit l'exactitude MM3 pour les petites molécules, et est applicable aux protéines et autres systèmes d'importance biologique.

Les différents champs de force se distinguent par trois aspects principaux :

1) La forme de la fonction de chaque terme énergétique.
2) Le nombre de termes croisés (qui reflètent le couplage entre coordonnées internes) inclus.
3) Le type d'information utilisé pour ajuster les paramètres.

II-c-4- Représentation simple d'un champ de force

Beaucoup de champs de forces utilisés actuellement pour la modélisation moléculaire peuvent s'interpréter en termes d'une représentation relativement simple à quatre composantes des forces intra et intermoléculaires internes au système considéré. Les pénalités énergétiques sont associées aux écarts des liaisons et des angles par apport à leurs valeurs de 'référence' ou 'd'équilibre', il y a une fonction qui décrit la façon dont l'énergie change lors de la rotation des liaisons, et finalement le champ de force contient des termes décrivant l'interaction entre des parties non liées du système. Des champs de forces plus sophistiqués peuvent comporter des termes additionnels, mais ils présentent invariablement ces quatre composantes. Un fait intéressant lié à cette représentation est qu'on peut imputer les variations à des coordonnées internes spécifiques telles que les longueurs et les angles de liaisons, la rotation des liaisons ou les mouvements des atomes relativement les uns aux autres. Ce qui permet de comprendre facilement comment les changements des paramètres du champ de force affectent ses performances, et aident également dans le processus de paramètrisation.

Pour des molécules seules ou des ensembles d'atomes et/ ou de molécules, un tel champ de force a la forme fonctionnelle suivante :

$$\upsilon(r^N) = \sum_{liaisons} \frac{k_i}{2}(I_i - I_{i,0})^2 + \sum_{angles} \frac{k_i}{2}(\theta_i - \theta_{i,0})^2 \sum_{torsions} \frac{V_n}{2}(1 - \cos(n\omega - \gamma)) + \sum_{i=1}^{N} \sum_{j=i+1}^{N} \left(4\varepsilon_{ij} \left[\left(\frac{\sigma_{ij}}{r_{ij}} \right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}} \right)^{6} \right] + \frac{q_i q_j}{4\pi\varepsilon_0 r_{ij}} \right) \quad (1.II.53)$$

$\upsilon(\mathrm{r}^{\mathrm{N}})$ représente l'énergie potentielle qui est fonction des positions (**r**) des N particules (habituellement les atomes). Les diverses contributions sont représentées schématiquement sur la figure suivante :

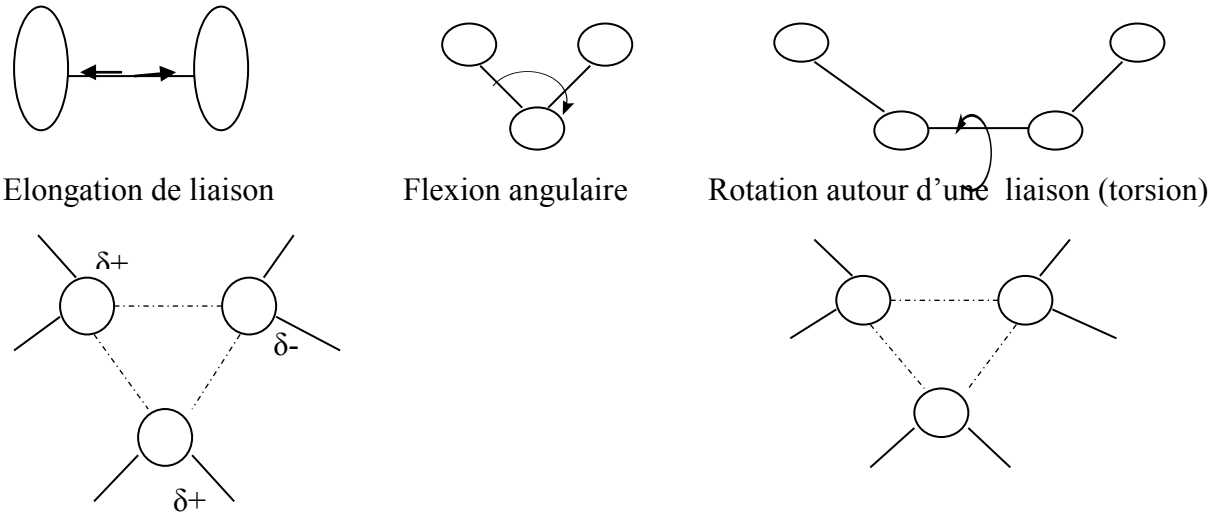


Figure II.3: Représentation schématique des quatre contributions d'un champ de force de MM : élongation de liaison, flexion angulaire, termes de torsion et interactions non liées.

Le premier terme de l'équation (1.II.53) modélise l'interaction entre paires d'atomes liés, représentée dans ce cas par un potentiel harmonique qui fournit la variation de l'énergie lorsque la longueur de liaison l_i dévie de sa valeur de référence (à l'équilibre) $l_{i,0}$. Le second terme est une sommation sur tous les angles de valence de la molécule, encore modélisé par un potentiel harmonique (un angle de valence est l'angle formé par trois atomes A- B- C où A et C sont toux deux liés à B). Le troisième terme dans l'équation (1.II.53) renseigne sur la variation d'énergie lorsqu'une liaison tourne. La quatrième contribution est le terme de non liaison, qui est calculé entre toutes les paires d'atomes (i et j) localisés sur différentes molécules ou appartenant à la même molécule mais séparés par au moins 3 liaisons (c'est-à-dire avec une relation l, n où $n \geq 4$). Dans un champ de force simple le terme de non liaison comprend un terme de potentiel coulombien pour les interactions électrostatiques et le potentiel de Lennard-Jones pour les interactions de Van der Waals.

II-c-5-Champ de force MM2 et MM+ [73]

II-c-5-1-Champ de force MM2

* Elongation des liaisons : MM2 a été paramétré pour ajuster les distances moyennes calculées à partir du mouvement vibrationnel à température ambiante à celles obtenues par diffraction électronique.

Pour mieux reproduire la courbe de Morse, MM2 fait intervenir un terme quadratique et un autre cubique :

$$v(l) = \frac{k}{2}(l - l_0)^2 \left[1 - k'(l - l_0)\right]^2 \qquad (1.II.54)$$

* Variation des angles : les déviations des angles de leurs valeurs de références sont souvent exprimées en utilisant la loi de Hooke ou potentiel harmonique :

$$v(\theta) = \frac{k}{2}(\theta - \theta_0)^2 \qquad (1.II.55)$$

Comme pour les termes d'élongation des liaisons, la précision du champ de force peut être augmentée par l'incorporation de termes d'ordres supérieurs. MM2 contient un terme d'ordre 4 en plus du terme quadratique.

$$v(\theta) = \frac{k}{2}(\theta - \theta_0)^2 \left[1 - k'(\theta - \theta_0)\right]^2 \qquad (1.II.56)$$

* Torsion des angles dièdres : correspond à la rotation d'une liaison selon l'angle dièdre ω formé par 4 atomes. Le champ de force MM2 utilise 3 termes d'une série de Fourier :

$$v(\omega) = \frac{v_1}{2}(1 + \cos\omega) + \frac{v_2}{2}(1 - \cos 2\omega) + \frac{v_3}{2}(1 + \cos 3\omega) \qquad (1.II.57)$$

Une interprétation physique a été attribuée à chacun des 3 termes à partir de l'analyse de calculs *ab initio* effectués sur des hydrocarbures fluorés simples.

* Angle dièdre impropre ou déviation extra- planaire. Examinons comment la cyclobutanone serait modélisée en utilisant uniquement un champ de force comportant des termes standards pour l'élongation des liaisons et les variations angulaires du type de ceux apparaissant dans l'équation (1.II.57). La structure d'équilibre obtenue avec un tel champ de force sera caractérisée par la localisation de l'atome d'oxygène hors du plan formé par l'atome de carbone contigu (à l'oxygène) et les 2 carbones qui lui sont liés (Figure II.4).

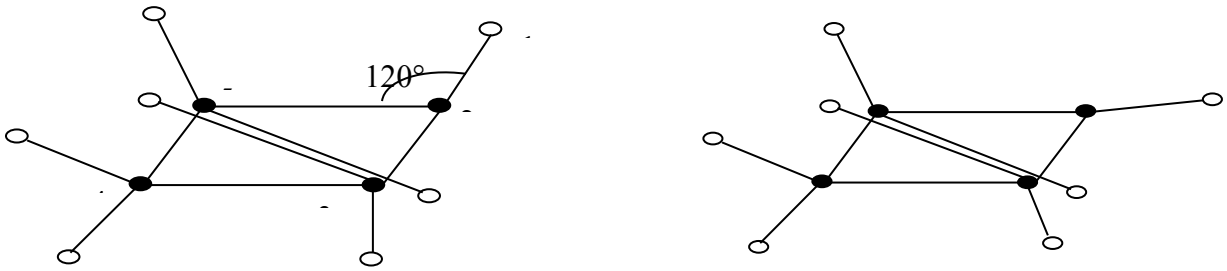


Figure II.4: Sous un terme extra-planaire, l'atome d'oxygène de la cyclobutanone est prévu de se situer hors du plan du cycle/gauche plutôt que dans le plan.

Dans cette configuration, les angles vers l'oxygène adoptent des valeurs proches de la valeur de référence 120°. Expérimentalement, il est établi que l'atome d'oxygène demeure dans le plan du cyclobutane bien que les angles C-C=O soient grands (133°). Ceci parce que l'énergie de liaison π, qui est maximalisée dans l'arrangement coplanaire, sera plus réduite si l'atome d'oxygène est dévié hors du plan. Pour réaliser la géométrie désirée il est nécessaire d'ajouter un (ou des) terme (s) additionnel (s) dans le champ de force qui maintienne (nt) le carbone sp^2 et les 3 atomes qui lui sont liés dans le même plan. La plus simple façon de le réaliser est d'utiliser un terme de variation angulaire extra- planaire.

Il y a plusieurs façons d'incorporer des termes de variation extra- planaire dans un champ de force. Une approche possible est de traiter les 4 atomes comme un angle de torsion impropre pour lequel les 4 atomes (Figure II.5) ne sont pas liés dans la séquence 1- 2- 3- 4. Une façon de définir, dans ce cas, une torsion impropre consiste à impliquer les atomes 1- 5- 3- 2 de la figure.

Un potentiel de torsion de la forme suivante :

$$v(\omega) = k(1 - \cos 2\omega) \qquad (1.II.58)$$

peut être utilisé pour maintenir l'angle de rotation impropre à 0° ou 180°.

Il existe diverses autres façons pour inclure la contribution de la variation d'angle extra-planaire. Par exemple, une définition qui est la plus proche de la notion de « variation extra-planaire » implique le calcul de l'angle entre une liaison à partir de l'atome central et le plan défini par l'atome central et les 2 autres atomes (Figure II.5). La valeur 0° correspond aux 4 atomes coplanaires. Une troisième approche consiste à calculer la hauteur de l'atome central au- dessus du plan défini par les 3 autres atomes (Figure II.5). Avec ces deux définitions la déviation de la coordonnée extra- planaire (angle ou distance) peut être modélisée à l'aide d'un potentiel harmonique de la forme :

$$v(\theta) = \frac{k}{2}\theta^2 \quad ; \quad v(h) = \frac{k}{2}h^2 \qquad (1.II.59)$$

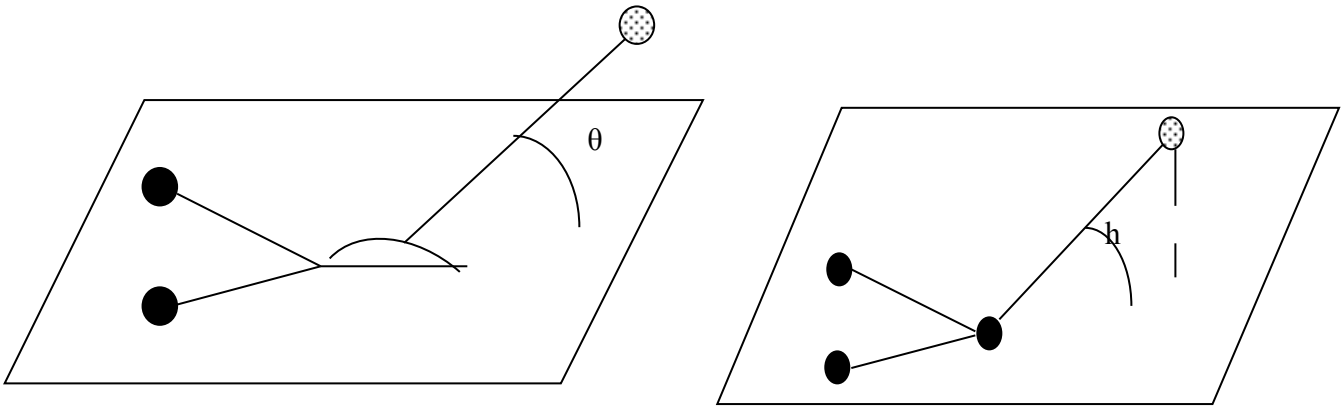


Figure II.5: Deux façons pour modéliser les contributions de la variation d'angle extra-planaire.

* Termes de croisement : Les termes de croisement permettent de tenir compte des interactions entre l'élongation des liaisons, la variation des angles et la torsion des angles dièdres. Par exemple, si les liaisons C-O et O-H de l'angle C-O-H sont étirées, alors la distance entre les atomes terminaux (C et H) est augmentée, ce qui rend plus facile la diminution de l'angle COH. Pareillement, la diminution de l'angle COH tend a être accompagnée d'une augmentation des longueurs des liaisons O-H et C-O. Pour tenir compte de cette interaction, on peut ajouter un terme de croisement « élongation-variation angulaire ». (stretch- bend) de la forme :

$$v_{\Delta\theta} = \frac{1}{2} k_{12}(\Delta l_1 + \Delta l_2)\,\Delta\theta \qquad (1.II.60)$$

avec $\Delta l_1 = l_1 - l_{10}$; $\Delta l_2 = l_2 - l_{20}$ et $\Delta\theta = \theta - \theta_0$

l_{10}, l_{20} et θ_0 représentent les valeurs de références pour l_1, l_2 et θ respectivement.

Les termes de croisement les plus utilisés sont (Figure II.6) :

* élongation- élongation et élongation- variation angulaire, pour deux liaisons à un même atome ;

* élongation- torsion angle dièdre, variation angulaire- torsion angle dièdre et variation angulaire- variation angulaire pour 2 angles avec un atome central commun.

Le champ de force MM2 fait intervenir uniquement un terme de croisement élongation-variation angulaire.

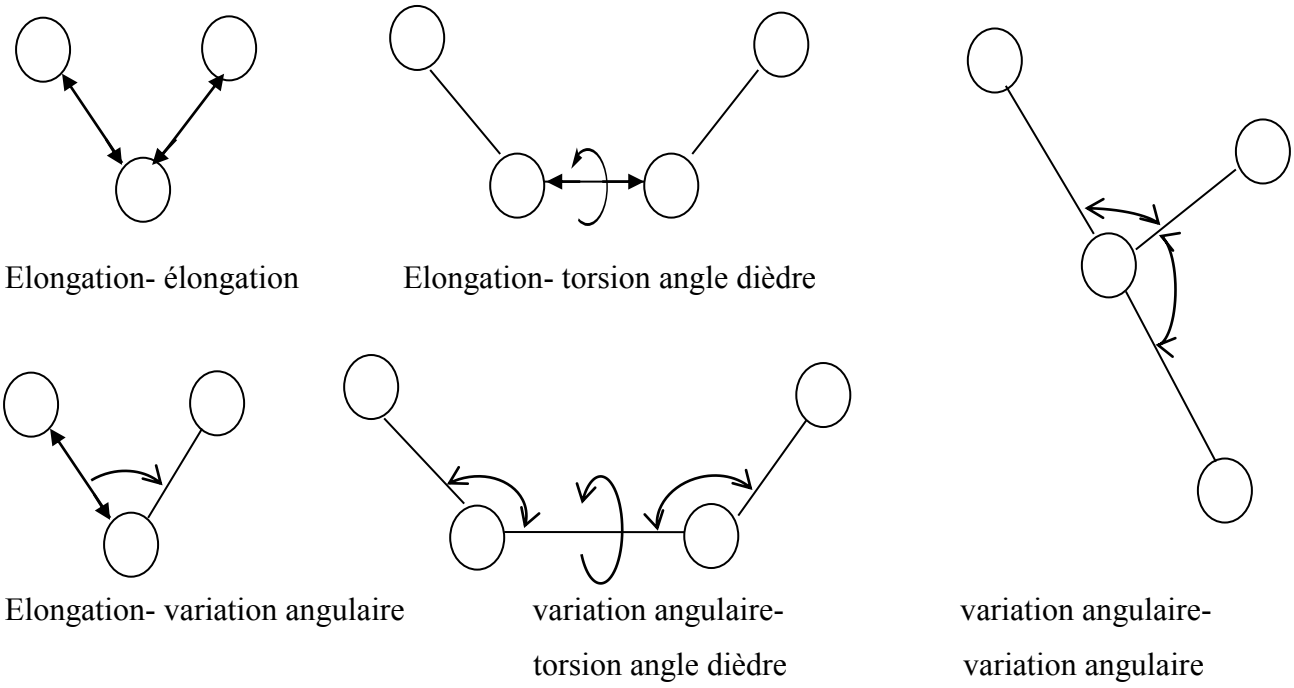


Figure II.6: Illustration schématique des termes de croisement supposés les plus importants pour les champs de forces.

* Interactions électrostatiques : Le terme électrostatique v_{es} est généralement défini par la somme des interactions électrostatiques entre toutes les paires d'atomes exceptées les paires 1, 2 et 1, 3 : $v_{es} = \sum l \geq 4\ v_{es,ij}$, où les atomes i, j vérifient la relation ($l \geq 4$).

v_{es} est généralement calculé en affectant des charges atomiques partielles à chaque atome et en appliquant un potentiel coulombien.

MM2 n'utilise pas cette procédure mais associe un moment dipolaire avec chaque liaison puis calcule v_{es} comme somme des énergies potentielles d'interactions entre moments de liaisons. Chaque moment dipolaire étant localisé au centre, et dirigé le long de cette liaison. Les valeurs des moments de liaisons de certains types de liaisons ont été choisies pour ajuster les moments dipolaires expérimentaux de petites molécules.

L'équation (1.II.61) décrit un modèle de charge, basé essentiellement sur les dipôles centrés, utilisé dans MM2 [62].

$$v_{es} = \frac{\mu_i \mu_j}{kr^3}(\cos \chi - 3 \cos \alpha_i \cos \alpha_j) \qquad (1.II.61)$$

χ et α_i, α_j désignent respectivement les angles entre les dipôles et les angles entre chaque dipôle et le vecteur de connexion.

* Interactions de Van der Waals : la plupart des champs de forces utilisent le potentiel 12- 6 de Lennard Jones.

MM2 utilise le potentiel de Buckingham, avec un terme attractif proportionnel à r^{-6} et un terme répulsif proportionnel à $e^{-\alpha r}$ où α est un paramètre :

$$v_{vdw} = A\, e^{-\alpha r} - \frac{B}{r^6} \qquad (1.II.62)$$

* Liaisons conjuguées : MM2 utilise une procédure générale qui consiste à effectuer des calculs semi- empiriques sur les électrons π pour en tirer les ordres de liaisons, qui sont ensuite utilisés pour affecter des longueurs de liaisons et des constantes de force de référence pour les liaisons conjuguées.

II-c-5-2-Champ de force MM+

C'est une extension du Champ de force MM2, avec l'ajout de quelques paramètres additionnels. 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.

Le tableau suivant [56] compare les trois techniques computationnelles majeures évoquées.

Tableau II.7: Etude comparative des techniques *ab initio*, semi- empirique et mécanique moléculaire.

ab initio	**Semi- empirique**	**Mécanique moléculaire**
-Prise en compte de tous les électrons. -Limité à quelques dizaines d'atomes. Nécessite un super calculateur -Peut être appliquée à des composés inorganiques, organométalliques,	-Ignore certains électrons (simplification). -Limité à quelques centaines d'atomes. -Peut être appliquée à des composés sinorganiques, organiques, organométalliques et de petits oligomères.	-Ignore tous les électrons. Seuls les noyaux sont considérés. -Molécules contenant des milliers d'atomes -Peut être appliquée aux

Tableau II.7: Etude comparative des techniques *ab initio*, semi- empirique et mécanique moléculaire.(Suite)

ab initio	**Semi- empirique**	**Mécanique moléculaire**
et aux fragments moléculaires (composants catalytiques d'enzymes). -Vide, solvatation implicite. -Applicable à l'état fondamental, et aux états de transition et excité.	(peptides, nucléotides, saccharides). -Vide, solvatation implicite. -Applicable à l'état fondamental, et aux états de transition et excité.	composés inorganiques, organiques, oligonucléotides, peptides, saccharides, métallo-organiques et inorganiques. -Vide, solvatation implicite ou explicite. -Applicable uniquement à l'état fondamental.

II-d-LA DYNAMIQUE MOLECULAIRE

La dynamique moléculaire a débuté avec l'arrivée, en 1957, des premiers ordinateurs [74]. Mais les premières simulations réelles ont été faites par Rahman [75], grâce à ses travaux sur la simulation de l'argon liquide, en 1964, avec un temps de simulation de 10^{-11} s, puis de l'eau liquide [76] en 1971.

II-d-1- Principe de la dynamique moléculaire

Chaque atome de la molécule est considéré comme une masse ponctuelle obéissant à la loi d'action de masse et dont le mouvement est déterminé par l'ensemble des forces exercées sur lui par les autres atomes en fonction du temps.

$$\vec{F}_i = m_i \vec{a}_i = m_i \frac{d^2 \vec{r}_i(t)}{dt^2} \qquad (1.II.63)$$

$\vec{F}_i$: vecteurs force agissant sur l'atome i.

m_i : masse de l'atome i.

$\vec{a}_i$: vecteur accélération de l'atome i.

$\vec{r}_i$: position de l'atome i.

Grace aux vitesses et aux positions de chaque atome au cours du temps, il est possible d'évaluer les données macroscopiques, comme l'énergie cinétique et la température. L'énergie cinétique est fournie par la relation :

$$E_c = \sum_{i=1}^{N} \frac{|\vec{P}_i|}{2n_i} \quad (1.II.64)$$

où $\vec{P}_i$ est la quantité de mouvement de l'atome i.

La température s'obtient à partir de l'énergie cinétique en exploitant la relation :

$$E_c = \frac{3K_bT}{2}(3N - N_c) \quad (1.II.65)$$

où: K_b désigne la constante de Boltzmann ; N_c le nombre de contraintes, et $(3N\text{-}N_c)$ le nombre total de degrés de liberté.

La force $\vec{F}_i$ qui s'exerce sur un atome i, en position$\vec{r}_i(t)$, est déterminée par dérivation de la fonction potentielle :

$$\vec{F}_i = \frac{d\vec{E}(r_i \ldots r_n)}{dr_i(t)} \quad (1.II.66)$$

E : fonction de l'énergie potentielle d'interaction totale.

r_i : coordonnées cartésiennes de l'atome i.

Les vitesses de chaque atome sont calculées à partir de la connaissance des accélérations atomiques.

$$\vec{a}_i = \frac{d\vec{V}_i}{dt} \quad (1.II.67)$$

Et les positions des atomes sont déterminées à partir des vitesses atomiques par la relation :

$$\vec{V}_i = \frac{d\vec{r}_i}{dt} \quad (1.II.68)$$

L'intégration de ces équations se fait en subdivisant la trajectoire en une série d'états séparés par des intervalles de temps très courts dont la longueur définit le pas d'intégration t, ce qui conduit à une trajectoire en fonction du temps. Connaissant la vitesse et l'accélération de l'atome i à l'instant t, on peut connaître sa position à l'instant t+Δt :

$$r_i(t + \Delta t) = r_i(t) + v_i\Delta t + \frac{1}{2}a_i\Delta t^2 \quad (1.II.69)$$

II-d-2- Application de la dynamique moléculaire

Une application importante de la dynamique moléculaire est l'analyse des modes normaux de vibration de long de la trajectoire. Une autre application est l'optimisation et le raffinement des structures 3D d'après les données de la cristallographie et/ou de la RMN. La mise en œuvre de cette méthode requiert néanmoins des moyens de calcul particulièrement

puissants et elle est coûteuse en temps et en argent. Elle se généralise cependant pour les études de peptides et de petites protéines [77].

II-e-LES ÉTUDES QSAR/QSPR

Les méthodes « QSAR (Quantitative Structure- Activity Relationships) /QSPR (Quantitative Structure- Property Relationships) » relient l'activité biologique / ou une propriété à la structure d'une molécule en proposant une relation mathématique plus ou moins complexe. Ces méthodes [78] reposent sur le principe selon lequel: des molécules ayant les mêmes propriétés sont proches dans l'espace chimique. L'espace chimique est un espace à n dimensions qui correspondent à des variables décrivant la structure moléculaire. Ce sont ces variables, appelées descripteurs, qui vont être reliées à la propriété étudiée par l'intermédiaire de différentes méthodes statistiques.

II-e-1-Les descripteurs moléculaires : Que sont-ils ?

II-e-1-1-Définition : Un descripteur moléculaire est le résultat final d'une procédure logique ou mathématique transformant l'information chimique encodée dans une représentation de la molécule en un nombre. Des milliers de descripteurs moléculaires (~ 10 000) ont été développés au cours des dernières décennies. Quatre familles de descripteurs constitutionnels, topologiques, géométriques et quantiques, impliquant différents niveaux de complexité.

Ainsi, en partant de la structure, des descripteurs sont calculés puis utilisés pour développer un modèle permettant de prédire la propriété, Y, ciblée :

$$Y = f(structure) = f(descripteurs) \quad (1.II.70)$$

II-e-1-2- Caractéristiques d'un descripteur idéal.

Pour permettre la construction d'un modèle QSAR/ QSPR fiable un descripteur idéal devrait présenter les caractéristiques suivantes [79] :

*- Etre pertinent pour une large catégorie de composés.

*- Etre corrélé avec les réponses étudiées tout en présentant une corrélation insignifiante avec d'autres descripteurs.

*- Le calcul du descripteur doit être rapide et indépendant des propriétés expérimentales.

*- Un descripteur doit fournir des valeurs différentes pour des molécules structurellement dissemblables, même si ces différences sont faibles.

*- Un descripteur doit posséder une interprétation physique pour pouvoir déterminer les caractéristiques requises pour les composés étudiés.

II-e-2-Les types de descripteurs

Les descripteurs peuvent être de différents types selon leur mode de calcul ou de détermination : physicochimique (hydrophobe, stérique ou électronique), structural (fréquence d'occurrence d'une sous-structure), topologique, électronique (calcul par orbitale moléculaire), géométrique (calcul de l'aire de la surface moléculaire) ou simples paramètres indicateurs [80,81]. Les « constantes de substitution » sont fondamentalement des descripteurs physicochimiques conçus sur la base de facteurs qui régissent les propriétés physicochimiques des entités chimiques. L'ensemble des descripteurs moléculaires sont des développements de l'approche des constantes de substitution, quoique beaucoup d'entre eux sont aussi déduits à partir d'approches expérimentales.

Les descripteurs peuvent également être classés selon les dimensions. Le tableau II.8 donne une illustration utile des descripteurs moléculaires couramment basés sur les dimensions.

Tableau II.8 : Différents descripteurs, employés dans les études QSAR, basés sur la dimension.

Dimension des descripteurs	Paramètres
Descripteurs 0D	Indices constitutionnels, propriété moléculaires, dénombrement d'atomes et de liaisons.
Descripteurs 1D	Nombre de fragments, empreintes digitales.
Descripteurs 2D	Paramètres topologiques, paramètres structuraux, paramètres physico-chimiques incluant des descripteurs thermodynamiques

Tableau II.8 : Différents descripteurs, employés dans les études QSAR, basés sur la dimension. (Suite)

Dimension des descripteurs	Paramètres
Descripteurs 3D	Paramètres électroniques, paramètres spatiaux, paramètres d'analyse de forme moléculaire, paramètres d'analyse de champ moléculaire et paramètres d'analyse de surface de récepteur

Remarque :

La dimension a été limitée à des descripteurs 0D-3D, bien que des descripteurs de dimensions plus élevées soient également disponibles [79].

II-e-3-Analyse des descripteurs

En règle générale, tous les descripteurs calculés ne peuvent être utilisés directement dans la construction du modèle, pour trois raisons principales : 1/ les différents éléments du jeu de descripteurs peuvent s'intercorréler, c'est-à-dire codent fondamentalement le même aspect structural ; 2/ les descripteurs peuvent coder les entités qui ne contribuent pas du tout à la propriété ; 3/ la taille globale d'un ensemble de descripteurs peut être si grande qu'elle devient ingérable. Chaque cas nécessite un prétraitement de l'ensemble des descripteurs défini de sorte que l'information essentielle soit extraite dans un ensemble réduit de descripteurs, avec une densité d'information, liée à la propriété cible, plus élevée. Deux paramètres statistiques sont principalement utilisés pour juger de la qualité des descripteurs. Le premier est une mesure de la variation d'un descripteur pour l'ensemble des données. Une faible variance caractérise le peu d'information contenue dans le descripteur. Le second est une mesure de redondance interne. Les descripteurs totalement indépendants présentent un coefficient de corrélation égal à 0,0 : ils sont dits orthogonaux. Le cas idéal est pratiquement inexistant, et le coefficient de corrélation entre deux descripteurs ne devrait pas dépasser 0,6, quoique des coefficients de corrélation entre descripteurs, compris entre 0,4 et 0,9, aient été jugés acceptables dans la littérature.

II-f-Relations quantitatives structures activités/propriétés (QSAR/QSPR)

II-f-1-Introduction

Le développement de nouvelles techniques de modélisation a permis la mise en place de nombreuses méthodes QSPR (en anglais QSPR : *Quantitative Structure Property Relationships*) et QSAR (en anglais QSAR : *Quantitative Structure-Activity Relationships*) ; elles reposent pour la plupart sur « la recherche d'une relation entre un ensemble de nombres réels, appelés descripteurs moléculaires, et la propriété ou l'activité que l'on souhaite prédire».

Ces méthodes permettent de confirmer les données expérimentales disponibles et de prédire les propriétés/activités pour de nouveaux composés ou des composés pour lesquels les données expérimentales ne sont pas disponibles.

II-f-2-Historique

Il y a plus d'un siècle et demi, en 1863, *Cros* [82] a observé que le point d'ébullition et le point de fusion des alcanes augmentent avec le nombre d'atomes de carbone et la masse moléculaire. Il a observé également une diminution de la solubilité dans l'eau des alcools avec l'augmentation du nombre d'atomes de carbone et la masse moléculaire, ce qui, depuis, est considéré comme la première formulation générale en QSPR.

Cinq ans après, en 1868, *Crum-Brown* et *Fraser* [83] postulèrent que « l'activité biologique d'une molécule est une fonction de sa constitution chimique ».

Quelques décennies plus tard, en 1893, *Richet* [84] a montré que la cytotoxicité de certains composés organiques était inversement proportionnelle à leur solubilité dans l'eau.

A la fin du 19ème siècle, *Meyer* en 1899 et *Overton* en 1901 [85-87], ont indépendamment observé « une relation linéaire entre l'activité des narcotiques et leur coefficient de partage huile-eau ».

Six ans après, en 1907, *Fühner* et *Neubauer* [88] ont montré pour une série de narcotiques homologues, que l'activité augmentait en fonction de la progression géométrique de la série de composés, ceci montrant l'importance de la contribution d'additivité de groupements fonctionnels pour l'activité biologique.

En 1962, *Hansen* [89] a montré l'existence d'une corrélation entre la toxicité des acides benzoïques substitués et les constantes électroniques « σ » des substituants.

L'année 1964 est considérée comme le début des méthodes QSAR modernes. *Hansch* et *Fujita* ont établi les premières corrélations entre les propriétés physico-chimiques (log P,

pKa, paramètres stériques et électroniques) et l'activité biologique (activité enzymatique, pharmacologique), Ces méthodes seront appelées par la suite "analyse de *Hansch"* et "analyse de *Free Wilson"* [90-91]). Sept ans plus tard, *Hansch* et *Lien* ont réalisé une étude QSAR sur différentes familles d'antifongiques : benzoquinones, sels d'alkylpyridinium, imidazoles et phénols. Ils ont observé que quels que soient la famille et le champignon utilisés, l'activité antifongique dépend du coefficient de partage Eau-Octanol, expérimental ou calculé [92].

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 des relations quantitatives structure / temps de rétention noté QSRR [93]**.**

Actuellement, des méthodes 3D comme l'étude CoMFA (*Comparative Molecular Field Analysis*) et CoMSIA (*Comparative Molecular Similarity Indices Analysis*) [94,95] permettent de traiter les relations structure-activité en trois dimensions, 3D-QSAR/QSPR.

II-f-3- Définition

Les méthodes QSAR/QSPR sont basées sur l'hypothèse que l'activité ou la propriété d'un composé chimique est liée à sa structure, plus précisément cette approche affirme que l'activité (ou la propriété) et la structure d'un composé chimique sont liées par un certain algorithme mathématique, cela est basé sur le postulat de base selon lequel ″les composés chimiques similaires ont des activités similaires″. De plus, lorsque les paramètres moléculaires sont exprimés par des chiffres, on peut proposer une relation mathématique, ou relation quantitative structure activité/propriété, entre les composés et leurs activités ou propriétés.

Par définition, une QSAR/QSPR est un modèle mathématique qui associe un ou plusieurs paramètres quantitatifs dérivés de la structure chimique, à une mesure quantitative d'une propriété ou d'une activité.

II-f-4- Principe

Le principe d'une étude QSAR/QSPR (Figure II.7), consiste à trouver une relation mathématique reliant de manière quantitative une activité biologique, ou une propriété, mesurée pour une série de composés similaires dans les mêmes conditions expérimentales, avec des descripteurs moléculaires à l'aide des méthodes statistiques. L'objectif de ces études est d'analyser les données structurales afin de détecter les facteurs déterminants pour l'activité ou la propriété étudiée. Pour ce faire, différents types de méthodes statistiques peuvent être employées (voir plus loin : les méthodes statistiques).

L'expression mathématique obtenue peut alors être utilisée comme moyen prédictif de l'activité/propriété étudiée pour de nouvelles molécules ou des molécules pour lesquelles les données expérimentales ne sont pas disponibles.

Ceci peut être traduit par le diagramme de la page suivante :

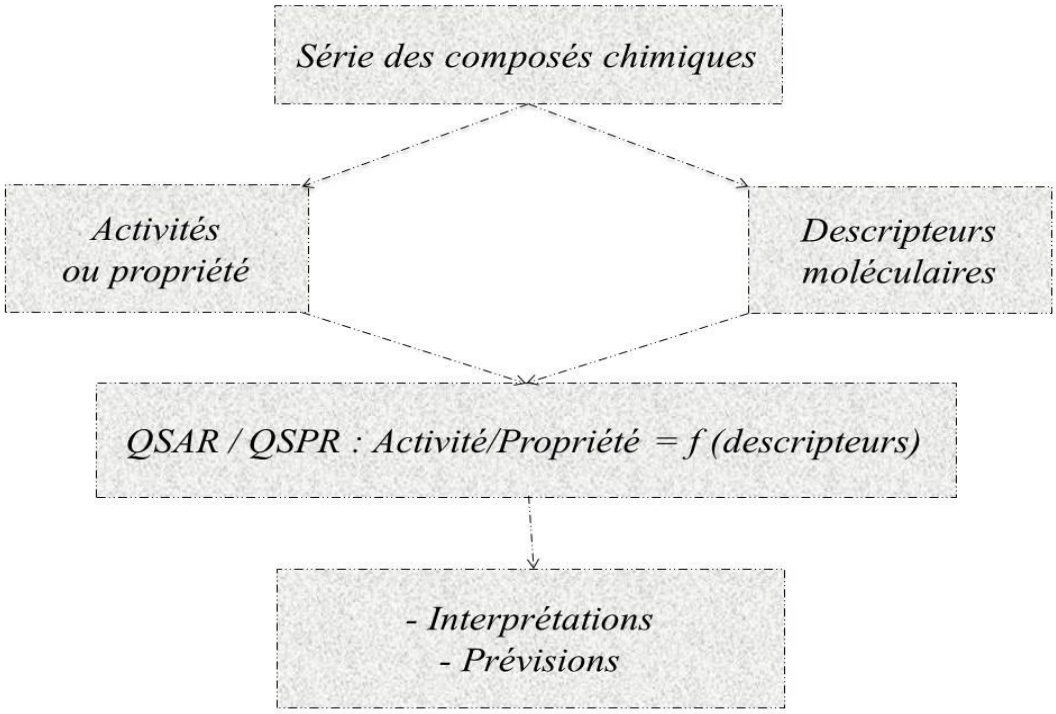


Figure II.7 : Modèle d'étude de relation quantitative structure activité/propriété

II-f-5-Stratégie globale

Le développement d'un modèle débute par la recherche du maximum possible de données expérimentales fiables. Ensuite, le développement d'une série de descripteurs qui caractérisent les structures moléculaires des composés de la base de données en vue de les relier à l'activité/propriété expérimentale étudiée. Une fois développé, le modèle doit être validé en termes de corrélation (sur le jeu de données d'entraînement). L'influence des composés du jeu d'entraînement sur le modèle (robustesse du modèle) est estimée par des méthodes de validation interne. Pour estimer le pouvoir prédictif du modèle, il est nécessaire de disposer de données expérimentales supplémentaires (jeu de données de validation externe) afin de déterminer la capacité du modèle à prédire ces valeurs. Enfin, pour tout modèle, il est important de savoir pour quel type de molécules il est utilisable ou non, c'est-à-dire connaître son domaine d'applicabilité.

Un modèle QSAR/QSPR relie, d'une manière quantitative, la structure des molécules à une activité ou propriété donnée. La stratégie de développement de tels modèles, en

respectant les cinq règles mises en place par l'OCDE (*Organisation de Coopération et de Développement Economique*) pour la validation des modèles QSAR/QSPR (voir plus loin : les principes OECD de validité des modèles QSAR/QSPR), fait intervenir les étapes suivantes :

- Constitution de la base de données structure – activité (ou propriété) à partir de mesures quantitatives, fiables et normalisées de l'activité (ou propriété) cible, pour chaque composé, et sélection des descripteurs moléculaires en relation avec l'activité (ou la propriété) cible afin de traduire de manière numérique la structure des molécules ;
- Division de ce jeu de données en un jeu d'apprentissage et un jeu de test ;
- Construction des modèles à partir du jeu d'apprentissage à l'aide des méthodes statistiques;
- Caractérisation de modèles par leurs indices statistiques et par une validation interne ;
- Validation des modèles avec le jeu de test et calcul de leur indice de corrélation externe ;
- Répétition de l'opération de division pour obtenir d'autres jeux d'apprentissage et de test, et répétition des mêmes étapes (facultative) ;
- Définition du domaine d'applicabilité des modèles proposés afin d'éviter des extrapolations hasardeuses ;
- Exploration et exploitation des modèles validés pour comprendre les mécanismes possibles et faire des prévisions d'activité/propriété pour de nouvelles molécules, si cela est possible.

II-g-Base de données

II-g-1-Source de données

Le choix de la base de données expérimentale initiale est une étape critique pour le développement des modèles QSAR/QSPR. Généralement, les composés testés ont deux origines possibles (dans la plupart des cas sont issus de la littérature), soit des produits de synthèse ou bien des produits d'extraction à partir de plantes. Quelle que soit son origine, il arrive qu'un échantillon ne soit pas pur mais corresponde à un mélange racémique. Le résultat test d'un tel échantillon pose problème : il est impossible de savoir quelle est la contribution de chaque énantiomère dans l'activité observée. Les structures dont la propriété étudiée est mesurée sur un mélange racémique ne peuvent pas être utilisées dans les études QSAR/QSPR [96].

Une base de données, pour être de qualité, doit comprendre des données expérimentales fiables, puisque les barres d'erreurs sur celles-ci se propageront sur le modèle final. Il est donc important de choisir des données présentant de faibles incertitudes afin de limiter les barres d'erreur expérimentales.

D'autre part, l'homogénéité des données est fondamentale. Si l'on veut comparer l'activité/propriété d'une série de molécules, il faut s'assurer, si cela est possible, qu'elle est le résultat de leur interaction avec une seule et même cible et plus précisément avec le même site actif, et l'activité doit être mesurée par un seul et même test, avec des conditions expérimentales identiques pour chaque molécule.

Enfin, la diversité des structures est un facteur important dans la qualité des modèles construits, elle définit l'espace chimique que l'analyse va couvrir.

II-g-2-Homogénéité de la distribution des valeurs

L'homogénéité de la distribution des valeurs mesurées doit être contrôlée. En effet, la plupart des méthodes statistiques reposent sur l'hypothèse que la distribution des valeurs observées suit une loi normale. Il est donc nécessaire de contrôler la normalité de cette distribution. Il existe pour cela des tests statistiques de normalité mais la simple représentation des données sur un histogramme de distribution permet d'évaluer cette caractéristique.
Dans le cas défavorable, des transformations mathématiques (test de *Box-Cox* [97], transformation de *Logit de x* [98], ...) permettent, parfois, de retrouver une distribution normale sans que l'information contenue dans le jeu de données ne soit modifiée.

II-h-Développement de modèles QSAR/QSPR

Nous avons utilisé le logiciel de modélisation moléculaire HyperChem 6.03 [99] pour représenter les molécules puis, à l'aide de la méthode semi-empirique PM3 [51], on a obtenu les géométries finales. Tous les calculs ont été menés dans le cadre du formalisme RHF [100] sans interaction de configuration. Les structures moléculaires ont été prèoptimisées à l'aide de l'algorithme Polak- Ribiere avec pour critère une racine du carré moyen du gradient égale à 0,001kcal/mol. Les géométries ainsi optimisées ont été transférées dans le logiciel informatique Dragon version 5.3 [101] pour le calcul de plus de 1200 descripteurs appartenant à différentes classes. Les descripteurs d'un même groupe, à valeur constante (écarts types inférieurs à 0,0001) ont été exclus. Pour un seuil de corrélation de $R \geq 0,95$ entre deux descripteurs ; celui qui présente le plus de corrélations avec les autres variables, est exclu.

I1-h-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 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 MOBY DIGS de Todeschini [102], en maximisant Q^2_{LOO}.

II-h-2- Méthodes utilisées pour le développement de modèles QSAR/QSPR

L'application pratique des gammes de descripteurs moléculaires dans le développement de modèles QSAR/QSPR n'est pas une tâche aisée [103]. Tout d'abord, un très grand nombre
(~10 000) 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 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 est basée sur la discrimination statistique entre de larges ensembles de descripteurs.

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

II-h-2-1-La régression linéaire multiple :

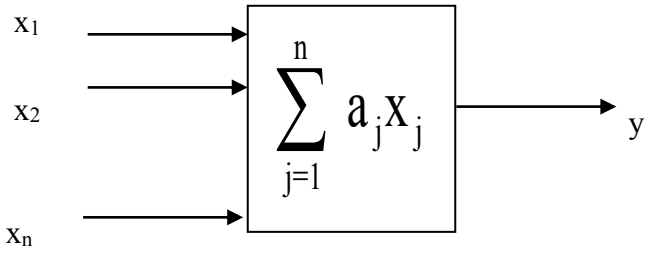
L'étude d'un phénomène peut, le plus souvent, être schématisé de la manière suivante: on s'intéresse à une grandeur y, que nous appellerons par la suite réponse ou variable expliquée, qui dépend d'un certain nombre de variables x_1; x_2; ….x_n que nous appellerons facteurs ou variables explicatives.

La régression est une des méthodes les plus connues et 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 entre les variables prises en compte dans le modèle [114].

La régression multi-linéaire (MLR, pour Multiple Linear Regression) [115] 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 :

$$y = a_0 + a_1x_1 + a_2x_2 + \ldots + a_nx_n \qquad (1.II.71)$$

où les a_i sont les coefficients de la régression.



La détermination de l'équation (1.II.71) 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.

$$\begin{aligned} \hat{y}_1 &= a_0 + a_1 x_{1,1} + a_2 x_{2,1} + \ldots + a_n x_{n,1} + \varepsilon_1 \\ \hat{y}_2 &= a_0 + a_1 x_{1,2} + a_2 x_{2,2} + \ldots + a_n x_{n,2} + \varepsilon_2 \\ \hat{y}_p &= a_0 + a_1 x_{1,p} + a_2 x_{2,p} + \ldots + a_n x_{n,p} + \varepsilon_p \end{aligned} \qquad (1.II.72)$$

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 :

$$\begin{pmatrix} y_1 \\ . \\ . \\ . \\ y_p \end{pmatrix} = \begin{pmatrix} 1 & x_{1,1} & \ldots & x_{n,1} \\ . & . & . & . \\ . & . & . & . \\ . & . & . & . \\ 1 & x_{1,p} & \ldots & x_{1,p} \end{pmatrix} \begin{pmatrix} a_0 \\ . \\ . \\ . \\ a_p \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ . \\ . \\ . \\ \varepsilon_p \end{pmatrix} \qquad (1.II.73)$$

soit de manière condensée :

$$\mathbf{Y} = \mathbf{X}\,\mathbf{A} + \boldsymbol{\varepsilon} \qquad (1.II.74)$$

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 :

$$\hat{y}_i = \hat{a}_0 + \hat{a}_1 x_{1,i} + \ldots + \hat{a}_n x_{n,i} \qquad (1.II.75)$$

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

$$\varepsilon_i = y_i - \hat{y}_i \qquad (1.II.76)$$

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.

$$\min [\sum(\varepsilon_i)^2] = \min [\sum(y_i - \hat{y}_i)^2] = \min [\sum(y_i - \hat{a}_0 - \hat{a}_1 x1_{,i} - \ldots - \hat{a}_n x_{n,i})^2]$$

$$= \min (Y - X\hat{A})^T (Y - X\hat{A}) \qquad (1.II.77)$$

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

$$\mathbf{A} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y} \qquad (1.II.78)$$

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 [116].

II-2-2-Réseaux de Neurones Artificiels

Les réseaux de neurones ont été étudiés depuis les années 40 [117]. Les idées de base de cette technique viennent de la recherche cognitive, d'où vient le nom ″réseaux de neurones″.

La technique inspirait beaucoup de chercheurs à cette époque, mais beaucoup de l'intérêt disparaît après un article de Minsky et Papert [118]. Finalement relancée au début des années 80 après un quasi-oubli d'une vingtaine d'années. La cause de l'intérêt soudain était l'apparition de nouvelles architectures de réseaux de neurones.

II-2-2-1-Le neurone artificiel :

L'élément de base d'un réseau de neurones est, bien entendu, le neurone artificiel. Un neurone (figure II.8) contient deux éléments principaux :

- Un ensemble de poids associés aux connections du neurone, et
- Une fonction d'activation (Figure II.9).

Les valeurs d'entrée sont multipliées par leur poids correspondant et additionnées pour obtenir la somme S.

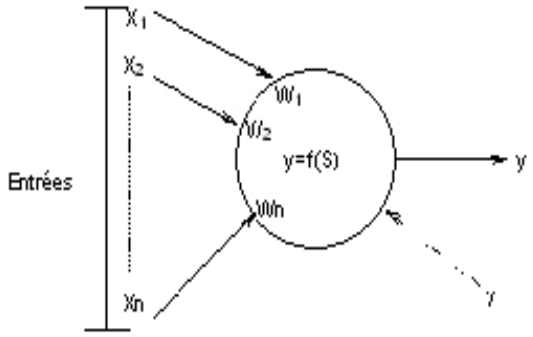
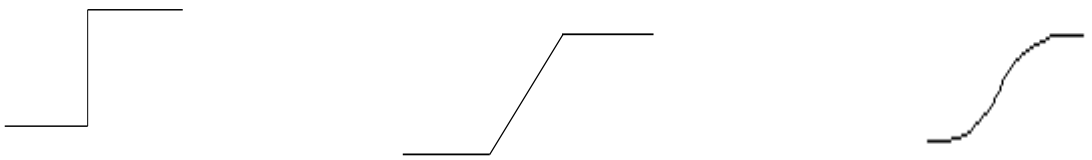


Figure II.8 : le neurone artificiel générique.

Cette somme devient l'argument de la fonction d'activation, qui est le plus souvent une des formes présentées ci- dessous. Une fonction d'activation importante est la simple multiplication avec un, c'est-à-dire que la sortie est simplement une somme pondérée.



Fonction à seuil **Fonction à saturation** **Fonction sigmoïde**

Figure II.9 : Fonctions d'activation.

Le choix de la fonction d'activation dépend de l'application. S'il faut avoir des sorties binaires c'est la première fonction que l'on choisit habituellement.

Une entrée spéciale est pratiquement toujours introduite pour chaque neurone. Cette entrée, normalement appelée biais (bias en anglais), sert pour déplacer le pas de la fonction d'activation sur l'axe S. La valeur de cette entrée est toujours 1 et le déplacement dépend alors seulement du poids de cette entrée spéciale.

II-2-2-2-Propriétés des réseaux de neurones :

Un réseau de neurones se compose de neurones qui sont interconnectés de façon à ce que la sortie d'un neurone puisse être l'entrée d'un ou plusieurs autres neurones. Ensuite il y a des entrées de l'extérieur et des sorties vers l'extérieur [119].

Rumelbart et al. [119] donnent huit composants principaux d'un réseau de neurones:

- Un ensemble de neurones.
- Un état d'activation pour chaque neurone (actif, inactif,…).
- Une fonction de sortie pour chaque neurone (f(S)).
- Un modèle de connectivité entre les neurones (chaque neurone est connecté à tous les autres, par exemple).
- Une règle de propagation pour propager les valeurs d'entrée à travers le réseau vers les sorties.
- Une règle d'activation pour combiner les entrées d'un neurone (très souvent une somme pondérée).
- Une règle d'apprentissage.
- Un environnement d'opération (le système d'exploitation, par exemple).

Le comportement d'un réseau et les possibilités d'application dépendent complètement de ces huit facteurs et le changement d'un seul d'entre eux peut changer complètement le comportement de réseau.

Les réseaux de neurones sont souvent appelés des "boites noires" car la fonction mathématique qui est représentée devient vite trop complexe pour l'analyser et la comprendre directement. Cela est notamment le cas si le réseau développe des représentations distribuées [119], c'est-à-dire que plusieurs neurones sont plus ou moins actifs et contribuent à une décision. Une autre possibilité est d'avoir des représentations localisées, ce qui permet d'identifier le rôle de chaque neurone plus facilement. Les réseaux de neurones ont quand même une tendance à produire des présentations distribuées.

II-h-2-2-3-Les différents types de réseaux de neurones

Plusieurs types de réseaux de neurones ont été développés qui ont des domaines d'application souvent très variés. Notamment quatre types de réseaux sont bien connus :

- Le réseau de Hopfield (et sa version incluant l'apprentissage, la machine de Boltzmann).
- Les cartes auto-organisatrices de Kohonen .
- Les réseaux à fonction radiale que l'on nomme aussi RBF (pour " Radial Basic Functions ").
- Les réseaux multicouches ou perceptron multicouches PMC

Le réseau de Hopfield [120] est un réseau avec des sorties binaires où tous les neurones sont interconnectés avec des poids symétriques. C'est-à-dire que le poids du neurone N_i au neurone N_j est égal au poids du neurone N_j au neurone N_i. Les poids sont donnés par l'utilisateur. Les poids et les états des neurones permettent de définir l'«énergie» du réseau.

C'est cette énergie que le réseau tente de minimiser pour trouver une solution. La machine de Boltzmann est en principe un réseau de Hopfield, mais qui permet l'apprentissage grâce à la minimisation de cette énergie.

Les cartes auto-organisatrices de Kohonen [121] sont utilisées pour faire des classifications automatiques des vecteurs d'entrées.

Les réseaux à fonction radiale sont des réseaux multicouches, à une couche cachée. Cependant, contrairement aux perceptrons multicouches, les fonctions de transfert de la couche cachée dépendent de la distance entre le vecteur d'entrée et le vecteur centre.

Les réseaux multicouches (PMC) sont les réseaux les plus puissants des réseaux de neurones qui utilisent l'apprentissage supervisé.

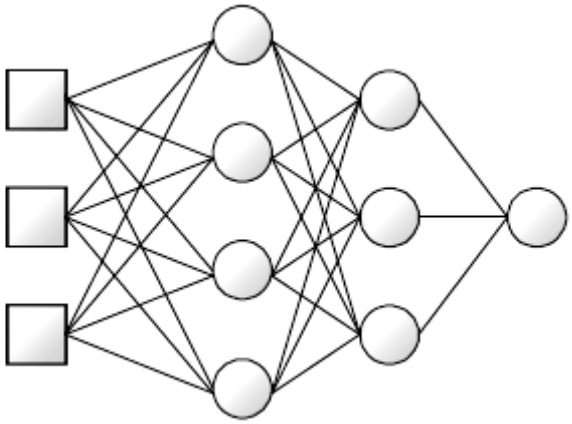
II-h-2-2-4-Les réseaux multicouches ou perceptron multicouches (PMC)

Les réseaux multicouches (PMC) (figure II.10) se composent des entrées, une couche de sortie et zéro ou plusieurs couches cachées [119]. Les connexions sont permises seulement d'une couche inférieure (plus proche des entrées) vers une couche supérieure (plus proche de la couche de sortie). Il est aussi interdit d'avoir des connexions entre des neurones de la même couche.

Les entrées servent à distribuer les valeurs d'entrée aux neurones des couches supérieures, éventuellement multipliées ou modifiées d'une façon ou d'une autre.

La couche de sortie se compose normalement des neurones linéaires qui calculent seulement une somme pondérée de toutes ses entrées.

Les couches cachées contiennent des neurones avec des fonctions d'activation non linéaires, normalement la fonction sigmoïde.



Les entrées Couches cachées Couche de sortie

Figure II.10 : Structure générale du perceptron multicouche

Il a été prouvé [122] qu'il existe toujours un réseau de neurones de ce type avec trois couches seulement (les entrées, couche de sortie et une couche cachée) qui peut approximer une fonction f : $[0\text{-}1]^n \Rightarrow R^n$ avec n'importe quelle précision $\varepsilon > 0$ désirée. Un problème consiste à trouver combien de neurones cachés sont nécessaires pour obtenir cette précision. Un autre problème est de s'assurer a priori qu'il est possible d'apprendre cette fonction.

Initialement tous les poids peuvent avoir des valeurs aléatoires, qui sont normalement très petites avant de commencer l'apprentissage.

II-h-2-2-5-Apprentissage :

L'apprentissage d'un réseau de neurones signifie qu'il change son comportement de façon à lui permettre de se rapprocher d'un but défini. Ce but est normalement l'approximation d'un ensemble d'exemples ou l'optimisation de l'état du réseau en fonction de ses poids pour atteindre l'optimum d'une fonction économique fixée a priori.

Il existe trois types d'apprentissages principaux. Ce sont l'apprentissage supervisé, l'apprentissage non-supervisé, et l'apprentissage par tentative (graded training en anglais) [122].

On parle d'apprentissage supervisé quand le réseau est alimenté avec la bonne réponse pour les exemples d'entrées donnés. Le réseau a alors comme but d'approximer ces exemples aussi bien que possible et de développer à la fois la bonne représentation mathématique qui lui permet de généraliser ces exemples pour ensuite traiter de nouvelles situations (qui n'étaient pas pressenties dans les exemples).

Dans le cas de l'apprentissage non-supervisé le réseau décide lui-même quelles sont les bonnes sorties. Cette décision guidée par un but interne au réseau qui exprime une configuration idéale à atteindre par rapport aux exemples introduits. Les cartes auto-organisatrices de Kohonen sont un exemple de ce type de réseau [121].

'Graded learning' est un apprentissage de type essai-erreur où le réseau donne une solution en étant seulement alimenté avec une information indiquant si la réponse était correcte, ou si elle était au moins meilleure que la précédente.

Il existe plusieurs règles pour chaque type d'apprentissage. L'apprentissage supervisé est le type le plus utilisé. Pour ce type d'apprentissage la règle la plus exploitée est celle de Widrow-Hof. D'autres règles d'apprentissage sont par exemple la règle de Hebb, la règle de perceptron, la règle de Grossberg etc …. [119, 122, 123].

II-h-2-2-5-1-L'apprentissage de Widrow-Hof :

La règle d'apprentissage de Widrow-Hof est une règle qui permet d'ajuster les poids d'un réseau de neurones pour diminuer à chaque étape l'erreur commise par ce réseau de neurones (à condition que le facteur d'apprentissage soit bien choisi).
Un poids est modifié en utilisant la formule suivante :

$$w_{k+1} = w_k - \alpha \delta_k x_k \qquad (1.II.79)$$

Où :

w_k est le poids à l'instant k ;

w_{k+1} le poids à l'instant k+1 ;

α est le facteur d'apprentissage ;

δ_k caractérise la différence entre la sortie attendue et la sortie effective d'un neurone à l'instant k ;

x_k la valeur de l'entrée avec laquelle le poids w est associé à l'instant k.

Ainsi, si δ_k et x_k sont positifs tous les deux, alors le poids doit être augmenté. L'ampleur du changement dépend avant tout de la grandeur de δ_k mais aussi de celle de x_k. Le coefficient α sert à diminuer les changements pour éviter qu'ils deviennent trop grands, ce qui peut entraîner des oscillations du poids.

Deux versions améliorées de cet apprentissage existent, la version « par lois » et la version « par inertie » (momentum en anglais) [122], dont l'une utilise plusieurs exemples

pour calculer la moyenne des changements requis avant de modifier le poids et l'autre empêche que le changement du poids au moment k ne devienne beaucoup plus grand qu'au moment k-1.

II-h-2-2-5-2-L'apprentissage par rétro-propagation du gradient (Levenberg-Marquardt back-propagation)

L'algorithme d'apprentissage par rétro-propagation du gradient (figure II.11) est un algorithme itératif qui a pour objectif de trouver le poids des connexions minimisant l'écart commis par le réseau sur l'ensemble d'apprentissage. Cette minimisation par une méthode du gradient conduit à l'algorithme d'apprentissage de rétro-propagation.

La procédure d'apprentissage se décompose en deux étapes. Pour commencer, les valeurs d'entrées sont présentées au réseau, qui propage ensuite ces valeurs jusqu'à la couche de sortie et donne ainsi la réponse au réseau. A la deuxième étape les bonnes sorties correspondantes sont présentées aux neurones de la couche de sortie qui calculent l'écart, modifient leurs poids et rétro-propagent l'erreur jusqu'aux entrées pour permettre aux neurones cachés de modifier leurs poids de la même façon. Le principe de modification des poids est normalement l'apprentissage de Widrow-Hoff.

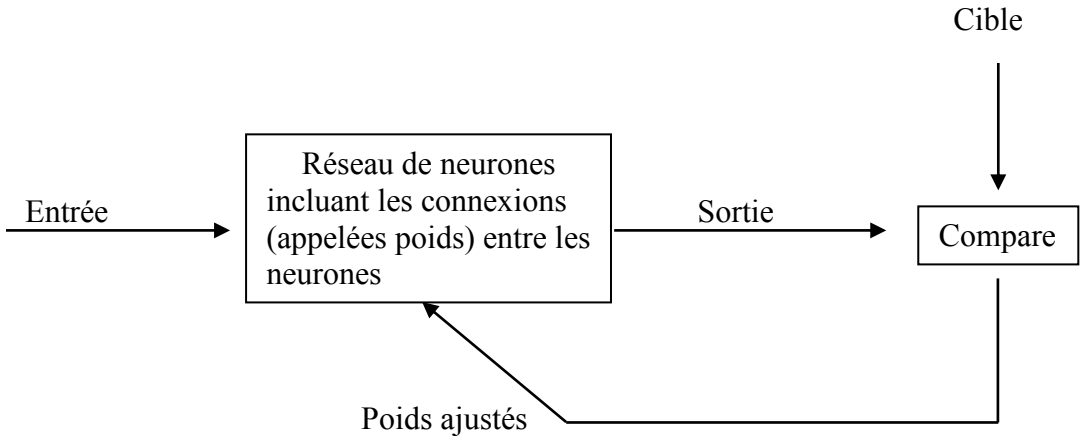


Figure II.11 : Apprentissage par un algorithme de rétro-propagation

Généralement pour le calcul de l'écart on utilise l'erreur quadratique moyenne *MSE* (*Mean Square Error*) définie par la relation :

$$MSE = \frac{\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2}{n} \qquad (1.II.80)$$

y_i est la valeur observée, $\hat{y}_i$ est la valeur estimée, et n le nombre d'observations.

II-h-2-2-6-Critères d'arrêt

Plusieurs critères d'arrêt peuvent être utilisés avec l'algorithme d'apprentissage. Le premier critère consiste à fixer un nombre préalable de cycles ou d'itérations, mais il est difficile de savoir a priori combien d'itérations seraient appropriées pour arriver au but fixé.

Un deuxième critère consiste à fixer une borne inférieure sur l'erreur quadratique moyenne (MSE), il est parfois possible de fixer a priori un objectif à atteindre. Lorsque l'indice de performance choisi diminue en dessous de cet objectif, on considère simplement que le réseau a suffisamment bien appris ses données et on arrête l'apprentissage. L'inconvénient de ce critère est qu'il peut engendrer un phénomène de sur-apprentissage indésirable dans la pratique.

Le troisième critère est "l'arrêt précoce", qui consiste à suivre l'évolution des performances du réseau de généralisation durant le déroulement de l'apprentissage et à stopper celui-ci juste avant que ces performances ne se mettent à se dégrader, c'est-à-dire dès que l'indice de performance calculé sur les données de validation cesse de s'améliorer. Cette méthode, la plus utilisée pour éviter le sur-apprentissage, est celle pour laquelle nous avons optée dans ce travail. Le graphe suivant illustre ce critère :

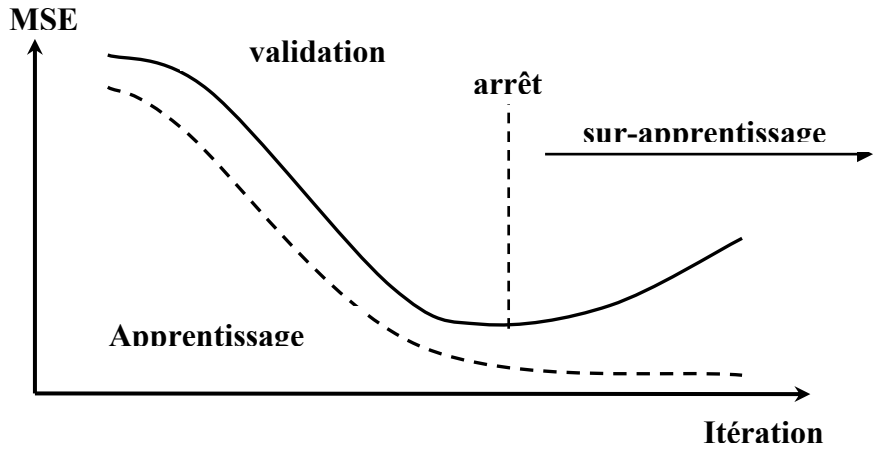


Figure II.12 : Illustration de l'arrêt précoce.

II-h-2-2-7-Construction d'un modèle

La construction d'un modèle implique dans un premier temps le choix des échantillons des données d'apprentissage, de test et de validation. Le choix du type de réseau intervient dans une seconde étape.

Les quatre grandes étapes de la création d'un réseau de neurones sont détaillées ci-après :

II-h-2-2-7-1-Construction de la base de données

Le processus d'élaboration d'un réseau de neurones commence par la construction d'une base de données.

Afin de développer une application à base de réseaux de neurones, il est nécessaire de disposer de deux bases de données, une pour effectuer l'apprentissage et l'autre pour tester le réseau obtenu et déterminer ses performances.

Généralement, les bases de données subissent un prétraitement qui consiste à effectuer une normalisation appropriée tenant compte de l'amplitude des valeurs acceptées par le réseau.

Les valeurs d'entrées et de sortie sont normalisées dans un intervalle spécifique afin de donner à chaque paramètre la même influence statistique. Les valeurs d'apprentissage et de test ont été normalisées dans la marge [- 1, 1], au moyen de l'équation :

$$x_{norm} = 2 \times \frac{(x_j - x_{\min})}{(x_{\max} - x_{\min})} - 1 \qquad (1.II.81)$$

où x_{norm} est la valeur normalisée ; x_j est la j^{ième} valeur ; $x_{\max}$ est la valeur maximale ; $x_{\min}$ est la valeur minimale

II-h-2-2-7-2-Définition de la structure du réseau

Nous avons retenu le Perceptron Multicouches comme base du modèle. Nous structurons ce réseau en précisant le nombre de couches et de neurones cachés pour que le réseau soit en mesure de reproduire ce qui est déterministe dans les données.

II-2-2-7-3-Nombre de couches et de neurones cachés

Les entrées et la couche de sortie mises à part, il faut décider du nombre de couches cachées. Sans couche cachée, le réseau n'offre que de faibles possibilités d'adaptation. Néanmoins, il a été démontré qu'un Perceptron Multicouches avec une seule couche cachée pourvue d'un nombre suffisant de neurones, peut approximer n'importe quelle fonction avec la précision souhaitée [124].

Chaque neurone peut prendre en compte des profils spécifiques de neurones d'entrée. Un nombre plus important permet donc de mieux "coller" aux données présentées mais diminue la capacité de généralisation du réseau. Il faut alors trouver le nombre adéquat de neurones cachés nécessaires pour obtenir une approximation satisfaisante.

II-h-2-2-7-4-Présentation de l'environnement utilisé

Dans cette optique, le logiciel MATLAB [125], qui contient un module consacré au développement de réseaux de neurones, a été retenu ; un PC Dell P4 avec une Ram de 512 et une vitesse de 3,4 GHZ a été utilisé.

Le réseau de neurones stocke l'information dans une chaîne d'interconnexions neuronales, en faisant appel à la notion de poids (poids entrée - couche cachée = *IW* (i*nitial weights)*, poids couche cachée - sortie = *LW(last weights*).

Une capacité d'apprentissage est nécessaire pour ajuster les poids des réseaux de neurones pendant la phase d'apprentissage au cours de laquelle toutes les données sont présentées au RNA à plusieurs reprises.

La fonction sigmoïde de transfert, tangente hyperbolique, a été adoptée comme fonction d'activation pour les couches cachées et de sortie.

II-h-2-3-Machines à vecteurs support

La régression par Machines à vecteurs support (SVR) [126] 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 [127]. 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 [126] 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 [128].

Les méthodes SVM 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; ⋋, qui correspond à la largeur du tube d'insensibilité ; les éventuels paramètres de la fonction noyau k (σ,γ,..). Ces hyperparamètres sont en générale 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 [129].

Cela implique de réaliser l'apprentissage pour différentes valeurs et d'estimer leur performance. 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.

II-i-Paramètres d'évaluation de la qualité de l'ajustement

L'ajustement des modèles QSPR 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 :

$$R^2 = 1 - \frac{SCE}{SCT} = 1 - \frac{\sum_1^n (y_i - \hat{y}_i)^2}{\sum_1^n (y_i - \bar{y})^2} \qquad (1.II.82)$$

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 SDEC) :

$$SDEC = \sqrt{\frac{\sum_{i=1}^{n} \left(y_i - \hat{y}_i \right)^2}{n}} \qquad (1.II.83)$$

II-i-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*) [130] et LMO (*Leave Many Out*) [131].

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 composés du jeu d'entraînement, puis le coefficient de prédiction noté Q^2 est calculé à l'aide de l'équation suivante :

$$Q^2_{LOO} = \frac{SCT - PRESS}{SCT} \qquad (1.II.84)$$

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

$$PRESS = \sum_1^n \left(y_i - \hat{y}_{(i)}\right)^2 \quad (1.II.85)$$

Et le SDEP par :

$$\sigma_N = SDEP = \sqrt{\frac{\sum_{i=1}^{n}\left(y_i - \hat{y}_{(i)}\right)^2}{n}} = \sqrt{\frac{PRESS}{n}} \quad (1.II.86)$$

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 [132].

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 [133]. 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.

II-i-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 II.13). On calcule alors sur le vecteur obtenu (considéré comme vecteur expérimental réel) un modèle QSPR, 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 [134,135] et celle qui ne le fait pas [136]. 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.

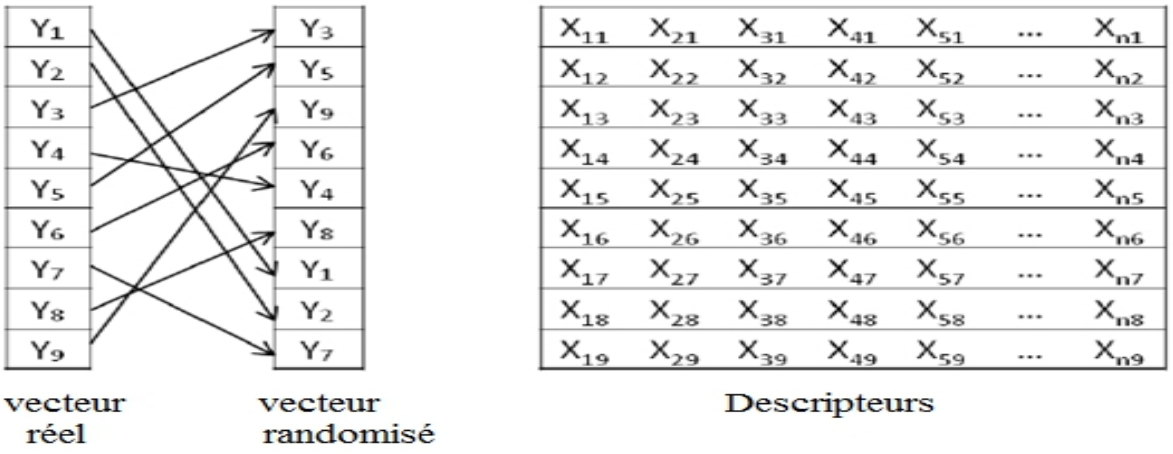


Figure II.13 : Illustration de la méthode du test de randomisation

II-i-3-Validation externe

La meilleure façon d'estimer la véritable puissance prédictive d'un modèle QSPR 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 [119, 120].

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

$$\boldsymbol{R}^2_{\mathrm{CV,ext}} = 1 - \frac{\sum_{i=1}^{n_{ext}} \left(y_i - \hat{y}_i\right)^2}{\sum_{i=1}^{n_{ext}} \left(y_i - \bar{y}_{TR}\right)^2} \qquad (1.\mathrm{II}.87)$$

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

$$\boldsymbol{SDEP}_{ext} = \sqrt{\frac{\sum_{i=1}^{n_{ext}} \left(y_i - \hat{y}_{(i)}\right)^2}{n_{ext}}} \qquad (1.\mathrm{II}.88)$$

L'article de Chirico et Gramatica [137] répertorie différents coefficients de calcul de la prédictivité présentés ci-après.

Le coefficient Q^2F1 proposé par Tropsha [133, 138, 139] n'est pas une vraie validation externe car il fait intervenir des informations sur le jeu de calibrage.

$$Q^2_{F1} = 1 - \frac{\sum_{i=1}^{n_{ext}} (\hat{y}_i - y_i)^2}{\sum_{i=1}^{n_{ext}} (y_i - \bar{y}_{TR})^2} \quad (1.II.89)$$

Avec y_i la valeur expérimentale de la propriété, $\hat{y}_i$ la valeur prédite/calculée de la propriété et $\bar{y}_{TR}$ la moyenne des valeurs y_i du jeu d'entraînement.

En 2008, une autre mesure de la prédictivité, proposée par Schüürmann [140], est le Q^2F2 qui se différencie de Q^2F1 par le fait que la moyenne utilisée au dénominateur est celle du jeu de validation et non celle du jeu d'entraînement : il s'agit donc bien d'une validation externe car aucune donnée du jeu d'entraînement n'est nécessaire. De plus, Q^2F1 est plus optimiste car supérieur ou égal à Q^2F2 et par conséquent accepte plus facilement les modèles. Le risque d'avoir un modèle non prédictif accepté est moins grand avec Q^2F2.

$$Q^2_{F2} = 1 - \frac{\sum_{i=1}^{n_{ext}} (\hat{y}_i - y_i)^2}{\sum_{i=1}^{n_{ext}} (y_i - \bar{y}_{EXT})^2} \quad (1.II.90)$$

$\bar{y}_{EXT}$ étant la moyenne des valeurs y_i du jeu de validation.

En 2009, le coefficient Q^2F3 a été proposé par Consonni [141] afin de supprimer le biais introduit par la distribution des données. De plus, selon Consonni, l'absence d'information sur le jeu d'entraînement est un désavantage. En effet, il a été observé que la valeur de Q^2F3 est identique quelle que soit la distribution du jeu de validation. Il semble également être insensible au nombre de composés. En effet, la valeur de Q^2F3 ne change pas avec la taille du jeu de validation, contrairement à Q^2F2 dont la valeur augmente avec le nombre de composés. Cependant, tout comme Q^2F1, ce coefficient n'est pas une vraie validation externe car il fait intervenir des informations sur le jeu d'entraînement.

$$Q^2_{F3} = 1 - \frac{\left[\sum_{i=1}^{n_{EXT}} (\hat{y}_i - y_i)^2\right] / n_{EXT}}{\left[\sum_{i=1}^{n_{TR}} (y_i - \bar{y}_{TR})^2\right] / n_{TR}} \quad (1.II.91)$$

où n_{TR} est le nombre de molécules du jeu d'entraînement et n_{EXT} le nombre de molécules dans le jeu de validation.

Le dernier coefficient CCC [142, 143] mesure à la fois la précision (distance par rapport à l'équation) et la justesse (c'est-à-dire à quel point la ligne de la régression dévie de la droite x=y dite « concordance line »). Il s'agit d'une validation externe car aucune information du jeu d'entraînement n'est nécessaire

$$CCC = \frac{2\sum_{i=1}^{n}(x_i - \bar{x}_i)(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x}_i)^2 + \sum_{i=1}^{n}(y_i - \bar{y})^2 + n(\bar{x} - \bar{y})^2} \qquad (1.II.92)$$

Tous ces coefficients ont pour but l'amélioration de la validation du modèle et ainsi d'augmenter la confiance en ce type de méthode. Le but étant de pouvoir utiliser les modèles QSPR avec assurance pour prédire les propriétés physico-chimiques.

Une validation externe supplémentaire selon [137] est appliquée uniquement à l'ensemble de validation. Selon les critères recommandés de Tropsha et *al*., un modèle QSPR prédictif, doit remplir les conditions suivantes:

1) $Q^2_{EXT} > 0.5$ (1.II.93-a)

2) $R^2 > 0.6$ (1.II.93-b)

3) $(R^2 - R_0^2)/R^2 < 0.1$ et $0.85 < k < 1.15$ (1.II.93-c)

$(R^2 - R'^2_0)/R^2 < 0.1$ et $0.85 < k' < 1.15$ (1.II.93-d)

où

$$R = \frac{\sum(y_i - \bar{y})(\tilde{\quad} \quad \tilde{\quad})}{\sqrt{\sum(y_i - \bar{y})^2 \sum(\tilde{\quad} \quad \tilde{\quad})^2}} \qquad (1.II.94\text{-a})$$

$$R_0^2 = 1 - \frac{\sum(y_i - y_i^{r_0})^2}{\sum(y_i - \bar{y})^2} \qquad (1.II.94\text{-b})$$

$$R'^2_0 = 1 - \frac{\sum(\tilde{\quad} \quad \tilde{\quad})^2}{\sum(\tilde{\quad} \quad \tilde{\quad})^2} \qquad (1.II.94\text{-c})$$

$$k = \frac{\sum(y_i \tilde{\quad})}{\sum(\tilde{\quad})^2} \qquad (1.II.94\text{-d})$$

$$k' = \frac{\sum(y_i \tilde{\quad})}{\sum(y_i)^2} \qquad (1.II.94\text{-e})$$

R est le coefficient de corrélation entre les valeurs calculées et expérimentales dans l'ensemble de test; R_0^2 (valeurs calculées par rapport à celles observées) et $R_0'^2$ (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 $\tilde{}$. sont définis respectivement par : $y_i^{r_0} = k\,\tilde{}$ et, $\tilde{}$. _ ; 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.

II-j-Les méthodes de choix des échantillons de calibrage et de validation [144]

Pour le choix des modèles représentatifs, plusieurs méthodes de sélection d'échantillons peuvent être utilisées ; nous avons appliqué, le choix aléatoire, l'algorithme de Kennard-Stone (CADEX) et l'algorithme (DUPLEX).

II-j-1-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 [145]. Cependant, le problème avec cette approche est qu'il ya une chance que la scission de données souffre de la variance, ou de partialité, en particulier lorsque les données ne sont pas réparties uniformément [146].

II-j-2-Algorithme de Kennard-Stone (CADEX) [147]

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.

II-j-3-Algorithme (DUPLEX) [148]

Une version améliorée appelée DUPLEX a été proposée par Snee [149]; il est largement utilisé dans le domaine de la chimiométrie, y compris plusieurs applications ANN [150,151]. 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.* [152], 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 :

$$z_{ij} = \frac{x_{ij} - \overline{x_j}}{s_j \sqrt{n-1}} \quad \text{i=1,…, ℓ}$$

Où

$s\,j$: Ecart-type du j^{ème} régresseur.

$x\,j$: 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 $Z = (z_{ij})$ par sa transposée Z', sous la forme : Z'.Z =T'.T

T est une matrice (ℓ × ℓ) triangulaire supérieure unique, dont les éléments peuvent être obtenus par la méthode de Cholesky [153]. On opère alors la transformation : W= Z.T^{-1} 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^2n 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 [154].

PARTIE.2 Applications/Résultats et discussions

Chapitre I
Modélisation du temps de rétention relatif

I-1- Introduction :

L'atmosphère est en permanence l'objet d'une importante contamination par de nombreux polluants d'origines naturelles ou anthropiques, présents à l'état gazeux ou sous forme particulaire. Parmi ceux-ci, les COV « une large famille de polluants, dont une majorité a un effet direct ou indirect sur la santé humaine et l'environnement » tiennent une place remarquable du fait des quantités présentes, de la diversité de leurs origines, de leurs structures, et de leurs caractéristiques vis-à-vis des écosystèmes [155].

Toutefois, la voie expérimentale, en utilisant des tests sur les animaux est généralement un processus long, coûteux et techniquement difficile et même éthiquement discutable [155]. Pour ces raisons, les techniques de modélisation capables d'estimer les propriétés biologiques d'une manière plus économique, plus rapide et plus facile sont devenues d'un intérêt potentiel. L'objectif majeur de l'analyse QSAR est de trouver une relation mathématique entre l'activité et les descripteurs liés à la structure de la molécule [156, 157]. Ces études ont été extrapolées aux techniques séparatives en corrélant les propriétés physico-chimiques des COV avec les temps de rétention relatifs ; c'est l'étude Relations Quantitatives Structure/Rétention noté QSRR [158].

Lors du passage d'un soluté dans une colonne chromatographique, les molécules de soluté vont se partager entre la phase mobile et la phase stationnaire. Ce partage, lié à la volatilité du soluté d'une part, et à son affinité pour la phase stationnaire d'autre part, va exister tout au long de la colonne et conduire à la « rétention » du soluté. Cette rétention s'exprime alors par rapport au temps que mettrait un soluté qui n'irait jamais dans la phase stationnaire et qui se déplacerait donc à la vitesse de la phase mobile [159].

Le temps de rétention chromatographique t_r représentant le temps que le soluté passe dans la colonne, il est donc égal à la somme des temps passés dans la phase mobile et dans la phase stationnaire [159].

Il dépend :

-du couple soluté - phase stationnaire (plus le coefficient de partage est grand c'est à dire plus la substance est retenue par la phase stationnaire et plus la substance sort tardivement et plus le pic est large),

-de l'étendue des volumes vides de la colonne.

-du débit du gaz vecteur,

-de la masse de phase liquide stationnaire dans la colonne (longueur de la colonne et taux d'imprégnation),

-de la température de la colonne (action sur le coefficient de partage : plus la température de la colonne est élevée plus la substance sort rapidement et plus le pic est étroit).

Il est indépendant:

- de la quantité injectée (tant qu'elle est faible),
- de la nature et de l'abondance des autres constituants du mélange,
- de la nature du gaz vecteur,
- de sa pression.

Cependant les molécules de soluté ne passent pas tout le temps t_R dans la phase stationnaire. Elles consacrent le temps t_m à parcourir les vides et interstices de la colonne. En pratique, cette durée comporte aussi le temps de balayage des volumes de l'injecteur et des tuyaux de raccordement, volumes qu'il convient de réduire au minimum. Le temps t_m apparaît ainsi comme le temps de rétention d'un composé non retenu par la phase stationnaire. Comme c'est généralement le cas de l'air en C G L : on l'appelle temps de rétention de l'air t_0.

La différence est appelée temps de rétention réduit :

$$t_r^{'} = t_r - t_0 \qquad (2.I.95)$$

Temps de rétention relatif : Il est souvent plus aisé de comparer les temps de rétention des substances à celui d'une substance de référence.

Le temps de rétention relatif d'une substance devient donc [160]:

$$t_r\text{Relatif} = t_r^{'}\text{substance}/t_r^{'}\text{ref} \qquad (2.I.96)$$

L'objectif de cette étude est d'établir des modèles QSRR fiables pour la prédiction des temps de rétention relatifs de composés organiques volatils.

I-2-Résultats et discussion

I-2-1-La régression linéaire multiple :

I-2-1-1-Calcul et sélection des descripteurs moléculaires :

Les données ont été prélevées dans la littérature [161] et les molécules ont été dessinées en utilisant le logiciel Hyperchem 6, 03 [99]. Les géométries finales ont été optimisées par la méthode PM3. Les géométries obtenues ont été transférées dans le logiciel Dragon version 5,3 [101] pour le calcul des descripteurs de types géométrique, topologique et auto corrélation 2D. Les descripteurs fortement corrélés entre eux ($R \geq 0,95$) ont été éliminés. Les descripteurs moléculaires du modèle ont été sélectionnés à l'aide de l'algorithme

génétique, dans la version MOBY DIGS de Todeschini [102] en maximisant le coefficient de prédiction Q^2_{LOO}.

I-2-1-2-Calcul des modèles :

L'utilisation de l'algorithme génétique conduit à de nombreux modèles de différentes dimensions. Parmi ces modèles un modèle à cinq descripteurs a été choisi. Le modèle développé pour 92 composés organiques qui appartiennent à 13 classes différentes est comme suit :

$$t_{rr} = f\,(J, PCR, X\,3v, X0sol, \mu) \qquad (2.I.97)$$

J : Blaban distance index de connectivité.

PCR : Ratio du nombre de chemins multiples sur le chemin d'accès.

X3v : Indice de connectivité de valence chi-3

X0sol : Indice de connectivité de solvatation chi-2

μ : Moment dipolaire.

La matrice de corrélation montre que les descripteurs du modèle choisi ne sont pas fortement corrélés entre eux.

Tableau I.9 : Matrice de corrélation des descripteurs du modèle.

	t_{rr}	J	PCR	X3v	X0sol
J	0,030				
	0,778				
PCR	0,379	-0,161			
	0,000	0,125			
X3v	0,708	0,318	0,028		
	0,000	0,002	0,791		
X0sol	0,747	0,553	0,168	0,685	
	0,000	0,000	0,110	0,000	
μ	-0,210	-0,071	0,045	-0,495	-0,304
	0,045	0,501	0,667	0,000	0,003

Les valeurs expérimentales des temps de rétention relatifs (t_{rr}) des 122 composés organiques, tirées de la littérature [161] ont été divisées à l'aide de l'algorithme Kennard and

Stone (CADEX) en deux ensembles l'un de calibrage (92 composés) utilisé pour développer le modèle et l'autre de validation (30 composés) utilisé pour la prédiction externe.

Le modèle basé sur les descripteurs sélectionnés a pour équation :

t_{rr} = - 2,10 (± 0,47) - 1,08 (± 0,11) J + 1,43 (± 0,38) PCR + 0,72 (± 0,09) X3v + 0,75 (± 0,06) X0sol + 0,18(± 0,04) u (2.I.98)

Les paramètres statistiques du modèle optimal construit sur les 92 observations de l'ensemble d'estimation sont réunis dans le tableau I.10

Selon les valeurs du test t (| t |), on peut classer les descripteurs sélectionnés dans ce modèle d'après leur pourcentage de contribution qui se présente dans l'ordre: X0sol > J > X3v > u > PCR. 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.10 : Paramètres statistiques du modèle optimal.

Descripteur	x	Dx	t	Probabilité-t	VIF
Constante	-2,1037	0,47770	-4,40	0,000	
J	-1,0755	0,10920	-9,85	0,000	1,649
PCR	1,4286	0,38250	3,75	0,000	1,178
X3v	0,7179	0,09417	4,62	0,000	2,309
X0sol	0,7477	0,06421	11,64	0,000	2,789
u	0,1813	0,04137	4,38	0,000	1,351

Le Tableau I.11 regroupe les paramètres statistiques liés au modèle.

Tableau I.11 : Paramètres statistiques liés au modèle.

R^2	87, 3900
Q^2_{loo}	85, 5000
Q^2_{BOOT}	84, 3500
Q^2_{Ext}	90, 1200
R^2adj	86, 6600
SDEC	0, 3980
SDEP	0, 4270
$SDEP_{Ext}$	0, 3530
F	119, 2136
PRESS	16, 7972

$Q^2_{LMO\ /\ 10}$	84, 6017
$Q^2_{LM\ O/\ 20}$	84, 4700
$Q^2_{LMO\ /\ 40}$	86, 9628
$Q^2_{LMO\ /\ 30}$	85, 4673
$Q^2_{LMO\ /\ 50}$	84, 7707
K_x	34,1900
K_{xy}	42, 6800
DK	8, 5000
SE	0, 4121

Les statistiques calculées établissent la pertinence du modèle. En effet, la valeur de R^2 signifie que 87,39 % de la variabilité de t_{rr} est expliquée par les 5 descripteurs sélectionnés. Les valeurs de R^2 et R^2_{adj} établissent la qualité de l'ajustement du modèle, et les valeurs proches de R^2 et Q^2_{LOO} indiquent sa robustesse. D'après la valeur du paramètre de Fisher F le modèle est très significatif. La ressemblance de SDEC et SDEP, ainsi que les valeurs proches de Q^2_{LMO} montrent la stabilité du modèle. La validation par bootstrap (Q^2_{Boot}) confirme tout à la fois la capacité de prédiction interne et la stabilité du modèle.

I-2-1-3-Analyse des résidus et diagnostiques d'influence :

Les valeurs expérimentales et calculés des temps de rétention relatifs ainsi que les valeurs des résidus standardisés et des leviers des 92 composés de l'ensemble de calibrage sont présentées dans le tableau I.12 (page suivante).

I-2-1-4-Validation externe :

Pour vérifier les capacités prédictives de notre modèle nous l'avons validé sur un ensemble de 30 composés. Cet ensemble n'a pas été utilisé pour la construction de modèle. Le tableau I.12 (page suivante) donne quelques caractéristiques de l'ensemble des composés organiques de validation externe.

Tableau I.12 : Valeurs des t_{rr} expérimentales, calculées/prédites, h_{ii}, e_{iStd} et les descripteurs sélectionnés de 122 COVs.

Composé	t_{rrExp}	$t_{rrCalc/Pred}$	hi	$e_{iStd.}$	J	PCR	X3v	X0sol	u
Dibromomethane	1.2668	1.351	0.106	0.2416	1.633	1	0	4.707	1.449
1,1,1-trichloroethane	0.9191	0.8085	0.125	-0.3278	3.024	1	0	6	1.376
1,3-Dibromopropane	3.8305	2.8724	0.073	-2.6036	2.191	1	1.389	6.121	1.819
1,1,2-Trichloro-1,2,2-trifluoroethane	0.5878	0.3835	0.147	-0.6291	4.02	1	1.25	5.784	0.8814
Ethyl disulfide	3.7395	3.0099	0.147	-2.2483	2.339	1	2.871	5.536	0.0000229
3-Bromopentane	2.3527	2.7577	0.087	1.1263	2.754	1	2.181	5.992	1.922
Iodomethane	0.5803	1.128	0.129	1.6333	1	1	0	3.5	1.442
1-Pentanethiol	2.7509	1.922	0.033	-2.1147	2.339	1	1.078	5.328	1.957
Ethyl iodide	0.7939	1.0456	0.076	0.6873	1.633	1	0	4.207	1.827
1-Bromobutane	1.5094	1.8804	0.033	0.9465	2.191	1	1.048	5.121	1.822
1-Bromopentane	2.9502	2.4317	0.048	-1.3536	2.339	1	1.298	5.828	1.835
2-Hexanone	2.4268	1.9252	0.038	-1.2894	2.678	1.072	0.882	5.699	2.664
Cyclopentylchloride	2.272	2.1697	0.061	-0.2726	2.184	1	1.721	4.906	1.598
2-Methylheptane	1.966	2.1962	0.053	0.6065	2.716	1	1.385	6.406	0.04448
2-Nitropropane	1.2814	1.1962	0.09	-0.238	2.993	1.125	0.422	5.155	4.159
Butyl formate	1.5044	2.1591	0.081	1.8032	2.447	1.089	0.684	5.536	3.906
2-Ethylbutyraldehyde	1.9289	1.8095	0.045	-0.3102	2.992	1.065	1.241	5.699	2.522
2,3,4-Trimethylpentane	1.7248	2.1542	0.08	1.1807	3.464	1	2.103	6.732	0.06245
cis-1,2-Dichloroethene	0.7732	1.0364	0.053	0.6931	1.975	1.159	0.429	4.414	6.77E-07
Propyl sulfide	3.4429	2.9439	0.084	-1.3821	2.447	1	2.091	5.889	1.909
Cycloheptane	2.4215	2.0869	0.057	-0.887	2.042	1	1.75	4.95	0.002945
1,1-dichloroethane	0.6875	0.5409	0.05	-0.3841	2.324	1	0	4.577	1.616
Ethylcyclohexane	2.9354	3.053	0.086	0.3263	2.125	1	2.302	5.82	0.04974
Dipropyl ether	1.1711	1.5341	0.041	0.9372	2.447	1	0.697	5.536	1.109
Isopropyl acetate	1.0243	1.2378	0.059	0.5671	2.953	1.061	0.402	5.862	1.819

Tableau I.12 : Valeurs des t_{rr} expérimentales, calculées/prédites, h_{ii}, $e_{iStd.}$ et les descripteurs sélectionnés de 122 COVs. (Suite)

Composé	t_{rrExp}	$t_{rrCalc/Pred}$	h_{ii}	$e_{iStd.}$	J	PCR	X3v	X0sol	u
2,3-Butanedione	0.7187	0.4966	0.079	-0.6092	2.993	1.126	0.496	5.155	0.000009541
Nitromethane	0.6669	0.4797	0.076	-0.5118	2.324	1.18	0	3.577	3.984
Cyclopropylcyanide	1.5194	1.1244	0.056	-1.0454	1.999	1.172	0.471	3.699	3.307
Allylsulfide	3.1745	2.715	0.032	-1.1706	2.447	1.174	1.414	5.889	1.956
1,2-Dichlorobenzene	4.3976	4.0146	0.101	-1.0894	2.279	1.378	1.58	6.983	1.351
1-Methylcyclohexene	2.0159	2.1358	0.032	0.3054	2.123	1.106	1.518	5.113	0.1644
Methanol	0.4742	0.0145	0.124	-1.3613	1	1	0	2	1.487
3-Pentanone	1.2444	1.2286	0.03	-0.0401	2.754	1.065	0.789	4.992	2.612
1-Bromopropane	0.8358	1.5342	0.036	1.7894	1.975	1	0.982	4.414	1.808
Allyl acetate	1.2808	1.5334	0.044	0.6557	2.678	1.147	0.404	5.699	1.805
Valeraldehyde	1.2435	1.5127	0.022	0.6754	2.339	1.1	0.676	4.828	2.565
trans-2-Heptene	1.3048	1.7228	0.033	1.0666	2.447	1.134	0.96	5.536	0.05275
3-Ethyl-2-pentene	1.3151	1.4933	0.037	0.4571	2.992	1.095	1.316	5.699	0.2448
Acetaldehyde	0.4794	0.2456	0.062	-0.6249	1.633	1.145	0	2.707	2.458
1-Heptyne	1.4709	1.7924	0.028	0.8142	2.447	1.16	0.925	5.536	0.37
2-Methyl-2-butanol	0.9019	0.7301	0.045	-0.4469	3.168	1	0.865	5.207	1.643
2-Bromopropane	0.6951	0.6183	0.047	-0.2005	2.324	1	0	4.577	2.043
3,3-Dimethylpentane	0.9475	1.5183	0.078	1.563	3.36	1	1.914	5.914	0.05973
Toluene	1.9918	2.1141	0.079	0.336	2.123	1.369	0.94	5.113	0.261
Diisopropyl ether	0.7539	1.1603	0.059	1.0807	2.953	1	0.544	5.862	1.31
p-Xylene	3.3243	2.7815	0.067	-1.4607	2.192	1.32	1.218	5.983	0.04861
Trimethylacetonitrile	0.948	0.8348	0.075	-0.309	3.168	1.127	0.335	5.207	3.318
Ethyl benzene	3.2463	2.88	0.076	-1.0003	2.125	1.371	1.251	5.82	0.3341
o-Xylene	3.5036	2.9953	0.076	-1.3881	2.279	1.378	1.426	5.983	0.4634
n-Butylbenzene	4.4965	4.3373	0.129	-0.4749	2.017	1.362	1.662	7.234	0.343

Tableau I.12 : Valeurs des t_{rr} expérimentales, calculées/prédites, h_{ii}, $e_{iStd.}$ et les descripteurs sélectionnés de 122 COVs. (Suite)

Composé	t_{rrExp}	$t_{rrCalc/Pred}$	h_{ii}	$e_{iStd.}$	J	PCR	X3v	X0sol	u
3-Ethyl-1-pentene	0.9522	1.4871	0.037	1.3729	2.992	1.065	1.382	5.699	0.1857
Cyclohexene	1.1043	1.3668	0.039	0.6761	2	1.11	1.158	4.243	0.175
sec.-Butylbenzene	4.2867	4.3828	0.11	0.2777	2.24	1.332	1.981	7.397	0.218
Methyl tert.-butyl ether	0.6647	0.4917	0.052	-0.4547	3.168	1	0.612	5.207	1.33
Isopropanol	0.5505	-0.2232	0.043	-2.0047	2.324	1	0	3.577	1.526
Propionitrile	0.6943	0.7763	0.063	0.2196	1.975	1.224	0.158	3.414	3.25
1-Chloropropane	0.6223	0.8148	0.029	0.4886	1.975	1	0.567	3.914	1.546
2-Butanone	0.7413	0.7675	0.03	0.0665	2.54	1.087	0.498	4.284	2.698
Ethanol	0.5115	-0.1445	0.065	-1.76	1.633	1	0	2.707	1.449
Bicyclo[2,1]hepta-2,5-diene	1.3051	2.0796	0.063	2.0716	2.119	1.175	1.755	4.69	0.09269
Methylcyclopentane	0.8379	1.4575	0.061	1.6523	2.184	1	1.644	4.406	0.03699
Crotonaldehyde	0.9543	1.2066	0.059	0.6704	2.191	1.272	0.271	4.121	3.163
Hexane	0.7363	1.1064	0.035	0.9475	2.339	1	0.957	4.828	1.18E-08
Trifluoromethyl-benzene	1.3526	2.347	0.072	2.7002	2.389	1.29	1.101	5.113	3.108
1-Hexene	0.7123	1.1551	0.027	1.1195	2.339	1.1	0.762	4.828	0.2527
2,2-Dimethylbutane	0.5941	0.5851	0.063	-0.024	3.168	1	1.061	5.207	0.06733
2-Methyl-2-propanol	0.5852	-0.2835	0.072	-2.3579	3.024	1	0	4.5	1.539
Acetonitrile	0.5439	0.5327	0.094	-0.0315	1.633	1.251	0	2.707	3.206
Methacrylonitrile	0.7548	0.8681	0.062	0.3025	2.54	1.243	0.191	4.284	3.239
Cyclopentane	0.6568	0.6276	0.079	-0.08	2.083	1	1.25	3.536	0.008769
Pentane	0.5498	0.5579	0.045	0.021	2.191	1	0.707	4.121	0.002103
1,4-Difluorobenzene	1.1153	1.1744	0.092	0.1658	2.192	1.32	0.804	4.243	2.406E-09
Acrylonitrile	0.5884	0.907	0.104	0.9107	1.975	1.349	0.091	3.414	3.251
1-Pentene	0.5392	0.609	0.04	0.18	2.191	1.112	0.493	4.121	0.2489
Hexafluorobenzene	0.7762	0.7234	0.12	-0.1552	2.76	1.255	1.156	4.243	6.479E-13

Tableau I.12 : Valeurs des t_{rr} expérimentales, calculées/prédites, h_{ii}, $e_{iStd.}$ et des descripteurs sélectionnés de 122 COVs. (Suite)

Composé	t_{rrExp}	$t_{rrCalc/Pred}$	h_{ii}	$e_{iStd.Err.Pr.}$	J	PCR	X3v	X0sol	u
Dichlorométhane	0.6001	0.5876	0.062	-0.0333	1.633	1	0	3.707	1.363
1,2-dichloropropane	1.2406	1.6593	0.031	1.0651	2.54	1	0.988	5.284	2.238
Ethyl sulfide	1.2571	1.546	0.04	0.7453	2.191	1	1.225	4.475	1.941
3,3-Diethylpentane	3.3426	2.86	0.167	-1.5408	3.825	1	3	7.328	0.08688
Heptane	1.2393	1.6995	0.037	1.1809	2.447	1	1.207	5.536	0.00162
Butyronitrile	1.0892	1.193	0.047	0.2709	2.191	1.199	0.362	4.121	3.303
1-Octyne	2.8834	2.3917	0.037	-1.2628	2.53	1.146	1.175	6.243	0.3728
Propyl formate	0.8293	1.5503	0.064	1.9305	2.339	1.1	0.39	4.828	3.905
1-Propanol	0.6508	0.1777	0.038	-1.217	1.975	1	0.224	3.414	1.452
Tetrahydrofuran	0.8477	0.6248	0.045	-0.5791	2.083	1	0.827	3.536	1.668
Isobutyronitrile	0.8441	0.8102	0.047	-0.0884	2.54	1.163	0.258	4.284	3.285
2-Hexyne	1.1062	1.1791	0.061	0.1945	2.339	1.247	0.552	4.828	0.05825
Propyl benzene	3.9369	3.5495	0.093	-1.089	2.078	1.368	1.382	6.527	0.3371
Diethyl ether	0.5592	0.551	0.028	-0.0208	2.191	1	0.408	4.121	1.148
Acetone	0.546	0.1284	0.039	-1.076	2.324	1.093	0	3.577	2.732
Cyclopentene	0.6371	0.5674	0.063	-0.1865	2.083	1.112	0.908	3.536	0.1482
4-Bromo-m-xylene	4.9433	4.9782	0.141	0.1064	2.346	1.333	2.203	7.853	1.363
Methyl propionate *	0.8665	0.903	0.025	0.0896	2.754	1.065	0.516	4.992	1.897
trans-1,2-dichloroethene *	0.6553	1.2357	0.026	1.4271	1.975	1.159	0.429	4.414	1.099
3,3-Dimethyl-2-butanone *	1.3367	1.3828	0.085	0.1168	3.541	1.051	1.056	6.077	2.709
sec.-Butanol *	0.7565	0.4875	0.025	-0.6609	2.54	1	0.591	4.284	1.471
2-Bromo-p-xylene *	4.9562	4.9124	0.134	-0.1143	2.346	1.333	2.182	7.853	1.083
2-Methyl-2-butene *	0.5762	0.4254	0.049	-0.3754	2.54	1.125	0.577	4.284	0.199
1,1-Dimethylcyclohexane *	2.245	2.921	0.068	1.6987	2.328	1	2.207	6.036	0.01134

Tableau I.12 : Valeurs des t_{rr} expérimentales, calculées/prédites, h_{ii}, $e_{iStd.}$ et des descripteurs sélectionnés de 122 COVs. (Suite et fin)

Composé	t_{rrExp}	$t_{rrCalc/Pred}$	h_{ii}	$e_{iStd.}$	J	PCR	X3v	X0sol	u
m-Xylene*	3.3081	2.7979	0.056	-1.2744	2.231	1.318	1.174	5.983	0.56
2-Methyl-1-pentene *	0.7097	0.9008	0.032	0.4712	2.627	1.079	0.677	4.992	0.384
1-Hexyne *	0.8176	1.2246	0.033	1.0041	2.339	1.178	0.675	4.828	0.3658
2-Chloropropane *	0.5565	0.1753	0.039	-0.9434	2.324	1	0	4.077	1.662
Propionaldehyde *	0.5513	0.5109	0.036	-0.0998	1.975	1.128	0.167	3.414	2.507
Butyraldehyde *	0.7281	0.9634	0.023	0.5778	2.191	1.112	0.407	4.121	2.544
3-Hexyne *	1.0247	1.1419	0.072	0.295	2.339	1.265	0.479	4.828	0.000001215
Cumene *	3.7404	3.4923	0.08	-0.6276	2.228	1.326	1.466	6.69	0.2374
2,4-Dimethylpentane *	0.8326	1.215	0.056	0.9547	2.953	1	0.943	5.862	0.03182
1-Heptene *	1.1771	1.7325	0.026	1.3657	2.447	1.089	1.012	5.536	0.2546
2-Pentanone *	1.1689	1.2633	0.027	0.2322	2.627	1.079	0.602	4.992	2.68
3-Ethylpentane *	1.1358	1.6196	0.048	1.2029	2.992	1	1.732	5.699	0.04307
2,2,4-Trimethylpentane *	1.1403	1.5	0.101	0.9203	3.389	1	1.021	6.784	0.0791
Ethyl acetate *	0.8026	0.9366	0.027	0.3296	2.627	1.079	0.348	4.992	1.884
1-Ethylcyclopentene *	1.8348	2.1591	0.034	0.8006	2.14	1.102	1.589	5.113	0.1441
4-Methylcyclohexene*	1.6085	2.1557	0.032	1.3494	2.123	1.113	1.531	5.113	0.1674
2-Methyl-1-propanol *	0.8618	0.3096	0.029	-1.3597	2.54	1	0.365	4.284	1.385
1,3-Dichlorobenzene *	4.2362	3.6631	0.096	-1.4624	2.231	1.318	1.257	6.983	0.8795
Propyl acetate *	1.4114	1.516	0.04	0.2589	2.678	1.072	0.509	5.699	1.884
1,2-dichloroethane*	0.9423	0.9629	0.05	0.0514	1.975	1	0.643	4.414	1.405E-11
1-Butanol *	1.0562	0.6744	0.024	-0.9375	2.191	1	0.512	4.121	1.417
2-Bromopentane*	2.2239	2.1587	0.043	-0.1617	2.627	1	1.144	5.992	1.971
1,1,2-Trichloro-ethane *	2.0898	1.8601	0.033	-0.5668	2.54	1	1.05	5.784	1.038

I-2-1-5-Domaine d'application du modèle RLM :

Avant qu'un modèle QSRR ne soit exploité, le domaine d'application doit être défini pour que la prédiction des composés qui tombent dans ce domaine puisse être considérée comme fiable.

Le domaine d'application du modèle RLM a été analysé dans le cadre du diagramme de Williams qui représente les résidus de prédiction standardisés e_{istd} en fonction des valeurs des leviers h_{ii}.

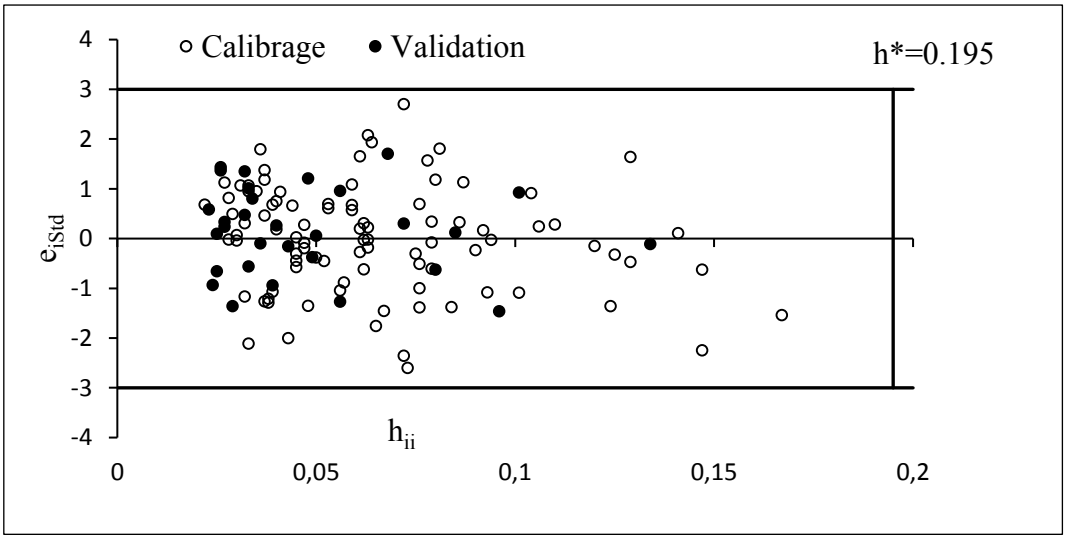


Figure I.14 : Diagramme de Williams.

Toutes les valeurs de leviers h_{ii} sont inférieures à la valeur critique :

$$h^* = \frac{3(p+1)}{n} = 0,195 \qquad (2.5)$$

Tous les résidus standardisés (e_{istd}) sont compris entre les limites ± 3.

I-2-1-6-Qualité de l'ajustement :

La figure I.15 reproduit les valeurs calculées $t_{rr\ Cal}$ et prédits $t_{rr\ Pred}$ en fonction de celles expérimentales $t_{rr\ Exp}$ pour l'ensemble de calibrage et de validation.

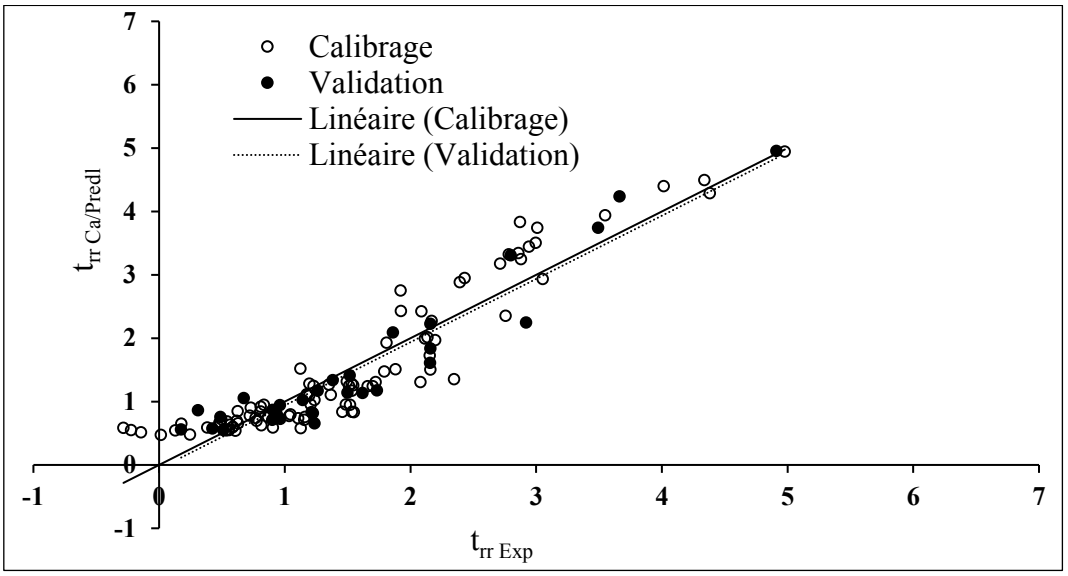


Figure I.15 : Droite d'ajustement du modèle.

D'après cette figure I.15 : on remarque une faible dispersion autour de la droite d'ajustement. Ce qui montre la faiblesse des erreurs lors du calcul (calibrage) et de la prédiction (validation) ; ce qui traduit un bon ajustement.

La figure I.16 qui représente les paramètres statistiques Q^2 et R^2 permet de comparer les résultats obtenus pour les modèles randomisés (carrés) au modèle réel de départ (cercle).

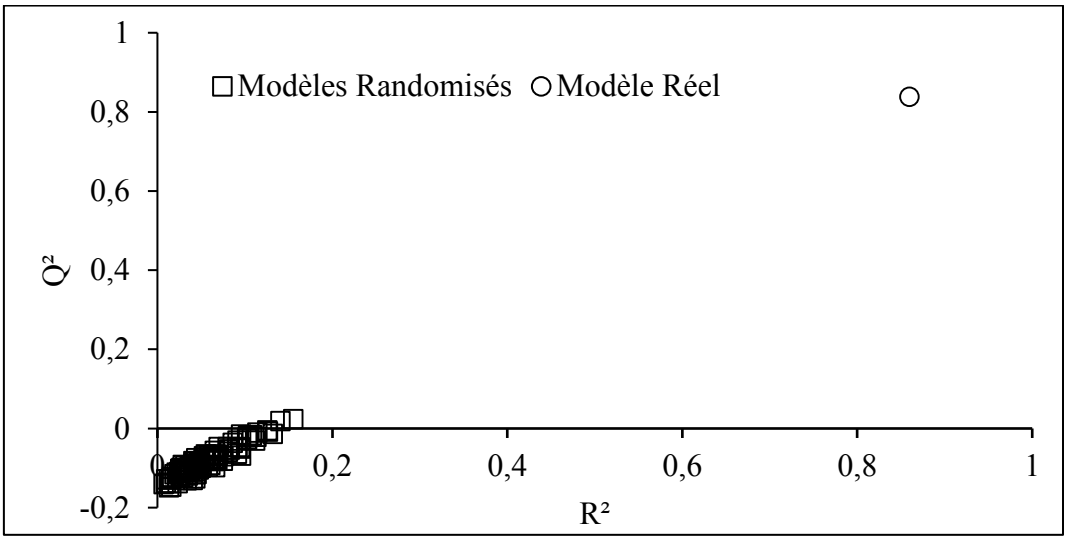


Figure I.16 : Test de randomisation associé au modèle QSAR.

Il est clair que les statiques obtenues pour les vecteurs modifiés de temps de rétention relatif sont plus petites que celles du modèle QSAR réel et pour la majeure partie on obtient même un $Q^2 \leq 0{,}0234$ et un $R^2 \leq 0{,}155$ ce qui montre que le modèle n'est pas dû au hasard.

I-2-2-Machine à vecteur support

Une régression SVM a été utilisée pour développer le modèle sur les composés de l'ensemble de calibrage, sur la base du même sous-ensemble et les mêmes descripteurs.

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 (RMSE) liée au meilleur paramètre de régression en utilisant le leave one-out (LOO) en tenant compte du RMSE de l'ensemble de test, les valeurs des résultats de la régression sont : *c=17.02 y=0.2 ε =0.2*

$$R^2 = 0,9570 \qquad Q^2_{Loo} = 0,9310 \qquad Q^2_{Ext} = 0,9560$$

$$RMSE = 0,3273 \qquad RMSE_{Ext} = 0,3401$$

Les t_{rr} observés et prédits de l'ensemble de calibrage et de l'ensemble de validation sont présentés dans la figure I.17. Les valeurs calculées sont, en général, en bon accord avec les valeurs expérimentales correspondantes.

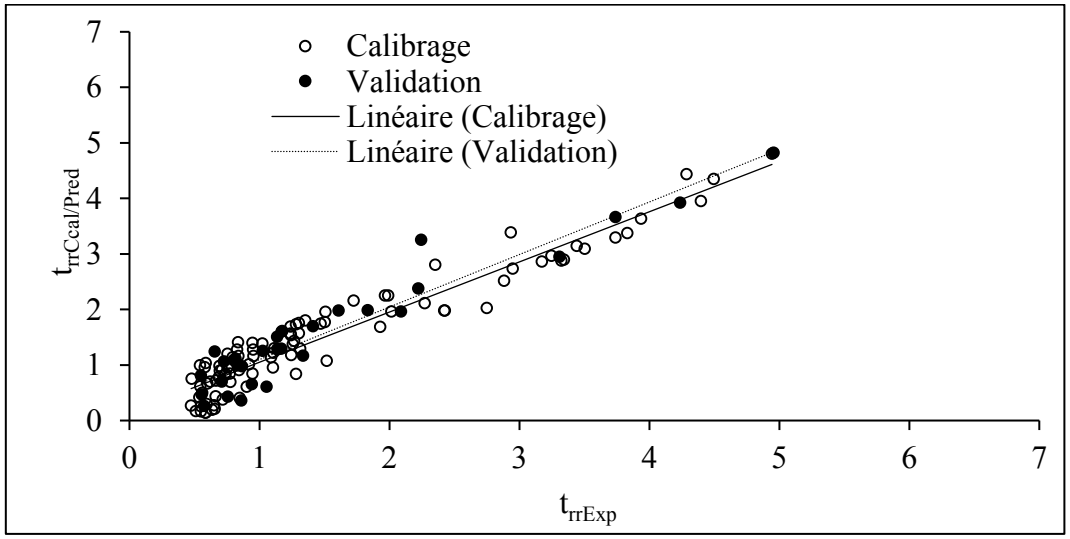


Figure I.17 : Graphe des valeurs calculées, prédites en fonction des valeurs expérimentales.

I-2-3-Les réseaux de neurones artificiels (RNA) :

Le choix du nombre de neurones de la couche cachée est fixé à 4 et le nombre d'itérations à 400. Le graphe suivant explicite ce choix.

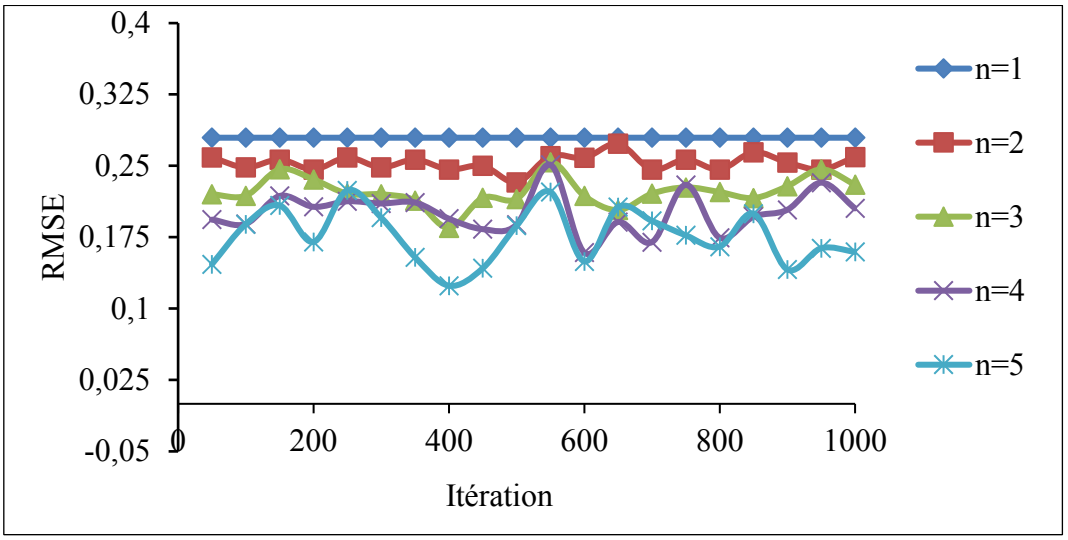


Figure I.18 : Choix du nombre de neurones de la couche cachée.

La structure optimale adoptée est reproduite dans le tableau I.13.

Tableau I.13 : Structure optimale adoptée pour le réseau de neurones.

Entrées	05 (les descripteurs)
Sortie	01 (t_{rr})
Couche cachée	Une couche cachée
Nombre de neurones dans la couche cachée	04
Algorithme d'apprentissage	Rétro propagation du gradient de l'erreur
Fonction d'apprentissage	Tangente hyperbolique (couche cachée)
	Linéaire (couche de sortie)

$R^2 = 0,97$ $RMSE = 0,194$ $RMSE_{Ext} = 0,291$ $Q^2_{Ext} = 0,93$

avec : $n_{tr} = 92$ $n_{test} = 30$

La validation statistique externe (Q^2_{Ext}) atteste de la bonne capacité prédictive des composés n'ayant pas participé au calcul du modèle.

La figure I.19 suivante représente les valeurs t_{rr} théoriques (calculées/prédites) en fonction des valeurs expérimentales (mesurées).

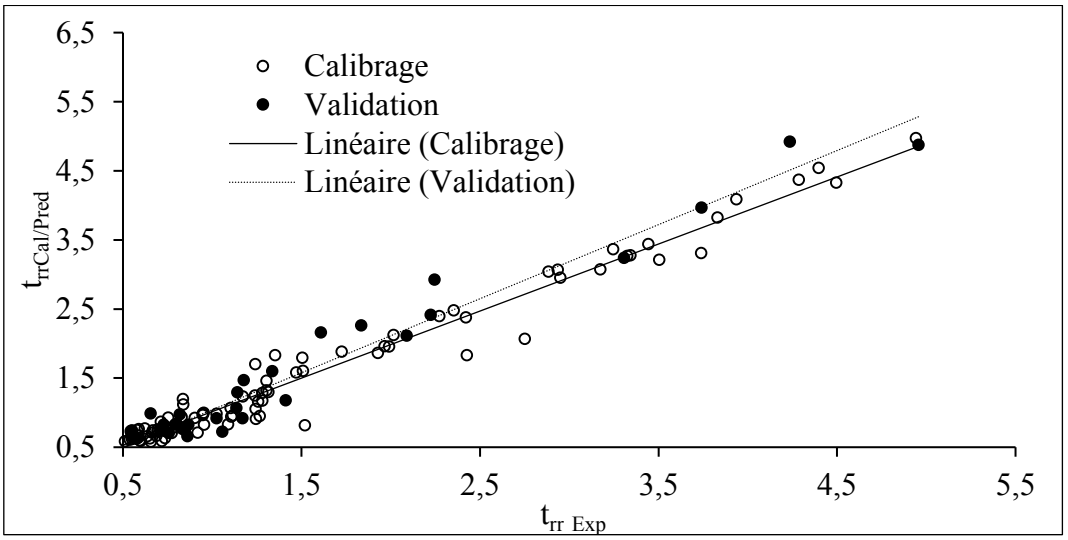


Figure I.19 : Graphe des valeurs calculées en fonction des valeurs expérimentales.

Les t_{rr} observés et prédits de l'ensemble de calibrage (cercles vides) et de l'ensemble de validation (cercles noirs) sont présentés dans la Figure I.19. Les valeurs calculées sont, en général, en bon accord avec les valeurs expérimentales correspondantes.

I-2-4-Contribution des descripteurs et interprétation :

En se basant sur une procédure décrite dans la littérature [162,163], les contributions relatives des cinq descripteurs du modèle ont été déterminées. Elles diminuent selon l'ordre suivant : Xosol (23, 71%) > J (22, 57%) > X3v (19, 58%) > u (17, 39%) > PCR (16, 72 %). Il convient de noter que la différence dans les contributions au modèle de deux descripteurs utilisés dans le modèle n'est pas significative, ce qui prouve que les cinq descripteurs sont indispensables pour générer le modèle prédictif.

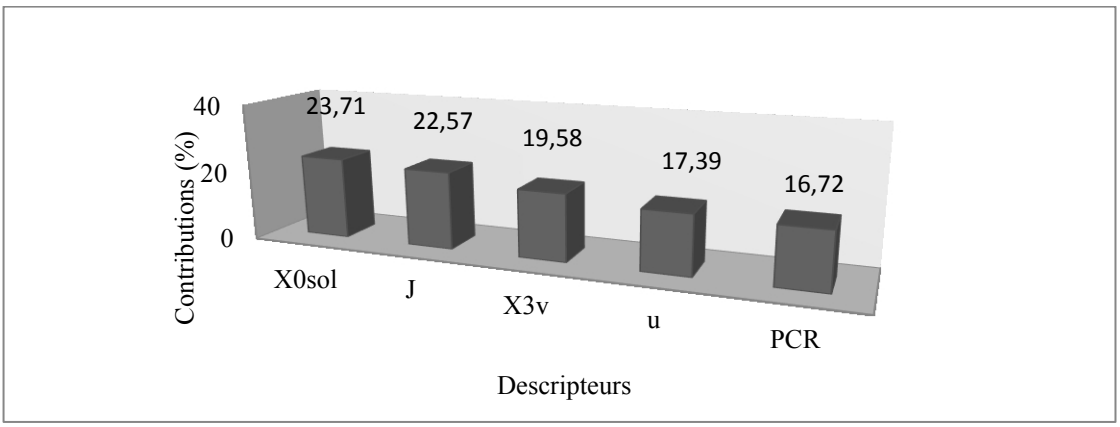


Figure I.20 : Contributions (%) des descripteurs du modèle RLM.

I-2-5-Comparaison des résultats des méthodes RLM, SVM et RNA : (Tableau I.14)

Trois méthodes RLM, SVM et RNA ont été utilisées pour prédire le temps de rétention relatif (t_{rr}) des composés organiques volatils. Le modèle a été développé par une sélection par algorithme génétique de descripteurs moléculaires théoriques parmi un large éventail obtenu avec le logiciel Dragon.

Les données ont été séparées à l'aide de l'algorithme Kennard and Stone (CADEX) en deux sous ensembles de 92 éléments pour le calibrage et 30 pour la validation externe.

Les modèles proposés (RLM, SVM et RNA) sont stables, robustes et prédictifs. L'approche par réseaux de neurones conduit au meilleur modèle à tous les points de vue : capacités prédictives interne et externe, qualité de l'ajustement…, ce qui prouve dans ce cas que les corrélations variable dépendante / variables explicatives sont fondamentalement non linéaires et il en résulte donc que la meilleure approche QSRR est basée sur la méthode RNA.

Tableau I.14 : Comparaison des résultats des méthodes : RLM, RNA et SVM.

Calibrage ; n= 92	Validation ; n= 30		
Méthodes/Statistiques	RLM	RNA	SVM
R^2	87, 39%	97, 01%	95, 30%
Q^2_{Ext}	90, 12%	93, 24%	95, 60%
SDEP	0, 398	0, 194	0, 3188
$SDEP_{Ext}$	0, 427	0, 2916	0, 3401
$(R^2-R^2_0)/ R^2<0, 1$	-0, 1274	-0, 1030	-0, 089
$(R^2-R'^2_0)/ R^2<0, 1$	-0, 1172	-0, 1055	-0, 094
0, 85 < k< 1, 15	1, 0092	1, 0316	0, 959
0, 85< k'< 1, 15	0, 9500	-0, 993	0, 101

I-3-Conclusion

Nous avons établi trois modèles QSRR basés sur trois approches RLM, RNA et SVM. Toutes ces approches ont été utilisées pour prédire le temps de rétention relatif de quelques composés organiques volatils. Les modèles ont été développés par un algorithme génétique

sélectionnant des descripteurs moléculaires théoriques parmi un large ensemble obtenu avec le logiciel Dragon.

L'ensemble de données est séparé en utilisant l'algorithme de Kennard et Stone en deux sous-ensembles de 92 éléments pour le calcul du modèle et 30 éléments pour la validation externe.

Les modèles proposés (MLR, RNA et SVM) sont stables, robustes et prédictifs. Le domaine d'applicabilité chimique du modèle MLR étudié et la fiabilité des prédictions ont été vérifiés par l'approche ‘‘effet de levier’’.

Chapitre II Modélisation du coefficient de partage octanol/eau

II-1-Introduction

La lipophilie d'une molécule caractérise son aptitude à se distribuer dans un système biphasique soit liquide-liquide soit solide-liquide. Le coefficient de partage dans le système *n*-octanol/eau est connu depuis longtemps comme étant un des paramètres physico-chimiques quantitatifs qui est le mieux corrélé à l'activité des molécules organiques [165]. Le coefficient de partage n-octanol / eau est le rapport de la concentration d'un produit chimique dans le n-octanol à celle dans l'eau du système à deux phases à l'équilibre [165].

Les propriétés physico-chimiques d'un composé chimique organique jouent un rôle important dans la détermination de sa distribution et de son devenir dans l'environnement.

L'élaboration des modèles mathématiques QSPR/QSAR reliant les propriétés physicochimiques et les activités biologiques à la structure moléculaire permet, d'une part, d'expliquer l'origine de ces activités/propriétés et, d'autre part, de les prédire pour des molécules pour lesquelles les données expérimentales ne sont pas disponibles [166].

Le but est de trouver un modèle statistique pour la prédiction du coefficient n-octanol / eau (K_{ow}) de quelques composés organiques volatils.

Le modèle QSPR a été construit en utilisant la régression linéaire multiple (RLM). Le modèle obtenu montre quels descripteurs jouent un rôle important dans la variation de K_{ow} de ces composés chimiques.

II-2-Résultats et discussion

Les 64 composés sont des produits chimiques organiques volatils. Les valeurs observées de leurs K_{ow} respectifs on été prélevées dans la littérature [167].

L'application de la procédure GA-VSS conduit à plusieurs modèles pour la prédiction de K_{ow} en fonction de différents ensembles de descripteurs moléculaires.

La performance du modèle est décrite à l'aide des paramètres liés à la capacité prédictive du modèle (Q^2_{LOO}, Q^2_{LMO}) et la capacité d'ajustement (R^2). Les déviations standards des erreurs de prédiction (SDEP) et de calcul (SDEC) sont également rapportées.

Le meilleur modèle obtenu en utilisant 48 éléments de calibrage est un modèle à deux dimensions :

$$K_{ow} = - 2,\ 01 \pm (0,260) + (0,\ 189\ \pm (0,01))\ MR - (0,\ 412\ \pm (0,04))Chi0\_EA\ (dm) \qquad (2.II.99)$$

La réfractivité moléculaire, notée (MR), en m^3/mol, est le volume de la substance absorbée par mole de cette substance. Elle est définie selon Lorentz-Lorenz [168] par la formule suivante :

$$MR = \frac{n^2-1}{n^2+2}\frac{MW}{d} = \frac{n^2-1}{n^2+2}MV \qquad (2.II.100)$$

Où : MW est le poids moléculaire ; d est la densité ; n est l'indice de réfraction ; MV est le volume molaire.

La réfractivité moléculaire est également proportionnelle à la polarisabilité αe, selon la relation suivante [169] :

$$MR = 4/3\pi\ NA\ \alpha e \qquad (2.II.101)$$

Où : NA est le nombre d'Avogadro qui est, le nombre de molécules dans une mole de substance NA = 6, 022 140 10 ^{23}mol^{-1}.

Chi0_EA (dm): Indice semblable à la connectivité d'ordre 0 du tapis d'adjacence de bord pondéré par moment dipolaire.

Les paramètres statistiques du modèle sont rapportés ci-aprés :

$R^2 = 87, 58\ \%$ $Q^2_{LOO} = 86, 22\%$ $Q^2_{Boot} = 85, 28\%$ $Q^2_{Ext} = 90, 02\ \%$

$n_{test} = 16$ $n_{tr} = 48$

$SDEP_{Ext} = 0, 426$ $SDEP = 0, 5$ $SDEC = 0, 475$ $S = 0, 49$

$K_{XX} = 35, 94$ $K_{XY} = 47, 34$ $s = 0, 49$ $F = 158, 72$

La base de données utilisée a été scindée à l'aide de l'algorithme Duplex en deux sous-ensembles respectivement de 48 éléments pour la sélection des variables explicatives puis le calcul du modèle et de 16 éléments pour la validation externe.

Les paramètres d'ajustement et de validation rapportés ont des valeurs élevées indiquant que le modèle réalise une très bonne performance prédictive et que les descripteurs impliqués décrivent très bien le coefficient de partage K_{ow}.

Le modèle est très hautement significatif (grande valeur du paramètre de Fisher : F=158,72). Le coefficient de prédiction Q² supérieur à 86,22%, indique un modèle robuste, c'est-à-dire dont les paramètres ne changent pas beaucoup lorsqu'on utilise d'autres ensembles de calibrage extraits de la population totale.

La matrice de corrélation (Tableau II.15) suggère que ces descripteurs sont faiblement corrélés entre eux. Ainsi, le modèle peut être considéré comme une équation de régression optimale.

Tableau II.15 : Matrice de corrélation.

	K_{ow}	réfractivité
réfractivité	0, 824	
	0, 000	
Ci0_EA (dm)	-0, 122	0, 359
	0, 403	0, 011

Les valeurs absolues élevées de t indiquées dans le tableau II.16 expriment que les coefficients de régression des descripteurs impliqués dans le modèle sont significativement plus grands que l'écart-type. Les valeurs de la probabilité (p) de t pour chaque descripteur sont très faibles, ce qui indique que chacun des descripteurs est très significatif.

Tableau II.16 : Caractéristiques des descripteurs sélectionnés dans le meilleur modèle RLM.

Descripteurs	x	Dx	t	p	VIF
Constante	-2,0125	0,2603	-7,73	0,000	
Réfractivité	0,18944	0,01043	18,16	0,000	1,148
Chi0_ EA (dm)	-0,41205	0,04708	-8,75	0,000	1,148

Tableau II.17 : Valeurs des K_{ow} observées et des deux descripteurs sélectionnés des 48 COVs de calibrage.

Composé	$K_{ow\ Obs}$	MR	Chi0_EA (dm)	Composé	K_{ow}	MR	Chi0_EA (dm)
1-Butanol	0, 88	22, 135	1, 078	Methyltert,-butyl ether	0, 94	26, 816	5, 392
1-Bromobutane	2, 75	28, 169	0, 822	Isopropanol	0, 05	17, 428	2, 157
1-Bromopentane	3, 37	32, 77	0, 822	Propionitrile	0, 16	16, 455	2, 171
2-Hexanone	1, 38	30, 024	3, 21	1-Chloropropane	2, 04	20, 585	0, 801
cis-1,2-Dichloroethene	1, 48	19, 642	0, 566	2-Butanone	0, 29	20, 822	3, 21
1,1-dichloroethane	1, 79	21, 28	2, 471	Ethanol	-0, 22	13, 009	1, 078
Dipropyl ether	2, 03	31, 557	4, 313	Methylcyclopentane	3, 37	27, 554	0
2-Methyl-1-propanol	0, 76	22, 005	1, 078	2-Chloropropane	1, 9	20, 479	1, 601
1,2-Dichlorobenzene	3, 43	35, 668	2, 167	1-Hexyne	2, 73	27, 87	1, 644

Tableau II.18 : Valeurs des K_{ow} observés et les deux descripteurs sélectionnés des 16 COVs de l'ensemble de validation.(Suite)

Composé	$K_{ow\ Obs}$	MR	Chi0_EA (dm)	Composé	K_{ow}	MR	Chi0_EA (dm)
3-Pentanone	0, 99	25, 449	4, 422	1-Hexene	3, 39	29, 452	2, 425
1-Bromopropane	2, 1	23, 568	0, 822	2, 2-Dimethylbutane	3, 82	29, 23	0
trans-1,2-Dichloroethene	2, 09	19, 642	0, 566	m-Xylene	3, 2	36, 14	4, 851
sec,-Butanol	0, 61	21, 952	2, 157	2-Methyl-2-propanol	0, 35	22, 065	3, 235
2-Pentanone	0, 91	25, 423	3, 21	Acetonitrile	-0, 34	11, 931	1, 349
2-Bromopropane	2, 14	23, 462	1, 644	Cyclopentane	3	23, 005	0
1-Heptene	3, 99	34, 053	2, 425	Pentane	3, 39	24, 807	0
Toluene	2, 73	31, 099	2, 425	Acrylonitrile	0, 25	16, 346	2, 392
Diisopropyl ether	1, 52	31, 345	6, 47	Hexafluorobenzene	2, 54	27, 356	3, 453
p-Xylene	3, 15	36, 14	4, 851	Heptane	4, 66	34, 009	0
Ethyl benzene	3, 15	35, 7	3, 638	Butyronitrile	0, 53	21, 056	2, 171
o-Xylene	3, 12	36, 14	3, 283	Propyl formate	0, 83	22, 613	3, 789
Cumene	3, 66	40, 249	4, 851	1-Propanol	0, 25	17, 534	1, 078
n-Butylbenzene	4, 38	44, 902	3, 638	Tetrahydrofuran	0, 46	20, 553	4, 313
Butyraldehyde	0, 88	20, 946	3, 071	Diethyl ether	0, 89	22, 509	4, 313

Tableau II.18 : Valeurs des K_{ow} observés et les deux descripteurs sélectionnés des 16 COVs de l'ensemble de validation.

Composé	$K_{ow\ Obs}$	MR	Chi0_EA (dm)	Composé	K_{ow}	MR	Chi0_EA (dm)
Acetone	-0, 24	16, 195	1, 997	1,3-Dichlorobenzene	3, 53	35, 668	3, 203
Dibromomethane	1, 88	22, 072	1, 644	Methanol	-0, 77	8, 261	0
1,1,2-Trichloro-ethane	1, 89	25, 71	2, 328	Ethyl acetate	0, 73	22, 161	4, 02
1,1,2-Trichloro-1,2,2-trifluoroethane	1, 65	27, 374	3, 921	Cyclohexene	2, 86	28, 723	3, 283
Ethyl iodide	2	24, 391	0, 88	Propionaldehyde	0, 59	16, 345	3, 071

Tableau II.18 : Valeurs des K_{ow} observés et les deux descripteurs sélectionnés des 16 COVs de l'ensemble de validation. (Suite)

Composé	$K_{ow\ Obs}$	MR	Chi0_EA (dm)	Composé	K_{ow}	MR	Chi0_EA (dm)
Ethylene dichloride	1, 48	20, 656	0, 566	Hexane	3, 9	29, 408	0
Cycloheptane	4	32, 207	0	dichlorométhane	1, 25	16, 438	1, 826
Propyl acetate	1, 24	26, 685	4, 02	Propyl benzene	3, 72	40, 301	3, 638

Les valeurs des K_{ow} expérimentales, calculées et prédites pour l'ensemble de calibrage, ainsi que les valeurs des leviers et des erreurs standardisées sont regroupées dans le tableau II.19.

Tableau II.19 : Valeurs des K_{ow} expérimentales, calculées, prédites, leviers et résidus standardisés de l'ensemble de validation.

Composé	$K_{ow\ Exp.}$	K_{owCalc}	h_{ii}	$e_{iStd.}$	Composé	$K_{owExp.}$	$K_{ow\ Calc}$	h_{ii}	$e_{iStd.}$
1-Butanol	0,88	1,7517	0,036	1,8794	Methyl tert,-butyl ether	0,94	0,8585	0,1	-0,1945
1-Bromobutane	2,75	2,9877	0,052	0,5251	Isopropanol	0,05	0,4266	0,052	0,8323
1-Bromopentane	3,37	3,85	0,082	1,1141	Propionitrile	0,16	0,2385	0,06	0,1757
2-Hexanone	1,38	2,3553	0,03	2,082	1-Chloropropane	2,04	1,5748	0,045	-1,0169
cis-1,2-Dichloroethene	1,48	1,4945	0,053	0,0322	2-Butanone	0,29	0,6306	0,045	0,7445
1,1-dichloroethane	1,79	1,0197	0,031	-1,6481	Ethanol	-0,22	0,0412	0,09	0,6133
Dipropyl ether	2,03	2,19	0,053	0,3543	Methylcyclopentane	3,37	3,2097	0,082	-0,3714
2-Methyl-1-propanol	0,76	1,7273	0,036	2,0862	2-Chloropropane	1,9	1,2266	0,034	-1,4468
1,2-Dichlorobenzene	3,43	3,8412	0,07	0,9348	1-Hexyne	2,73	2,5943	0,03	-0,2899
3-Pentanone	0,99	1,0004	0,06	0,0233	1-Hexene	3,39	2,5703	0,027	-1,7414
1-Bromopropane	2,1	2,1253	0,041	0,055	2,2-Dimethylbutane	3,82	3,5239	0,091	-0,6976
trans-1,2-Dichloroethene	2,09	1,4945	0,053	-1,3185	m-Xylene	3,2	2,8281	0,088	-0,8708
sec,-Butanol	0,61	1,2745	0,027	1,4129	2-Methyl-2-propanol	0,35	0,8533	0,039	1,0896
2-Pentanone	0,91	1,493	0,027	1,2399	Acetonitrile	-0,34	-0,2721	0,101	0,1625
2-Bromopropane	2,14	1,7681	0,026	-0,7894	Cyclopentane	3	2,3571	0,068	-1,4569

1-Heptene	3,99	3,4326	0,052	-1,2316	Pentane	3,39	2,6949	0,071	-1,5833

Tableau II.19 : Valeurs des K_{ow} expérimentales, calculées, prédites, leviers et résidus standardisés de l'ensemble de validation.(Suite)

Composé	$K_{ow\ Exp.}$	K_{owCalc}	h_{ii}	$e_{iStd.}$	Composé	$K_{owExp.}$	$K_{ow\ Calc}$	h_{ii}	$e_{iStd.}$
Toluene	2,73	2,879	0,033	0,3197	Acrylonitrile	0,25	0,1274	0,064	-0,276
Diisopropyl ether	1,52	1,265	0,154	-0,6688	Hexafluorobenzene	2,54	1,7555	0,03	-1,6742
p-Xylene	3,15	2,8281	0,088	-0,7537	Heptane	4,66	4,4196	0,134	-0,609
Ethyl benzene	3,15	3,2435	0,062	0,2099	Butyronitrile	0,53	1,1009	0,031	1,2201
o-Xylene	3,12	3,4717	0,064	0,7919	Propyl formate	0,83	0,7287	0,05	-0,2234
Cumene	3,66	3,5983	0,12	-0,1526	1-Propanol	0,25	0,8893	0,053	1,4159
n-Butylbenzene	4,38	4,9682	0,17	1,5855	Tetrahydrofuran	0,46	0,1275	0,083	-0,7726
Butyraldehyde	0,88	0,7109	0,041	-0,3675	Diethyl ether	0,89	0,4941	0,07	-0,8997

Tableau II.19 : Valeurs des K_{ow} expérimentales, calculées, prédites, leviers et résidus standardisés de l'ensemble de validation. (Suite)

Composé	$K_{owExp,}$	K_{owCalc}	h_{ii}	$e_{iStd.}$
Acetone	-0,24	0,2612	0,06	1,0544
Dibromomethane	1,88	1,5076	0,029	-0,7708
1,1,2-Trichloro-ethane	1,89	1,9087	0,021	0,0386
1,1,2-Trichloro-1,2,2-trifluoroethane	1,65	1,5668	0,039	-0,1731
Ethyl iodide	2	2,2558	0,04	0,5324
Ethylene dichloride	1,48	1,6846	0,05	0,4282
Cycloheptane	4	4,0818	0,114	0,1773
Propyl acetate	1,24	1,3971	0,043	0,3275
1,3-Dichlorobenzene	3,53	3,416	0,059	-0,2397
Methanol	-0,77	-0,4063	0,154	0,8067
Ethyl acetate	0,73	0,5491	0,06	-0,3805
Cyclohexene	2,86	2,0815	0,028	-1,6106
Propionaldehyde	0,59	-0,1515	0,076	-1,5736
Hexane	3,9	3,5572	0,091	-0,7334
dichlorométhane	1,25	0,3769	0,057	-1,834

Propyl benzene	3,72	4,1058	0,104	0,8315

II-2-1-Qualité de l'ajustement :

La figure II.21 reproduit pour la totalité des données les valeurs calculées des K_{ow} en fonction des valeurs observées, elle montre une faible dispersion autour de la droite d'ajustement.

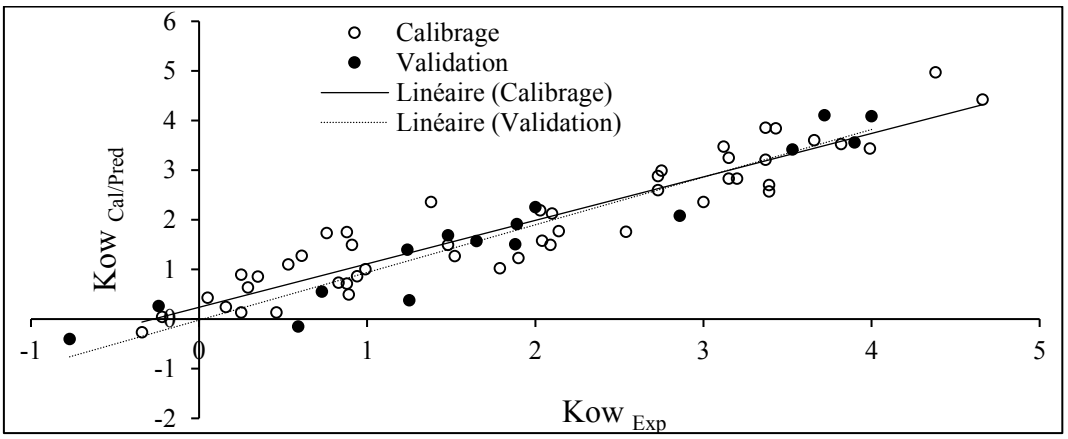


Figure II.21: Graphe des valeurs calculées des K_{ow} en fonction des valeurs observées.

II-2-2-Domaine d'application du modèle :

Le domaine d'application a été discuté à l'aide du diagramme de Williams (figure II.22) qui représente les résidus de prédiction standardisés en fonction des valeurs des leviers (hi), ième terme diagonal de la matrice de projection : $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\ \boldsymbol{X}'$ où $\boldsymbol{X}$ est la matrice des valeurs observées des variables explicatives et $\boldsymbol{X}'$ sa transposée, ces valeurs sont utiles pour le calcul des résidus caractéristiques.

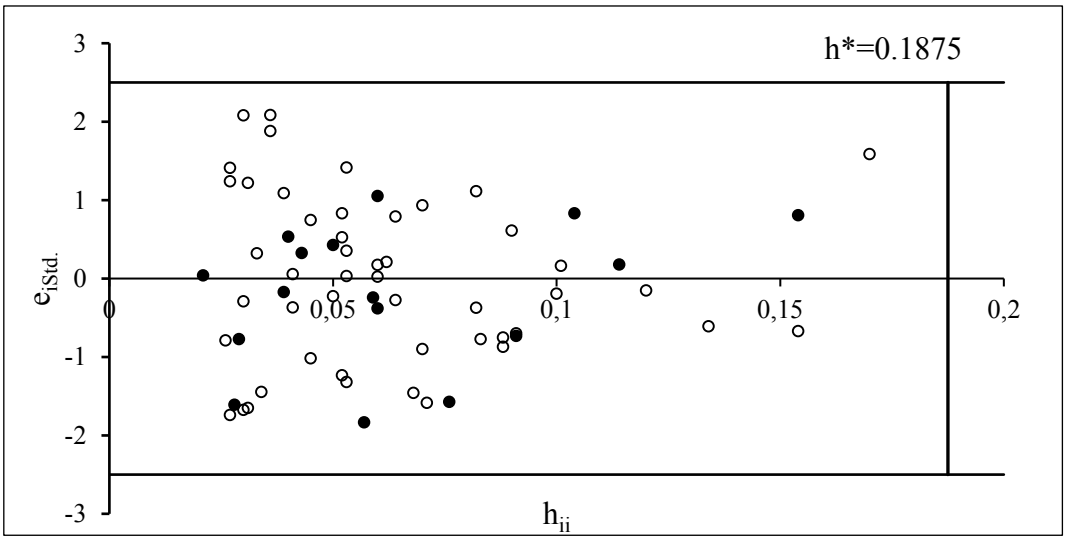


Figure II.22 : Diagramme de Williams.

Tous les points présentent un levier inférieur à la valeur critique h* = 0,1875 représentée par la ligne droite verticale. Tous les résidus de prédiction standardisés sont compris entre les limites ± 2,5.

II-2-3-Test de randomisation :

Dans le but d'établir que le modèle obtenu n'est pas dû au hasard ou à une sur-spécification, nous avons appliqué le test de randomisation. La figure II.23 fait ressortir des statistiques faibles (Q^2(%) < 20; R^2(%) < 30) pour les vecteurs modifiés alors que le point représentatif du modèle réel qui est isolé dans le graphe présente de bons paramètres statistiques ce qui garantit l'existence d'une relation multilinéaire entre K_{ow} et les descripteurs sélectionnés.

Les statistiques des vecteurs modifiés du coefficient de partage sont plus petites que celles du modèle réel.

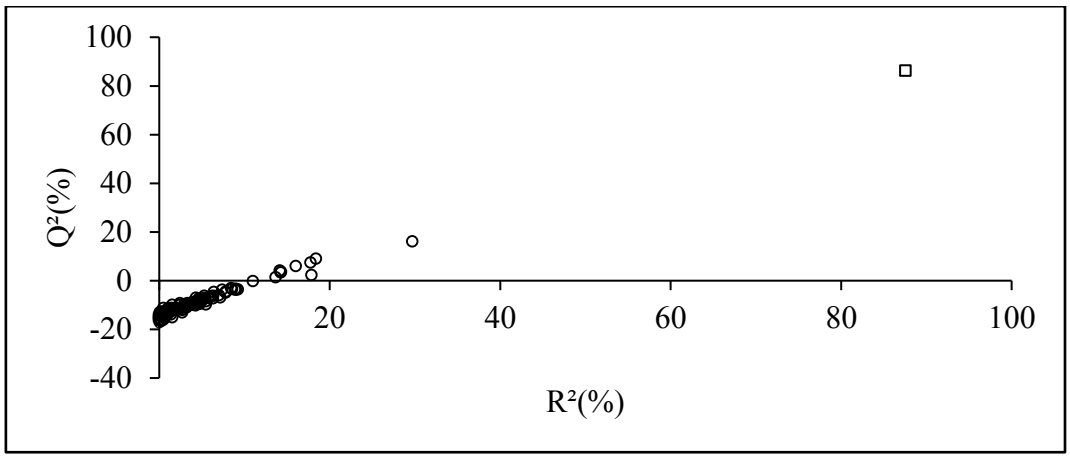


Figure II.23 : Test de randomisation.

II-3-Conclusion

La méthode RLM a été utilisée pour prédire le coefficient de partage octanol / eau (K_{ow}) d'une série de COVs. Le modèle a été développé par une sélection par algorithme génétique des descripteurs moléculaires théoriques parmi un large éventail obtenu avec plusieurs logiciels. Les données ont été séparées à l'aide de l'algorithme DUPLEX en deux sous ensembles de 48 éléments pour la construction du modèle et 16 pour la validation externe. Le modèle proposé est stable, robuste et prédictif. Le domaine d'applicabilité chimique du modèle RLM étudié et la fiabilité des prédictions ont été vérifiés par l'approche des leviers.

Chapitre III Modélisation de la pression de vapeur

III-1- Introduction

Les COVs sont largement utilisés en raison de leur capacité à s'évaporer dans l'air. Leur taux d'évaporation est grossièrement proportionnel à leur pression de vapeur. Lorsqu'une évaporation rapide est requise (pulvérisation de peinture par exemple) le solvant utilisé aura une pression de vapeur importante à température ambiante. Si une évaporation plus lente est requise (nettoyage de pièces mécaniques), le solvant utilisé aura une pression de vapeur plus basse à température ambiante [170].

La connaissance de la pression de vapeur est nécessaire pour la prédiction de l'évolution des polluants (COV) dans l'environnement.

Les méthodes QSPR sont souvent utilisées pour estimer les propriétés physicochimiques des composés organiques et prédire leur comportement dans l'environnement. Des méthodes chimiométriques peuvent être utilisées pour décrire la manière dont les propriétés physicochimiques varient en fonction des caractéristiques de la structure moléculaire exprimées en termes de descripteurs moléculaires appropriés. Les modèles QSPR peuvent également donner un aperçu général de la structure moléculaire qui influe sur ces propriétés [171].

III- 2- Résultats et discussion :

III-2-1-Régression linéaire multiple

Les valeurs observées des Pv de 51 composés organiques volatils ont été extraites de la littérature [167].

La régression linéaire multiple (MLR) a été utilisée dans les études QSPR. La sélection par algorithme génétique conduit à un bon modèle MLR à quatre descripteurs qui décrit au mieux la pression de vapeur. Le modèle retenu a pour équation :

$$\log Pv = 11,0 \pm (0,628) - 0,46 \pm (0,046) X0sol - 12,332 \pm (1,210) SpPosA\_H2 + 1,137 \pm (0,142) GATS2e - 1,23 \pm (0,127) Hy \quad (1.III.102)$$

$R^2 = 90,09\%$ $Q^2_{LOO} = 87,48\%$ $Q^2_{BOOT} = 85,55\%$ $SDEP = 0,256$

$SDEC = 0,227$ $K_{xx} = 38,51$ $K_{xy} = 45,57$

$n_{tr} = 39$ $s = 0,24$ $F = 77,25$ $n_{test} = 12$ $Q^2_{Ext} = 83,07$ $SDEP_{Ext} = 0,297$

La valeur de R^2 indique que 90,09 (%) de la variation totale est expliquée par le modèle, alors que la valeur élevée de Q^2_{Loo}, qui diffère peu de celle de R^2, renseigne sur la robustesse du modèle.

Tableau III.20 : Valeurs des log Pv mesurés et des descripteurs sélectionnés.

N°	Composés	log Pv	X0sol	SpPosA_H2	GATS2e	Hy
1	1,1,2-Trichloro-ethane	3,48	5,784	0,541	1,078	-0,359
2	Iodomethane	4,72	3,5	0,5	0,97	-0,315
3	Ethyl iodide	4,12	4,207	0,515	1,071	-0,528
4	1-Bromobutane	3,72	5,121	0,57	0,95	-0,719
5	2-Hexanone	2,71	5,699	0,56	0,534	-0,802
6	2-Methylheptane	3,41	6,406	0,576	1,317	-0,946
7	2,3,4-Trimethylpentane	3,53	6,732	0,548	1,128	-0,946
8	1,1-Dichloroethane	4,47	4,577	0,5	0,74	-0,431
9	Ethylcyclohexane	3,12	5,82	0,633	1,348	-0,946
10	Dipropyl ether	3,95	5,536	0,588	1,641	-0,802
11	trans-1,2-Dichloroethene	4,63	4,414	0,55	1,25	-0,431
12	2-Bromopropane	4,42	4,577	0,5	0,801	-0,646
13	3,3-Dimethylpentane	4,02	5,914	0,55	1,076	-0,936
14	Toluene	3,42	5,113	0,632	1,016	-0,936
15	p-Xylene	3,12	5,983	0,61	0,893	-0,946
16	Ethyl benzene	3,12	5,82	0,633	1,084	-0,946
17	o-Xylene	2,42	5,983	0,613	0,893	-0,946
18	Ethanol	3,87	2,707	0,515	1,026	0,638
19	Methylcyclopentane	4,25	4,406	0,599	1,299	-0,921
20	2-Methyl-2-propanol	3,7	4,5	0,482	0,728	0,132
21	Acrylonitrile	4,15	3,414	0,55	0,47	-0,646
22	1-Pentene	4,96	4,121	0,57	1,313	-0,898
23	Dichloromethane	4,75	3,707	0,515	1,232	-0,264
24	Butyronitrile	3,12	4,121	0,57	0,412	-0,719
25	1-Propanol	3,41	3,414	0,55	0,956	0,323

Tableau III.20 : Valeurs des log Pv mesurés et des descripteurs sélectionnés. (Suite)

N°	Composés	log Pv	X0sol	SpPosA_H2	GATS2e	Hy
26	1-Butanol	2,82	4,121	0,57	0,926	0,132
27	2-Methyl-1-propanol	3,12	4,284	0,541	0,912	0,132

28	1,2-Dichlorobenzene	2,13	6,983	0,613	0,642	-0,71
29	2,2,4-Trimethylpentane	3,79	6,784	0,53	1,034	-0,946
30	2-Pentanone	3,31	4,992	0,556	0,531	-0,767
31	2,4-Dimethylpentane	4,1	5,862	0,545	1,183	-0,936
32	Cumene	2,82	6,69	0,619	1,026	-0,954
33	Isopropanol	3,75	3,577	0,5	0,859	0,323
34	1-Chloropropane	4,66	3,914	0,55	0,934	-0,646
35	2,2-Dimethylbutane	4,63	5,207	0,523	1,005	-0,921
36	Cyclopentane	4,63	3,536	0,585	1,4	-0,898
37	2-Methyl-2-butene	4,72	4,284	0,541	0,963	-0,898
38	Propyl formate	4,02	4,828	0,57	0,906	-0,614
39	Diethyl ether	4,87	4,121	0,57	1,726	-0,719
40	Dibromomethane *	3,76	4,707	0,515	1,272	-0,264
41	1,3-Dichlorobenzene *	2,12	6,983	0,618	0,642	-0,71
42	3-Pentanone *	3,33	4,992	0,562	0,531	-0,767
43	Ethyl acetate *	4,12	4,992	0,556	0,799	-0,614
44	2-Butanone *	4,11	4,284	0,541	0,534	-0,719
45	Hexane *	4,29	4,828	0,57	1,382	-0,921
46	Pentane *	4,85	4,121	0,57	1,36	-0,898
47	1,2-Dichloropropane *	3,83	5,284	0,541	0,964	-0,539
48	Heptane *	3,76	5,536	0,588	1,398	-0,936
49	1,2-dichloroethane *	4,03	4,414	0,55	1,167	-0,431
50	1-Bromopropane *	4,12	4,414	0,55	0,956	-0,646
51	m-xylene *	2,42	5,983	0,618	0,893	-0,946

* Composés de validation

Tableau III.21 : Valeurs des log Pv calculés, prédits, h_{ii}, et $e_{istd.}$

Composé	log $Pv_{Calc/Pred}$	h_{ii}	$ei_{Std.}$	Composé	log $Pv_{Calc/Pred}$	h_{ii}	$ei_{Std.}$

1,1,2-Trichloro-ethane	3,3841	0,105	-0,4647	2-Methyl-1-propanol	3,2801	0,118	0,7935
Iodomethane	4,7638	0,101	0,211	1,2-Dichlorobenzene	1,8814	0,232	-1,5148
Ethyl iodide	4,631	0,072	2,3469	2,2,4-Trimethylpentane	3,7334	0,185	-0,3156
1-Bromobutane	3,63	0,03	-0,3867	2-Pentanone	3,4448	0,115	0,6638
2-Hexanone	3,1166	0,118	2,014	2,4-Dimethylpentane	4,1299	0,089	0,141
2-Methylheptane	3,662	0,108	1,2283	Cumene	2,6799	0,124	-0,7011
2,3,4-Trimethylpentane	3,6423	0,145	0,5827	Isopropanol	3,8153	0,15	0,3419
1,1-Dichloroethane	4,1496	0,1	-1,5394	1-Chloropropane	4,324	0,069	-1,5356
Ethylcyclohexane	3,264	0,136	0,7362	2,2-Dimethylbutane	4,4817	0,104	-0,717
Dipropyl ether	4,1052	0,163	0,8319	Cyclopentane	4,907	0,195	1,5733
trans-1,2-Dichloroethene	4,188	0,054	-1,9698	2-Methyl-2-butene	4,6084	0,097	-0,5337
2-Bromopropane	4,4842	0,12	0,3188	Propyl formate	3,5854	0,034	-1,8789
3,3-Dimethylpentane	3,9226	0,073	-0,4478	Diethyl ether	4,9728	0,211	0,6022
Toluene	3,2118	0,132	-1,0569	Dibromomethan*	4,3039	0,107	2,3622
p-Xylene	2,9552	0,085	-0,7732	1,3-Dichlorobenzene *	1,8198	0,237	-1,4105
Ethyl benzene	2,9638	0,118	-0,7736	3-Pentanone*	3,3708	0,113	0,1776
o-Xylene	2,9182	0,09	2,3562	Ethyl acetate*	3,5608	0,04	-2,342
Ethanol	3,8322	0,261	-0,2442	2-Butanone*	3,8998	0,136	-0,9284
Methylcyclopentane	4,2475	0,109	-0,0122	Hexane*	4,5053	0,085	0,9237

Tableau III.21 : Valeurs des log Pv calculés, prédits, h_{ii}, et $e_{istd.}$ (Suite)

Composé	log $Pv_{Calc/Pred}$	h_{ii}	$ei_{Std.}$	Composé	log	h_{ii}	$ei_{Std.}$

					$Pv_{Calc/Pred}$		
2-Methyl-2-propanol	3,6991	0,173	-0,0051	Pentane*	4,7773	0,116	-0,3174
Acrylonitrile	4,0264	0,23	-0,751	1,2-Dichloropropane*	3,7066	0,042	-0,5175
1-Pentene	4,7238	0,111	-1,1554	Heptane*	3,9942	0,082	1,0032
Dichloromethane	4,7186	0,094	-0,1493	1,2-dichloroethane*	4,0937	0,041	0,2668
Butyronitrile	3,4784	0,211	2,099	1-Bromopropane*	4,1189	0,039	-0,0048
1-Propanol	3,384	0,182	-0,144	m-xylene *	2,8565	0,098	1,886
1-Butanol	3,0135	0,158	1,0282				

* Composés de validation

La matrice de corrélation est reproduite ci-après :

Tableau III.22 : Matrice de corrélation.

	log Pv	X0sol	SpPosA_H2	GATS2e
X0sol	-0,537			
	0,000			
SpPosA_H2	-0,504	0,465		
	0,001	0,003		
GATS2e	0,405	0,045	0,218	
	0,010	0,784	0,182	
Hy	-0,005	-0,623	-0,552	-0,190
	0,977	0,000	0,000	0,247

Tableau III.23 : Caractéristiques des descripteurs sélectionnés pour le modèle MLR.

Descripteur	x	Dx	t	Probabilité-t	VIF
Constante	11,0490	0,6288	17,57	0,000	
X0sol	-0,4602	0,0460	-9,81	0,000	1,720
SpPosA_H2	-12,332	1,210	-10,19	0,000	1,528
GATS2e	1,1372	0,1423	7,99	0,000	1,074
Hy	-1,2333	0,1275	-9,68	0,000	1,944

X0sol : Indice de connectivité de solvatation chi-2 (1).

SpPosA_H2 : Somme positive spectrale normalisée de la matrice de distance au carré réciproque.

GATS2e : Auto-corrélation de lag retard 2 pondérée par l'électronégativité de Sanderson.

Hy: Facteur hydrophile.

Le diagramme de Williams représenté dans la figure III.24 permet d'afficher les valeurs des résidus de prédiction standardisés en fonction de leviers (h_{ii}), pour les deux ensembles (calibrage et validation). Tous les résidus sont situés dans l'intervalle de trois écarts-types, et tous les composés ont un hi < h*, ce qui met en évidence l'absence de point aberrant et /ou influent.

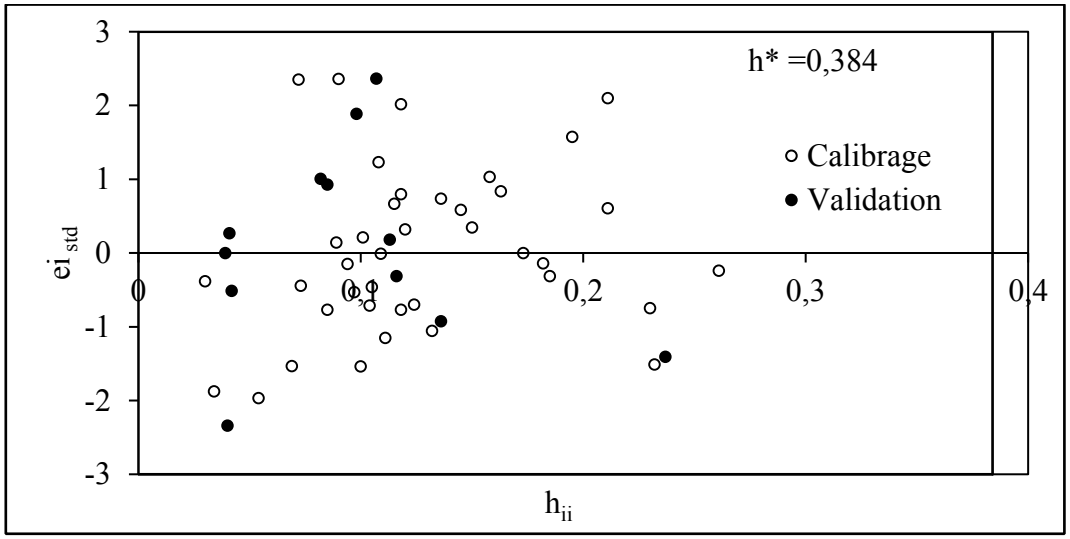


Figure III.24 : Diagramme de Williams.

Tous les résidus standardisés de prédiction sont compris entre les limites ± 3.

Vérification de la qualité de l'ajustement :

La qualité de l'ajustement a été vérifiée en procédant à une validation croisée par «Leave –one –out ». La figure III.25 qui reproduit les valeurs prédites log Pv en fonction de celles observées, fait ressortir une faible dispersion autour de la droite de corrélation caractéristique d'un bon ajustement, d'ailleurs confirmé par la grande valeur de Q^2 (87,48%).

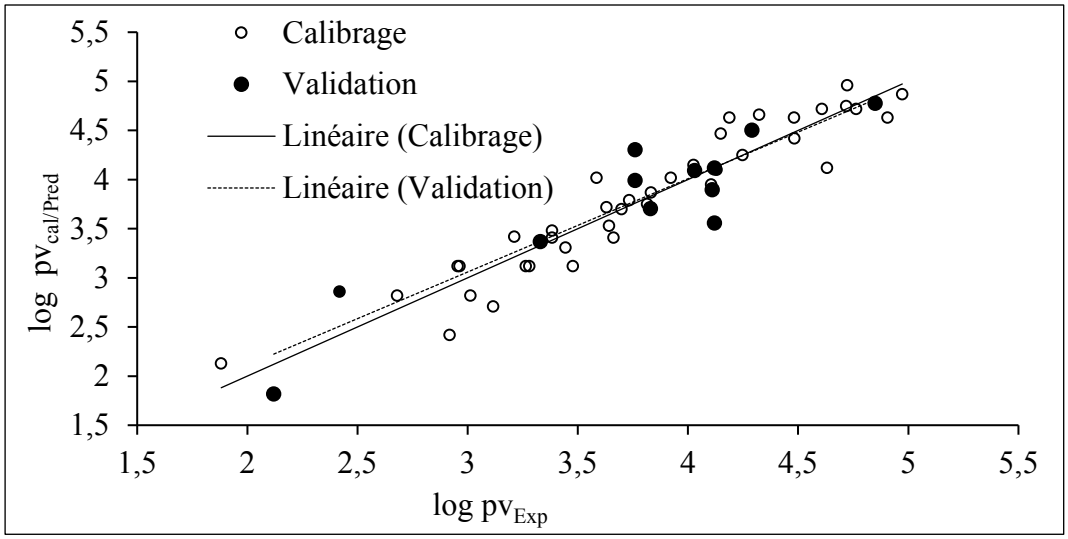


Figure III.25 : Graphe des valeurs prédites log Pv en fonction des valeurs observées.

La validité du modèle a été éprouvée par le test de randomisation de log Pv (Figure III.26).

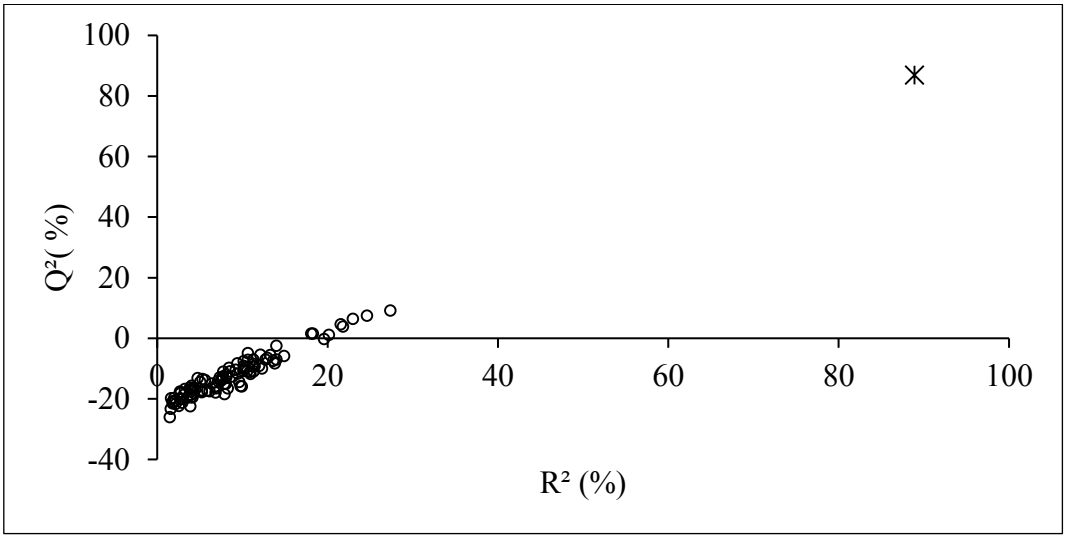


Figure III.26 : Test de randomisation.

La figure III.26 qui représente le graphe des coefficients statistiques Q^2 et R^2 permet de comparer les résultats obtenus pour les modèles randomisés à ceux du modèle réel de départ.

Il est clair que les statistiques obtenues pour les vecteurs modifiés de log Pv sont plus petites que celles du modèle QSPR réel, ce qui permet d'affirmer que le modèle proposé n'est pas aléatoire.

III-2-2-Les réseaux de neurones artificiels :

Le choix du nombre de neurones de la couche cachée est fixé à 4 et le nombre d'itérations à 24. Les graphes suivant explicitent ce choix.

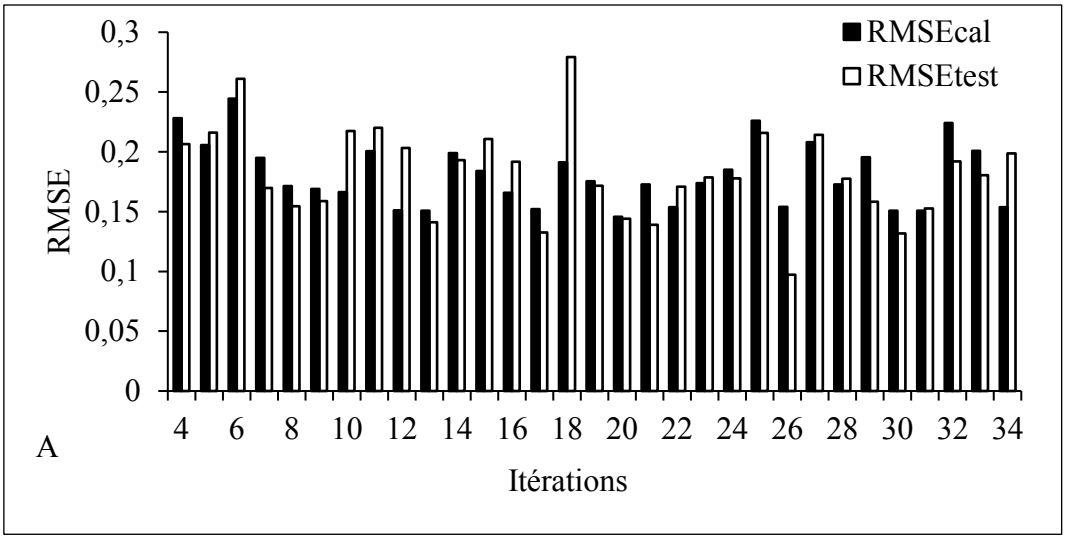


Figure III.27 : Variation des RMSE en fonction des itérations du deuxième neurone.

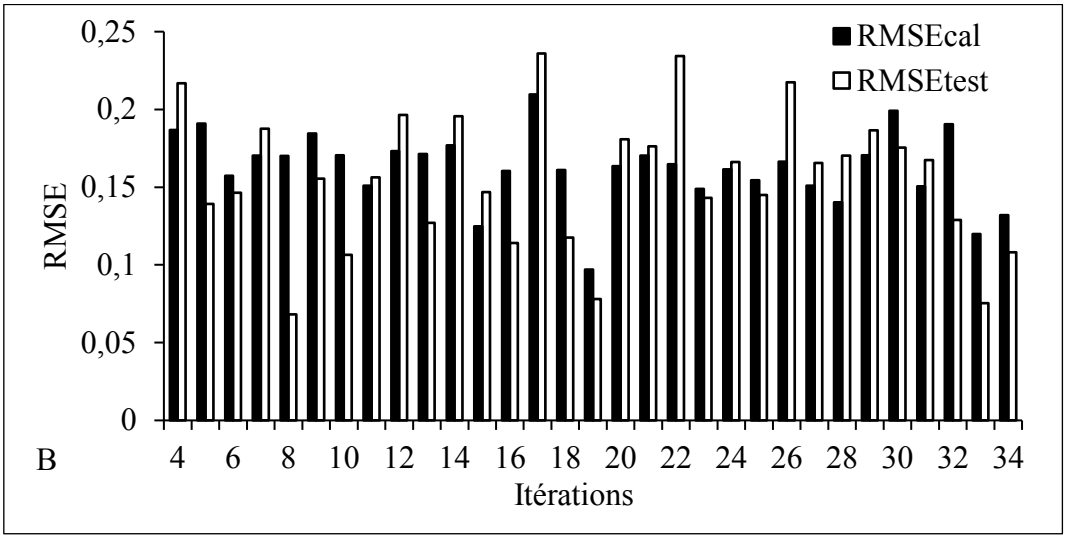


Figure III.28: Variation des RMSE en fonction des itérations du troisième neurone.

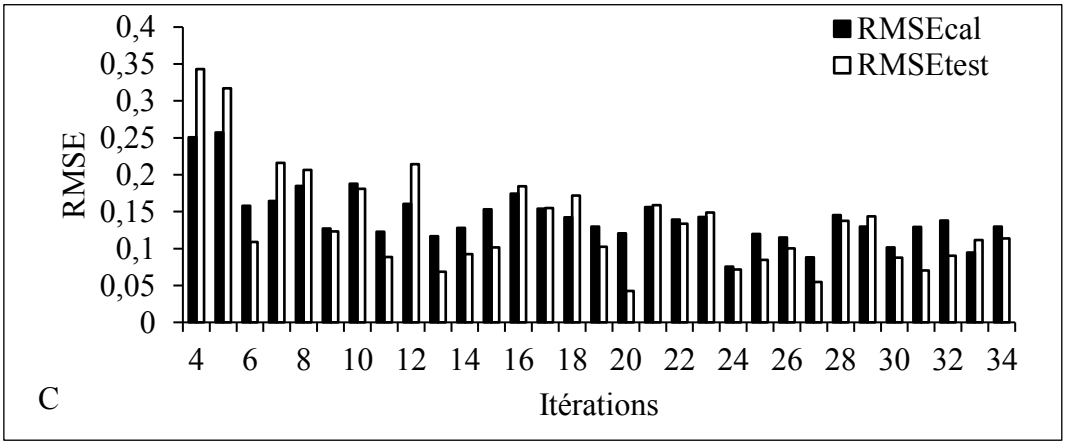


Figure III.29 : Variation des RMSE en fonction des itérations du quatrième neurone.

Tableau III.24 : Valeurs des paramètres statistiques (RNA).

Neurones	Itérations	R^2	Q^2_{Ext}	S	RMSE-cal	RMSE-test
2	20	95,93	87,16	0,147	0,145	0,144
3	19	98,20	89,41	0,098	0,096	0,078
4	24	98,91	92,07	0,076	0,075	0,071

La structure optimale adoptée est reproduite dans le tableau III.25.

Tableau III.25 : Structure optimale adopté pour le réseau de neurones.

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

$R^2 = 98,91$ $RMSE = 0,075$ $RMSE_{Ext} = 0,071$ $Q^2_{Ext} = 92,07$

La validation statistique externe (Q^2_{Ext}) atteste de la bonne capacité prédictive des composés n'ayant pas participé au calcul du modèle.

Tableau III.26 : Comparaison de la qualité des modèles RLM et RNA pour la pression de vapeur.

Méthodes	Calibrage n=39		Validation n=12		
	R^2	Q^2_{Ext}	RMSE	$RMSE_{Ext}$	S
MLR	90,09%	83,07%	0,227	0,297	0,24
RNA	98.91%	92.07%	0,075	0,071	0,076

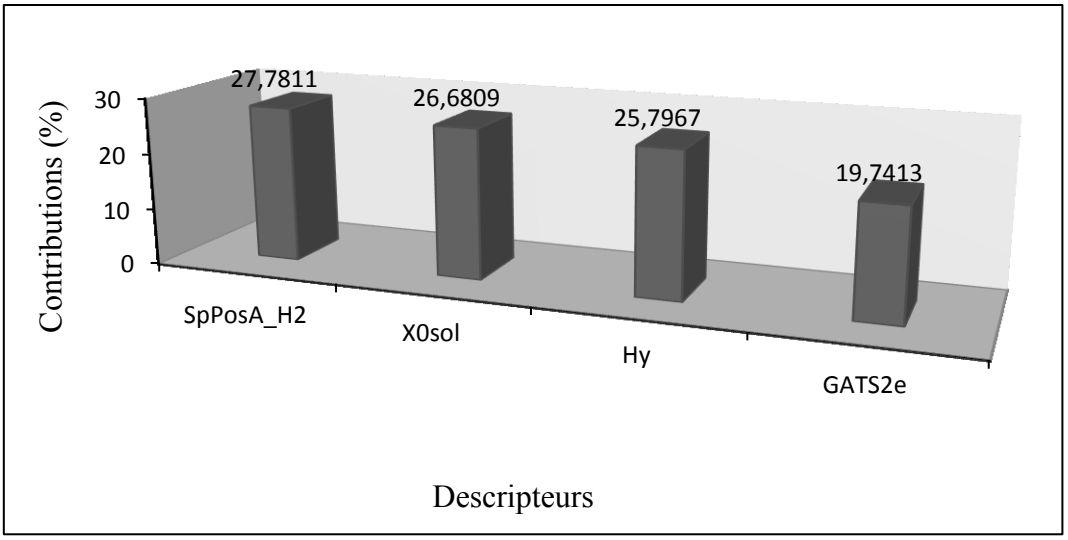


Figure III.30 : Contributions des descripteurs du modèle MLR.

Les contributions relatives des quatre descripteurs du modèle ont été déterminées. Elles diminuent selon l'ordre suivant : SpPosA_H2 (27,7811%) > X0sol (26,6809%) > Hy (25,7967%) > GATS2e (19,7413%). Il convient de noter que la différence de contribution entre deux descripteurs utilisés dans le modèle indique que tous les descripteurs sont indispensables pour générer le modèle prédictif.

III- 3-Conclusion

Les valeurs expérimentales de log Pv de 51 COVs ont été séparées, aléatoirement, en deux sous-ensembles disjoints :

- de 39 éléments réservés au calcul des modèles à partir de descripteurs théoriques reflétant la structure moléculaire et en adoptant une approche hybride soit AG/ MLR, soit AG/ RNA;
- de 12 éléments, exclusivement réservés à la validation externe.

L'approche par réseaux de neurones conduit au meilleur modèle à tous les points de vue : capacités prédictives interne et externe, qualité de l'ajustement…, ce qui prouve dans ce cas, que les corrélations variable dépendante/ variables explicatives sont fondamentalement non linéaires.

CONCLUSION GÉNÉRALE

Conclusion générale

L'objectif de cette thèse était de développer des modèles QSRR/QSPR fiables pour la prédiction de quelques propriétés de certains composés organiques volatils. Un grand nombre de descripteurs moléculaires a été calculé (Descripteurs constitutionnels, électroniques, topologiques, géométriques, physicochimiques, thermodynamiques, ...). Diverses méthodes statistiques ont été utilisées dans la construction de ces modèles (RLM, SVM, RNA...).

Les principales techniques de validation ont été utilisées (les tests statistiques standards, la validation interne, la validation externe, le test de randomisation, les domaines d'applicabilité...).

Dans ce cadre, nous avons présenté dans ce travail trois applications.

Dans la première application nous avons établis des modèles reliant certains descripteurs moléculaires avec les modèles QSAR utilisant l'analyse de régression linéaire multiple associant l'approche algorithme génétique (MLR/AG) pour la sélection de sous ensemble des variables significatives avec le logiciel MOBYDIGS, en maximisant la valeur du coefficient de prédiction Q^2_{LOO}.

Nous avons utilisé la méthodologie QSPR pour relier trois propriétés (Temps de rétention relatif, coefficient de partage octanol / eau et pression de vapeur,) d'un mélange hétérogène de composés organiques volatils ayant des propriétés chimiques et des origines diverses, à des descripteurs moléculaires théoriques caractéristiques de la molécule entière ou de ses fragments, calculés à l'aide de logiciels spécialisés du commerce.

Afin de vérifier que les modèles obtenus ne sont pas dûs au hasard ou à une sur-spécification, nous avons appliqué le test de randomisation. Les résultats statistiques obtenus pour les vecteurs modifiés des temps de rétention relatifs sont plus petits que ceux du modèle QSRR réel (Q^2=0,234 et R^2=0,155) ce qui confirme la relation structure / temps de rétention relatif.

En ce qui concerne la modélisation du coefficient de partage octanol/eau l'approche RLM s'est avérée.

L'approche par réseaux de neurones artificiels pour la modélisation de la pression de vapeur de 51 COVs conduit au meilleur modèle à tous les points de vue : capacités prédictives interne et externe, qualité de l'ajustement…, ce qui prouve dans ce cas, que les corrélations variable dépendante/variables explicatives sont fondamentalement non linéaires.

Les trois propriétés sont modélisées avec succès en utilisant l'analyse de régression linéaire multiple (MLR), les réseaux de neurones artificiels (RNA), les machines à vecteurs supports (SVM).

Les modèles QSRR/QSPR développés sont simples, interprétables et transparents en utilisant un nombre réduit de descripteurs. En outre, ils ont une bonne stabilité, une robustesse et un pouvoir prédictif élevés, vérifié par la validation interne qui est claire à partir de son coefficient de corrélation R^2 et le coefficient de validation croisée Q^2_{Loo} et plus précisément de sa validation externe Q^2_{Ext}.

Ainsi, les modèles sont considérés comme validés et applicables pour l'exploitation de la base de données. Le domaine d'applicabilité des meilleurs modèles étudiés obtenus avec les différentes propriétés calculés selon la méthode PM3 peut être servir comme un outil précieux pour filtrer les dissemblables et les valeurs aberrantes. Ainsi, il est applicable de faire des prédictions pour les nouveaux composés.

Ces descripteurs contribuent, non seulement, à obtenir des modèles robustes et prédictifs mais ils rendent également ces derniers chimiquement plus interprétables.

Références bibliographiques

Références bibliographiques

[1] Majoli, L. (2005). Elaboration, caractérisation et étude des performances de nouveaux adsorbants hydrophobes: application aux atmospheres odorantes et/ou chargées en composés organiques volatils (Doctoral dissertation).

[2] Kerbachi, R., Oucher, N., Bitouche, A., Berkouki, N., Demri, B., Boughédaoui, M., & Joumard, R. (2009, February). Pollution par les particules fines dans l'agglomération d'Alger.

[3] Chtita, S. (2017). Modélisation de molécules organiques hétérocycliques biologiquement actives par des méthodes QSAR/QSPR. Recherche de nouveaux médicaments (Doctoral dissertation).

[4] Les Composés organiques volatils /Agence de l'environnement et de la maitrise de l'énergie (ADEME) / Paris : Dunod - DL 2013.
http://portail-bu.univ-artois.fr/medias/doc/EXPLOITATION/ABSYS/735026/les-composes-organiques-volatils-reduction-des-emissions-de-cov-dans-l-industrie

[5] Directive n°2004/42/CE du 21/04/04 relative à la réduction des émissions de composés organiques volatils dues à l'utilisation de solvants organiques dans certains vernis et peintures et dans les produits de retouche de véhicules, et modifiant la directive n°1999/13/CE.
http://www.ineris.fr/aida/consultation_document/957

[6] Norme ISO 16000-6, 2011, Air intérieur — Partie 6: Dosage des composés organiques volatils dans l'air intérieur des locaux et chambres d'essai par échantillonnage actif sur le sorbant Tenax TA, désorption thermique et chromatographie en phase gazeuse.
https://www.iso.org/obp/ui/fr/#iso:std:iso:16000:-6:ed-2:v1:fr

[7] Eurofins France ; Eurofins Product Testing A/S. 2012.
www.eurofins.com/COV-fr

[8] Le Cloirec, P. (2004). COV (composés organiques volatils). Ed. Techniques Ingénieur.

[9] Agence française de sécurité sanitaire de l'environnement et du travail (AFSSET), rapport_COVîle de France, 2004.

[10] Airparif. Les gaz à effet de serre en Ile-de-France : par qui sont-ils émis ? Airparif Actualité. 2006, n°28, pp.1-8.
https://www.airparif.asso.fr/_pdf/publications/Rges_resume.pdf

[11] Grange, D., Host, S., & Gremy, I. (2007). Les composés organiques volatils (COV). In Etat des lieux: définition, sources d'émissions, exposition, effets sur la santé. Rapport ORS. Îlede-France. France.

[12] Institut National de l'Environnement industriel et des risques (INERIS).

http://www.ineris.fr/ressources/recherche/Guide%20GEIDE.

[13] Institut national de recherche et de sécurité (INRS), Fiche toxicologique Acide acétique.: http://www.inrs.fr/publications/bdd/fichetox/fiche.html?refINRS=FICHETOX_24

[14] Fabure, J. (2009). Étude de l'accumulation et des effets des composés organiques volatils (BTEX) chez les bryophytes (Doctoral dissertation). https://tel.archives-ouvertes.fr/tel-00557714/

[15] CHIMIE, 11e année : document de mise en œuvre. Manitoba Éducation, citoyenneté et jeunesse.2012,https://www.dref.mb.ca/notice?id=p%3A%3Ausmarcdef_0000085092&queryId=183adffc-c14c-4cd8-a0c3-7a162ac33f35&posInSet=3

[16] Dupuis, G. (2014). Alcènes et autres composés éthyléniques. https://www.faidherbe.org/site/cours/dupuis/ethyled.htm

[17] Boumendjel, A. (2012). Alcool, Université Joseph Fourier-Grenoble1. http://unf3s.cerimes.fr/media/paces/Grenoble_1112/boumendjel_ahcene/boumendjel_ahcene_p09/boumendjel_ahcene_p09.pdf

[18] Boust, C. (2009). Les cétones, ED 4221 Fiche Solvants 2ème édition .Institut National de la Recherche et de Sécurité(INRS) Paris. http://www.inrs.fr/dms/inrs/CataloguePapier/ED/TI-ED-4221/ed4221.pdf

[19] Boust, C. (2009). Les esters, ED 4227 Fiche Solvants 1ère édition .Institut National de la Recherche et de Sécurité (INRS) Paris. http://www.inrs.fr/dms/inrs/CataloguePapier/ED/TI-ED-4227/ed4227.pdf

[20] Boust, C. (2011), les hydrocarbures halogénés, ED 4223 Fiche Solvant 2ème édition. Institut National de la Recherche et de Sécurité (INRS) Paris. https://www.cancer-environnement.fr/LinkClick.aspx?fileticket=8iG_bTbVdcw%3D&tabid=320&portalid=0&mid=1744

[21] Ben Romdhane, H., Les fonctions chimiques, Faculté des Sciences de Tunis. http://www.orgapolym.com/pdf/cahier5/6_alcools

[22] Allinger, N.L., Cava, M.P., de Jongh, D.C., Johnson, C.R., Lebel, N.A., C.L. Stevens, (1976), Chimie Organique: Structure, McGraw-Hill, Montréal, 370 pages. ISBN 2-7042-0095-5

[23] OMS, Série de rapports techniques, (2006). Comité OMS d'experts des problèmes liés à la consommation d'alcool, 2éme ed, New York.

[24] Bourgeais, A. (2009). Cours de Chimie Organique (Alcanes), [Pdf], <URLhttp://eduscol.education.fr/rnchimie/chi_org/ab/chap6-alcanes.>.

[25] Air pollution by industries and households.
http://ec.europa.eu/eurostat/statistics_explained/index.php/Air_pollution_by_industries_and_h ous eholds#Source_data_for_tables_and_figures_.28MS_Excel.29
[26] Hansen, F. (2012). Institut national de la statistique et des études économiques, http://www.statistiques.public.lu/catalogue-publications/luxembourg/2012/PDF-17-12.pdf
[27] Ministère du Développement durable.
http://www.developpementdurable. gouv.fr/COV.html.
[28] Bisson, M. et al, (2005). Les composés organiques volatils (COV) dans l'air ambiant au Québec. http://www.mddep.gouv.qc.ca/air/cov/rapport89-99.
[29] Centre Interprofessionnel Technique d'Etudes de la Pollution Atmosphérique (CITEPA), Protocole de Göteborg, (2012). http://www.citepa.org/fr/actualites/198-30-avril-4-mai-2012-protocole-de-goeteborg.
[30] Centre Interprofessionnel Technique d'Etudes de la Pollution Atmosphérique (CITEPA), inventaire SECTEN, (2013).
http://www.citepa.org/fr/actualites/1113-4-juillet-2013-mise-enligne-de-la-mise-a-jour-de-l-inventaire-secten.
[31] Cancer et environnement.
http://www.cancer-environnement.fr/343-Composes-Organiques-Volatils-COV-dans-lair.ce.aspx#COV_et_canc_ro.
[32] Institut national de recherche et de sécurité (INRS), Fiche toxicologique Ozone.
http://www.inrs.fr/publications/bdd/doc/fichetox.html?refINRS=FT%2043.
[33] Institut national de recherche et de sécurité (INRS), Oxydation thermique et catalytique.
http://www.inrs.fr/media.html?refINRS=ED%204261
[34] SOLTYS, N. (1998). Procédés de traitement des COV ou composés organiques volatils. Techniques de l'Ingénieur J, 3, 928.
[35] Tatibouët, J M., (2015). « Plasma non thermique et traitement de l'air», Techniques de l'Ingénieur.http://www.techniquesingenieur.fr/basedocumentaire/environnement_securiteth5/traitements-de-l-air-42600210/plasma-non-thermiqueet-traitement-de-l-air-g1794/.
[36] Sun, L.M et Thonnelier, J-Y. (2015). « Perméation gazeuse», Ed. Techniques Ingénieur, http://www.techniques-ingenieur.fr/base-documentaire/procedes-chimie-bio-agroth2/operations-unitaires-techniques-separatives-sur-membranes-42331210/permeation gazeusej2810/ aspects-generaux-j2810niv10001.html#niv-sl2976403.

[37] Le Page, J. F. (1978). Catalyse de contact: conception, préparation et mise en œuvre des catalyseurs industriels, Technip.

[38] http://www.univ-bejaia.dz/dspace/handle/123456789/5101

[39]J. A. Pople, D. L. Beveridge, (1970). Approximate Molecular Theory , Mc Graw- Hill, New York.

[40] Hatree D. R., (1928). The Wave Mechanics of an Atom tvith a Non-Voulomb Ventral, Proc. Cambridge. Phill. Soc. Vol, 24, pp, 328.

[41] Fock V., (1930). Näherungsmethode zur lösung des quantenmechanischen mehrkörperproblems, Z. Physik. Vol, 61, pp, 126.

[42] Slater J. C., (1930). The self consistent field and the structure of atoms, Phys. Rev. Vol, 32, pp, 339.

[43] Löwdin P. O., (1950). On the Non Orthogonality Problem Connected with the Use of Atomic Wave Functions in the Theory of Molecules and Crystals, J. Chem. Phys, Vol, 18, pp, 365.

[44] Pariser R., Parr R. G., (1953). A Semi Empirical Theory of the Electronic Spectra and Electronic Structure of Complex Unsaturated Molecules I, J. Chem. Phys. Vol, 21, pp, 466-477.

[45] Pariser R., Parr R. G., (1953). A Semi Empirical Theory of the Electronic Spectra and Electronic Structure of Complex Unsaturated Molecules II, J. Chem. Phys. Vol, 21, pp, 767-776.

[46] Pople J. A., (1953). Electron interaction in unsaturated hydrocarbons, Trans. Faraday Soc. Vol, 49, pp, 1375-1385.

[47] Pople J. A., Santry D. P., Segal G. A., (1965). Approximate Self Consistent Molecular Orbital Theory. I. Invariant Procedures. J. Chem. Phys. Vol, 43, pp, 5129.

[48] Pople A., Beveridge D. L., Dobosh P. A., (1967). Approximate Self Consistent Molecular Orbital Theory. V. Intermediate Neglect of Differential Overlap, J. Chem. Phys. Vol, 47, pp, 2026-2033.

[49] Dewar M. J. S., Thiel W., (1977). Ground states of molecules. The MNDO method. Approximations and Parameters, J. Am. Chem. Soc., Vol, 99, pp, 4899-4907.

[50] Burstein K. Y., Isaev A.N., (1984). MNDO calculations on hydrogen bonds. Modified function for core-core repulsion, Theor. Chim. Acta. Vol, 64, pp, 397- 401.

[51] Dewar M. J. S., Zoebisch E. G., Healy E. F., Stewart J. J. P., (1985). Development and use of quantum mechanical molecular models. 76. AM1: a new general purpose quantum mechanical molecular model, J. Am. Chem. Soc, Vol, 107, pp, 3902- 3909.

[52] Stewart J. J., (1989). Optimization of parameters for semi empirical methods I. Method, J. Comput. Chem., Vol, 10, pp, 209-220 ; 221-264.

[53] Jensen F., (1998). Introduction to Computational Chemistry, Wiley, pp, 94-96.

[54] Stewart J. J. P., (1996). Application of Localized Molecular Orbitals to the Solution of Semiempirical Self-Consistent Field Equations, Int. J. Quantum. Chem., Vol, 58, pp, 133.

[55] Daniels A. D., Millam J. M., Scuseria G. E., (1997). Semi empirical methods with conjugate gradient density matrix search to replace diagonalization for molecular systems containing thousands of atoms, J. Chem. Phys., Vol, 107,pp, 425.

[56] Ramachandran K. I., Deepa G., Namboori K., (2008). Comptutional Chemistry and Molecular Modeling, Principles and Applications, Springer- Verlag Berlin Heidebberg.

[57] Coulson C. A., Longuet-Higgins H. C., (1947). The Electronic Structure of Conjugated Systems. I. General theory , Proc. Roy. Soc. (London) A 191, p, 39.

[58] Mulliken R. S., (1962). Criteria for the Construction of Good Self Consistent Field Molecular Orbital Wave Functions, and the Significance of LCAOMO Population Analysis″, J. Chem. Phys., Vol, 36, p, 3428.

[59] Kutzelnigg W., Delre G., Bertheir G.. (1971). σ and π Electrons in Theoretical Organic Chemistry, Springer Verlag, Berlin.

[60] Pullman B., (1969). La Biochimie Electronique, Collection Que sais-je ? PUF, n°1075, Deuxième édition, Paris.

[61] Boyd D. B., Lipkowitz K. B., (2000). eds. Reviews in Computational Chemistry, History of the Gordon Conferences on Computational Chemistry,Wiley- VCH, New York, pp, 399-439.

[62] Allinger N. L., (1977). Conformational analysis. 130. MM2. A hydrocarbon force field utilizing V1 and V2 torsional terms. J. Am. Chem. Soc., Vol, 99, pp, 8127- 8134.

[63] Burkert U., Allinger N. L., (1986). Molecular Mechanics, ACS Monograph No. 177, American Chemical Society, Washington, DC, 1982.

[64] Allinger N. L., Yuk Y. H., Lii J. –H., (1989). Molecular Mxhanics. The MM3 Force Field for Hydrocarbon 3. 1, J. Am. Chem. Soc., Vol, 111, pp, 8551- 8565.

[65] Allinger N. L.. Chem K., Katzenellbogen J. A., Wilson S. R., Anstead G. M., (1996). Hyperconjugative Effects on Carbon-Carbon Bond Lengths in Molecular Mechanics (MM4), J. Comput. Chem., Vol, 17, pp, 747- 755.

[66] Mac Kerell A. D., Bashford Jr., D., Bellott M., Dumbrack R. L., Evaseck Jr., J. D., Field M. J., Fischer S., Gao J., Gao H., He S., Joseph- Mac Carthy D., Kuchnir L., Kuczera K., Lau F. T. K., Mattos C., Michmick S., Nego T., Nguyen D. T., Prodhom B., Reiher III W. E., Roux B., Schlemkrich M., Smith J. C., Stote R., Straub J., Watanabe M., Wiorkiewicz-Kuczera J., Yin D., Karplus M., (1998). All-atom empirical potential for molecular modeling and dynamics studies of proteins, J. Phys. Chem. B, Vol,102, pp, 3586- 3616.

[67] Brooks B. R. et al., (1983). CHARMM: A Program for Macromolecular Energy, Minimization, and Dynamics Calculations, J. Comput. Chem., Vol, 4, pp, 187.

[68] Mackerell A. D. et al., (1995). An all-atom empirical energy function for the simulation of nucleic acids, J. Am. Chem. Soc., Vol, 117, pp, 11946.

[69] Mackerell A. D. et al., (1998). All-Atom Empirical Potential for Molecular Modeling and Dynamics Studies of Proteins. J. Phys. Chem. B, Vol, 102, pp, 3586.

[70] Momany F. A., Rone R., (1992). Validation of the General Purpose QUANTAa3.2/ CHARMm ® Force Field, J. Comput. Chem., Vol, 13, pp, 888.

[71] Halgren T. A., (1996).Merck Molecular Force Field I, J. Comput. Chem., Vol, 17, pp, 490, 520, 553, 616.

[72] Halgren T. A., Nachbar R. B., (1996). Merck Molecular Force Field. IV.Conformational Energies and Geometries for MMFF94, J. Comput. Chem., Vol, 17, pp, 587-615.

[73] Leach A. R., (2001). Molecular modeling- Principles and applications, Person, Prentice Hall, Second Edition, England, Chap.4.

[74] Alder B. J., Wainwright T. E., (1957). Phase Transition for a Hard Sphere System, J. Chem. Phys., Vol, 27, pp, 1208.

[75] Rahman A., (1964). Correlations in motion of atoms in liquid argon, Phys. Rev., 136, A405.

[76] Rahman A., Stillinger F. H., (1971). Molecular Dynamics Study of Liquid Water, J. Chem. Phys., Vol, 5, pp, 3336.

[77] Mc Cammon J. A., Harvey S. C., (1987). Dynamics of Proteins and Nucleic Acids, Cambridge Univ. Press.

[78] Leach A. R., (2007). Introduction to Chemoinformatics; Springer.

[79] Roy K., Kar S., Das R. N., (2015). A premier on QSAR / QSPR Modeling- Fundamental Concepts, Springer Breifs in Molecular Science- DOI 10. 1007 / 978- 3- 319- 17 281- 1.

[80] Todeschini R., Consonni V., (2000). Handbook of molecular descriptors. Wiley VCH, Weinhein

[81] Livingstone D. J.. (2000). The Characterization of chemical structures using molecular properties. A survey. I. Chem. Inf. Comput. Sci., Vol, 40, pp, 195- 209.

[82] Cros A. F. A., (1863). Action de l'alcool amylique sur l'organisme, Thèse de doctorat, faculté de médecine.

[83] Crum-Brown A. C. and Fraser T. R., (1868). On the Connection Between Chemical Constitution and Physiological Action, Part I: On the Physiological Action of the Salts of the Ammonium Bases, Derived from Strychnia, Brucia, Thebia, Codeia, Morphia, Nicotia, Earth an, Trans, Roy, Soc., Vol, 25, pp, 151-203.

[84] Richet M. C., (1893). Noté sur le rapport entre la toxicité et les propriétés physiques des corps, Comptes rendus des séances de la Société de biologie et de ses filiales, Paris, Vol, 45, pp, 775–6.

[85] Meyer H., (1899). Zur Theorie der Alkoholnarkose. Erste Mittheilung. Welche Eigenschaft der Anästhetica bedingt ihre narkotische Wirkung, Archiv für experimentelle Pathologie und Pharmakologie, Vol, 42, pp, 109–118.

[86] Overton E., (1901). Studien über die Narkose zugleich ein Beitrag zur allgemeinen. Pharmakologie, Ed. G. Fischer, Jena.

[87] Lipnick R. L., (1986). Charles Ernest Overton: narcosis studies and a contribution to general pharmacology, Trends in Pharmacological Sciences, Vol, 7, pp, 161–164.

[88 Fühner H. and Neubauer E., 1907. ămolyse durch Substanzen homologen Reihen, Archiv für experimentelle Pathologie und Pharmakologie, Vol, 56, pp,333–345.

[89] Hansen O. R., 1962. Hammett Series with Biological Activity, Acta Chemica Scandinavica, Vol, 16, pp, 1593–1600.

[90] Hansch C. and Fujita T., (1964). p-σ-π Analysis. A Method for the Correlation of Biological Activityand Chemical Structure, Journal of the American Chemical Society, Vol, 86(8), pp, 1616–1626.

[91] Free S. M. and Wilson J. W., (1964). A Mathematical Contribution to Structure-Activity Studies, Journal of Medicinal Chemistry, Vol, 7(4), pp, 395–399.

[92] Hansch C. and Lien E. J., (1971). Structure-activity relationships in antifungial agents. A survey, Journal of Medicinal Chemistry, Vol,14(8), pp, 653–670.

[93] Tham S. Y. and Agatonovic-Kustrin S., (2002). Application of the artificial neural network in quantitative structure-gradient elution retention relationship of phenylthiocarbamyl amino acids derivatives, Journal of Pharmaceutical and Biomedical Analysis, Vol, 28(3), pp, 581-590.

[94] Cramer R. D., Patterson D. E., and Bunce J. D., (1988). Comparative molecular field analysis, J. Am. Chem. Soc., Vol, 110(18), pp, 5959- 5967.

[95] Klebe G., Abraham U., and Mietzner T., (1994). Molecular similarity indices in a comparative analysis (CoMSIA) of drug molecules to correlate and predict their biological activity, Journal of Medicinal Chemistry, Vol, 37(24), pp, 4130-4146.

[96] Fortuné A., (2006). Techniques de Modélisation Moléculaire appliquées à l'étude et à l'optimisation de Molécules Immunogènes et de Modulateurs de la Chimiorésistance, Thèse de Doctorat,Université Joseph Fourier – Grenoble I, France.

[97] Box G. E. P. and Cox D. R., (1964). An analysis of distributions, Journal of the royal statistical society, Series B, Vol, 26(2), pp, 211-243.

[98] Armitage P., Berry G., (1994). Statistical Methods in Medical Research, 3rd ed., Blackwell.

[99] Hyperchem ^{TM} Release 6,03 for windows, Molecular Modeling System (2000).

[100] Levine I. N., (2000). Quantum Chemistry, 5thed, New Jersey: Prentice Hall.

[101] Todeschini R., Consonni V., Pavan M., (2005) DRAGON, Software for the Calculation of Molecular Descriptors, Release 5.3 for windows, Milano.

[102] Todeschini R., Ballabio D., Consonni V., Mauri A., Pavan M., (2009). MOBY DIGS Software for Multilinear Regression Analysis and Variable Subset Selection by Genetic Algorithm, Release 1,1 for windows, Milano.

[103] Kowalski B., Gerlach R., Wold H., (1982). Systems under indirect observation, (K, Jörenskoget H, Wold, eds,), North Holland, Amsterdam, pp, 191-206..

[104] Erikson L., Johannson E., Kettaneh-Wold N., (2001). Multi and megavariate data analysis- principles and applications″, Umetrics Academy, Umeå.

[105] Wold S., (1984). Chemometrics: Mathematics and Statistics in Chemistry, Reidel, Dordrecht, The Netherlands.

[106] Wold S., Ruhe A., Wold H., Dunn W., (1984). The Collinearity Problem in Linear Regression. The Partial Least Squares (PLS) Approach to Generalized Inverses,SIAMJ, Sci, Stat, Comput, Vol, 5, pp, 735- 743.

[107] Gelada P., Kowalski B. R., (1986). Partial least- squares regression: tutorial, Anal, Chim, Acta, Vol, 185, pp, 1- 17.

[108] Höskuldsson A., (1988). PLS regression methods, J, Chemometrics, Vol, 2, pp, 211-228.

[109] Burns J. A., Whiteside G. M., (1993). Feed- forward neural networks in chemistry: mathematical systems for classification and pattern recognition, Chem, Rev, Vol, 93(8), pp, 2583- 2601.

[110] Anker L. S., Jurs P. C., (1992). Prediction of carbon-13 nuclear magnetic resonance chemical shifts by artificial neural networks, Anal, Chem, Vol, 64, pp, 1157- 1164.

[111] Aoyama T., Suzuki Y., Ichikawa H., (1990). Neural networks applied to quantitative structure-activity relationship analysis, J, Med, Chem, Vol, 33, pp, 2583- 2590.

[112] Andrea T. A., (1991). Applications of neural networks in quantitative structure-activity relationships of dihydrofolate reductase inhibitors, J, Med, Chem, Vol, 34, pp, 2824 – 2836.

[113] Jurs P. C., (1996). Computer Software Applications in Chemistry, Second Edition, J, Wiely.

[114] Chouquet C., (2010). Modèles Linéaires, Laboratoire de Statistique et Probabilités-Université Paul Sabatier-Toulouse.

[115] Le jeune M, (2004). Statistiques : la théorie et ses applications, Springer-Verlag, Paris.

[116] Fayet G., (2010). Développement de modèles QSPR pour la prédiction des propriétés d'explosibilité des composes nitroaromatiques, Thèse de doctorat de l'université Pierre et Marie Curie.

[117] Mc Culloch-Pitts, (1943). A logical Calculus at the ideas imminent in Nervous Activity. Bulletin at math. Biophysics.Vol. 5, pp, 115-133.

[118] Minsky M., Papert S., (1969). Perceptrons. Massachusetts: MIT press.

[119] Rumelbart D. E., Mcclelland J. L. et al.,(1988). Parallel Distributed processing. Massachusetts: MIT press,Vol, 1, pp, 547.

[120] Hopfield J. J.. (1982). Neural Networks and physical systems with emergent collective computational abilities. Proceedings of the National Academy of sciences. USA.

[121] Kohonen T.. (1988). Self-organization and associative memory. Bulletin: Springer-Verlag. 984.

[122] Hecht-Nielson R.. (1991). Neurocomputing. Addison-Wesly Publishing Company, New York.

[123] Fogelmen-Soulié F.. (1988). Méthodes connexionnistes pour l'apprentissage. Actes des journées Nationales sur l'intelligence Artificielle. Paris: Teknea. pp, 275-293.

[124] Hornik K., (1991). Approximation capabilities of multilayer feedforward networks, Neural Networks, Vol, 4, pp, 251-257.

[125] Matlab Version 7.0.0.19920 (Release 14) The Language of Technical Computing The MathWorks, Inc. May 06, (2004).

[126] Vapnik, V. N. (1995). The nature of statistical learning. *Theory*.

[127] Smola, A. J., & Schölkopf, B. (2004). A tutorial on support vector regression. *Statistics and computing*, *14*(3), 199-222.

[128] Cristianini, N., & Shawe-Taylor, J. (2000). *An introduction to support vector machines and other kernel-based learning methods*. Cambridge university press.

[129] Bishop, C. M. (1995). *Neural networks for pattern recognition*. Oxford university press.

[130] Draper N. R., Smith H., (1998). Applied Regression Analysis, Third Edition, Wiley series in Probability and Statistics, New york.

[131] Gramatica P.. (2007). Principles of QSAR Models Validation: Internal and External. Qsar & Combinatorial Science, Vol, 26, pp, 694–701.

[132] Eriksson L., Jaworska J., Worth A. P., Cronin M.T.D., Mc Dowell R. M., Gramatica P., (2003). Methods for reliability and uncertainty assessment and for applicability evaluations of classification-and regression-based QSARs. Environmental health perspectives, Vol, 111(10), pp, 1361-1375.

[133] Golbraikh A.. and Tropsha A.. (2002). Beware of Q(2), Journal of Molecular Graphics & Modelling, Vol, 20, pp, 269–276.

[134] Dearden J. C., Cronin M. T. D. and 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, Vol, 20, pp, 241–266.

[135] Lozano S., Halm-Lemeille M.-P., Lepailleur A., Rault S., Bureau R.. (2010) Consensus QSAR Related to Global or MOA Models: Application to Acute Toxicity for Fish. Molecular Informatics, Vol, 29, pp, 803–813.

[136] Roy P. P., Kovarich S.. and 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 Vol , 32, pp, 2386-2396.

[137] Chirico N.. and Gramatica P., Real P.. (2011). External Predictivity of QSAR Models: How To Evaluate It? Comparison of Different Validation Criteria and Proposal of Using the Concordance Correlation Coefficient″. Journal of Chemical Information and Modeling, Vol, 51, pp, 232.

[138] Golbraikh A., Shen M., Xiao Z. Y., Xiao Y. D., Lee K. H., Tropsha A., (2003). Rational Selection of Training and Test Sets for the Development of Validated QSAR Models. Journal of Computer-Aided Molecular Design, Vol, 17, pp, 241–253.

[139] Tropsha A., Gramatica P.. and 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, Vol, 22, pp, 69–77.

[140] Schüürmann G., Ebert R. –U., Chen J., Wang B., Kühne R.. (2008). External Validation and Prediction Employing the Predictive Squared Correlation Coefficient - Test Set Activity Mean Vs Training Set Activity Mean. Journal of Chemical Information and Modeling, Vol, 48, pp, 2140-2145.

[141] Consonni V., Ballabio D.. and Todeschini R.. (2009). Comments on the Definition of the Q2 Parameter for QSAR Validation. Journal of Chemical Information and Modeling, Vol, 49, pp, 1669–1678.

[142] Lin L. I.–K.. (1989). A Concordance Correlation Coefficient to Evaluate Reproducibility. Biometrics, Vol, 45, pp, 255–268.

[143] Lin L. I.-K.. (1992). Assay Validation Using the Concordance Correlation Coefficient. Biometrics, Vol, 48, pp, 599–604.

[144]D.J. Bouveresse, et al, 2004. Sélection d'échantillons représentatifs par des méthodes chimiométriques, Spectra analyse, 33, [23-27].

[145] DE Knuth, 1997. "The art of computer programming", Vol, 2 (3rded,), Boston: Addison Wesley.

[146] G.D Tourassi, E. D Frederick, M. K Markey, E. Carey. Jr. Floyd, 2001. "Application of the mutual information criterion for feature selection in computer-aided diagnosis", Medical Physics, Vol, 28(12), pp, 2394-2402.

[147] R.W Kennard., L.A Stone, 1969, Computer Aided Design of Experiments, Technometrics, 11 (1), [137-148]

[148] A. BOUAKKADIA, 2016. Modélisation de quelques propriétés contrôlant l'évolution dans l'environnement d'une série d'herbicides.

[149] R D Snee, 1977. "Validation of Regression Models: Methods and Examples", Technometrics, Vol, 19, pp, 415- 428.

[150] F Despagne, DL. Massart, 1998. "Neural networks in multi variate calibrage", Analyst, Vol, 123, pp, 157-178.

[151] D. Sprevak, F. Azuaje, H. Wang, 2004. "Anon-random data sampling method For classification model assessment", In 17th international conference on pattern recognition,Vol,3,pp, 406-409.

[152] Y,Y Ren,, H, X, Liu, X, J, Yao, M, C, Liu. 2007. "Prediction of ozone tropospheric degradation rate constants by projection pursuit regression", Anal, Chim, Acta, Vol, 589, pp, 150–158.

[153] F A. Graybill, 1976. "Theory and Application of the Linear Model", Duxbury, North Scituate, Mass., pp, 231 – 236.

[154] C. Jin, B. Lei, J. Li, S. Li, Y. Shenb, X. Yao, 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 Comb, Sci*, Vol, 27, pp, 1318–1325.

[155] Hidalgo-Rodríguez, M., Fuguet, E., Ràfols, C., & Rosés, M. (2012). Modeling nonspecific toxicity of organic compounds to the fathead minnow fish by means of chromatographic systems. Analytical chemistry, 84(7), 3446-3452.

[156] Katritzky, A. R., Lobanov, V. S., & Karelson, M. (1995). QSPR: the correlation and quantitative prediction of chemical and physical properties from structure. Chemical Society Reviews, 24(4), 279-287.

[157] Katritzky, A. R., Karelson, M., & Lobanov, V. S. (1997). QSPR as a means of predicting and understanding chemical and physical properties in terms of structure. Pure and Applied Chemistry, 69(2), 245-248.

[158] Tham, S. Y., & Agatonovic-Kustrin, S. (2002). Application of the artificial neural network in quantitative structure–gradient elution retention relationship of phenylthiocarbamyl amino acids derivatives. Journal of pharmaceutical and biomedical analysis, 28(3-4), 581-590.

[159] Maret, L. (2013). Application de la technique de thermodésorption pour l'analyse de 93 COV et le screening des COV dans l'air des lieux de travail (Doctoral dissertation).

[160] Ecole Technique Supérieure du Laboratoire (ETSL), 2012, Chromatographie en phase gazeuse aspect théorique et appareillage, [Pdf],

<URL http://www.etsl.fr/Doc_pdf/Catalogue_FC2012>.

[161] Jalali-Heravi, M., & Garkani-Nejad, Z. (2002). Use of self-training artificial neural networks in modeling of gas chromatographic relative retention times of a variety of organic compounds. Journal of Chromatography A, **945** (1-2), 173-184.

[162] Zheng, F., Bayram, E., Sumithran, S. P., Ayers, J. T., Zhan, C. G., Schmitt, J. D., ... & Crooks, P. A. (2006). QSAR modeling of mono-and bis-quaternary ammonium salts that act as antagonists at neuronal nicotinic acetylcholine receptors mediating dopamine release. Bioorganic & medicinal chemistry, **14** (9), 3017-3037.

[163] Guha, R., & Jurs, P. C. (2005). Interpreting computational neural network QSAR models: a measure of descriptor importance. Journal of chemical information and modeling, 45(3), 800-806.

[164] Nguyen, N. T. D., Kummer, E., Dubost, J. P., Convard, T., Barbanton, J., & Carpy, A. (1999). Dossier-Importance of lipophilicity in molecular design-La probabilite d'hydratation moleculaire: un nouveau concept pour le calcul du log P d'une molecule a partir de sa structure 3D (text in. Analusis, 27(1), 29-31.

[165] Katritzky, A. R., Ramsden, C. A., Joule, J. A., & Zhdankin, V. V., (2010). Handbook of heterocyclic chemistry.Elsevier

[166] Michałowicz, J., & Duda, W. (2007). Phenols--Sources and Toxicity. Polish Journal of Environmental Studies, 16(3).

[167] Mackay, D., Shiu, W. Y., & Ma, K. C. (1997). Illustrated handbook of physical-chemical properties of environmental fate for organic chemicals (Vol. 5). CRC press.

[168] Lorentz, H. A. (1880). Ueber die Beziehung zwischen der Fortpflanzungsgeschwindigkeit des Lichtes und der Körperdichte.
Annalen der Physik, 245(4), 641-665.
https://onlinelibrary.wiley.com/doi/abs/10.1002/andp.18802450406

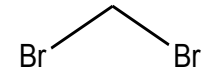
[169] Hansen, C., Telzer, B. R., & Zhang, L. (1995). Comparative QSAR in toxicology: examples from teratology and cancer chemotherapy of aniline mustards. Critical reviews in toxicology, 25(1),67-89. https://www.tandfonline.com/doi/abs/10.3109/10408449509089887

[170] EUROFORUM Traitement des effluents gazeux Journée du mercredi 23 janvier 2002 Réduire les émanations de COV dansl'atmosphère:Quelles solutions techniques ?

[171] Kubinyi, H. (Ed.). (1993). 3D QSAR in drug design: volume 1: theory methods and applications (Vol. 1). Springer Science & Business Media.

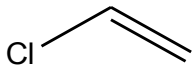
Annexes

I-L'ensemble des molécules étudiées :



Dibromomethan

74-95-3
(1)

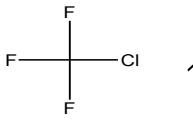


chloroethene

71-55-6
(2)

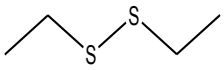
1,3-Dibromopropane

109-64-8
(3)



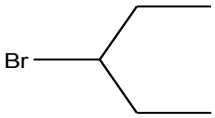
trifluoromethyl chloride

75-72-9
(4)



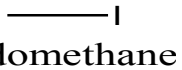
Ethyl disulfide

287-53-6
(5)



3-Bromopentane

1809-10-5
(6)

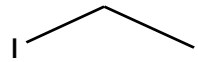


Iodomethane

74-88-4
(7)

1-Pentanethiol

110-66-7
(8)



Ethyl iodide

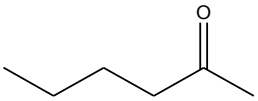
75-03-6
(9)

1-Bromobutane

109-65-9
(10)

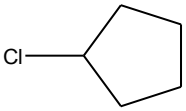
1-Bromopentane

110-53-2
(11)



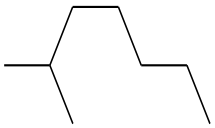
2-Hexanone

591-78-6
(12)



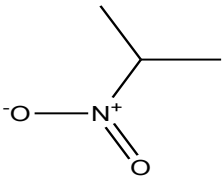
Cyclopentylchloride

930-28-9
(13)



2-Methylheptane

592-27-8
(14)

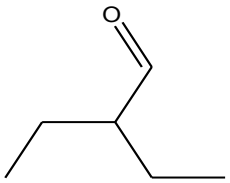


2-Nitropropane

79-46-9
(15)

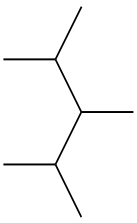
Butyl formate

592-84-7
(16)



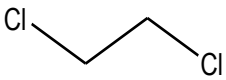
2-Ethylbutyraldehyde

97-96-1
(17)



2,3,4-Trimethylpentane

565-75-3
(18)

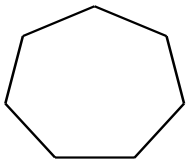


1,2-Dichloroethane

107-06-2
(19)

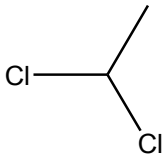
Propyl sulfide

111-47-7
(20)



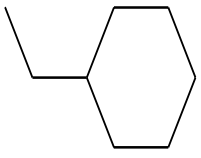
Cycloheptane

291-64-5
(21)



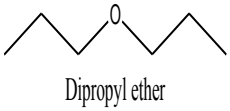
1,1-Dichloroethane

75-34-3
(22)



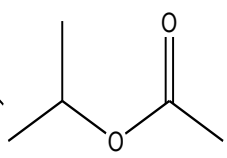
Ethylcyclohexane

1678-91-7
(23)



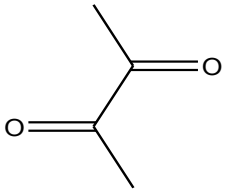
Dipropyl ether

111-43-3
(24)



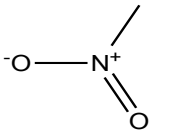
Isopropyl acetate

108-21-4
(25)



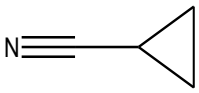
2,3-Butanedione

431-03-8
(26)



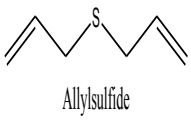
Nitromethane

75-52-5
(27)



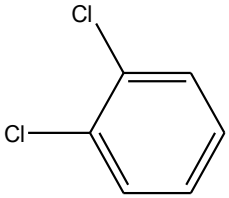
Cyclopropylcyanide

5500-21-0
(28)



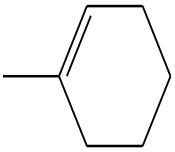
Allylsulfide

10152-76-8
(29)



1,2-Dichlorobenzene

95-50-1
(30)

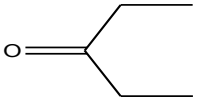


1-Methylcyclohexene

591-49-1
(31)

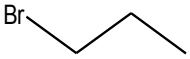
Methanol

67-56-1
(32)



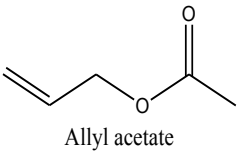
3-Pentanone

96-22-0
(33)



1-Bromopropane

106-94-5
(34)

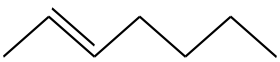


Allyl acetate

591-87-7
(35)

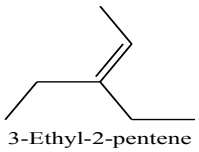
Valeraldehyde

110-62-3
(36)



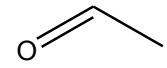
trans-2-Heptene

14686-13-6
(37)



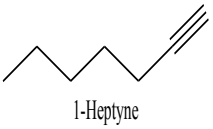
3-Ethyl-2-pentene

816-79-5
(38)



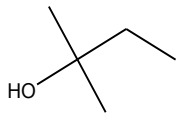
Acetaldehyde

75-07-0
(39)



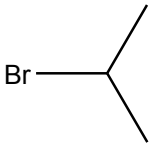
1-Heptyne

628-71-7
(40)



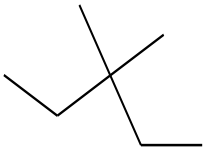
2-Methyl-2-butanol

75-85-4
(41)



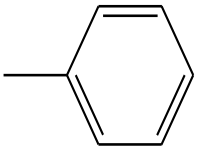
2-Bromopropane

75-26-3
(42)



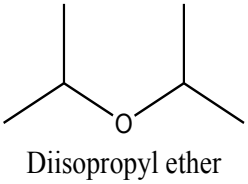
3,3-Dimethylpentane

562-49-2
(43)



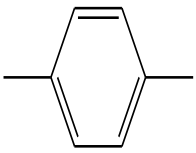
Toluene

108-88-3
(44)



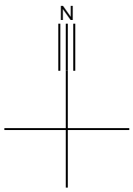
Diisopropyl ether

108-20-3
(45)



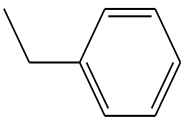
p-Xylene

106-42-3
(46)



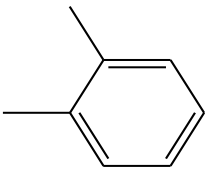
Trimethylacetonitrile

630-18-2
(47)



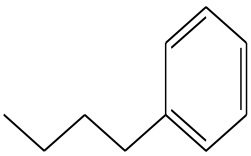
Ethyl benzene

100-41-4
(48)



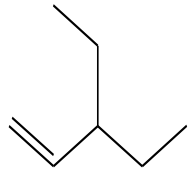
o-Xylene

95-47-6
(49)



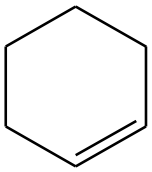
n-Butylbenzene

104-51-8
(50)



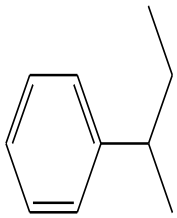
3-Ethyl-1-pentene

4038-04-4
(51)



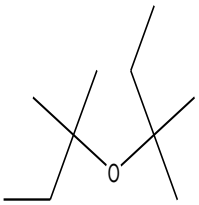
Cyclohexene

110-83-8
(52)



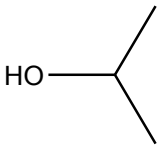
sec.-Butylbenzene

135-98-8
(53)



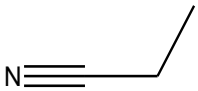
Methyltert.-butyl ether

1634-04-4
(54)



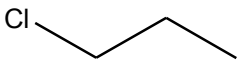
Isopropanol

67-63-0
(55)



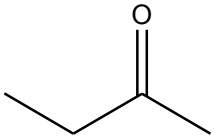
Propionitrile

107-12-0
(56)



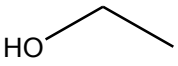
1-Chloropropane

540-54-5
(57)



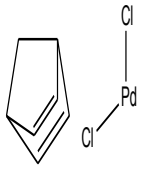
2-Butanone

78-93-3
(58)



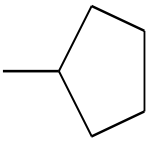
Ethanol

64-17-5
(59)



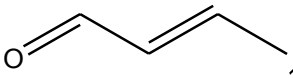
Dichloro(norbornadiene)palladium

12317-46-3
(60)



Methylcyclopentane

96-37-3
(61)

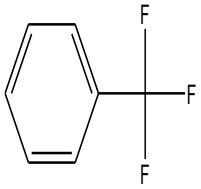


Crotonaldehyde

4170-30-3
(62)

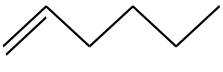
Hexane

110-54-3
(63)



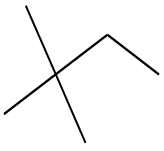
Trifluoromethyl-benzene

729-81-7
(64)



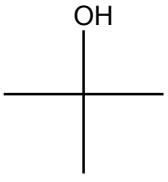
1-Hexene

592-41-6
(65)



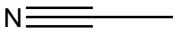
2,2-Dimethylbutane

75-83-2
(66)



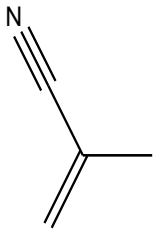
2-Methyl-2-propanol

75-65-0
(67)



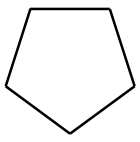
Acetonitrile

75-05-8
(68)



Methacrylonitrile

126-98-7
(69)

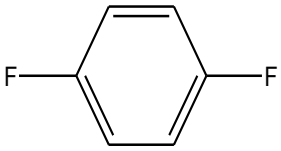


Cyclopentane

287-92-37
(70)

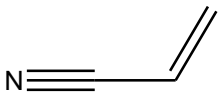
Pentane

109-66-0
(71)



1,4-Difluorobenzene

540-36-3
(72)

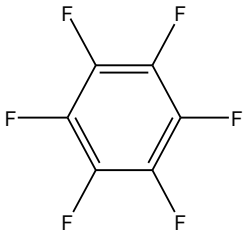


Acrylonitrile

107-13-1
(73)

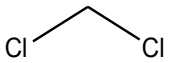
1-Pentene

109-67-1
(74)



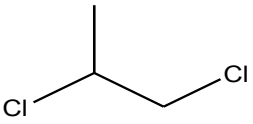
Hexafluorobenzene

392-56-3
(75)



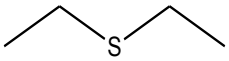
Dichloromethane

75-09-2
(76)



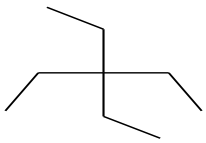
1,2-Dichloropropane

78-87-5
(77)



Ethyl sulfide

352-93-2
(78)

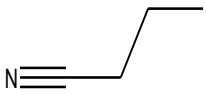


3,3-Diethylpentane

1067-20-5
(79)

Heptane

142-82-5
(80)

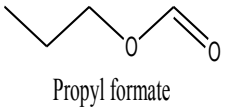


Butyronitrile

109-74-0
(81)

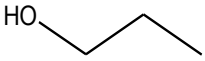
1-Octyne

629-05-0
(82)



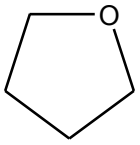
Propyl formate

110-74-7
(83)



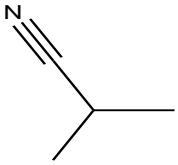
1-Propanol

71-23-8
(84)



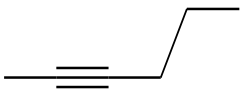
Tetrahydrofuran

109-99-9
(85)



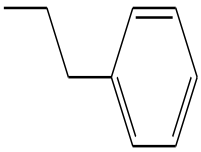
Isobutyronitrile

78-82-0
(86)



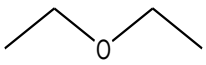
2-Hexyne

764-35-2
(87)



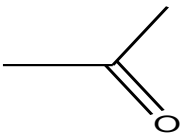
Propyl benzene

103-65-1
(88)



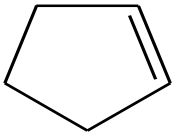
Diethyl ether

60-29-7
(89)



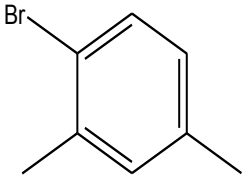
Acetone

67-64-1
(90)



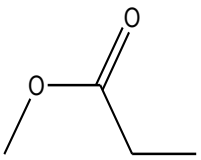
Cyclopentene

142-29-0
(91)



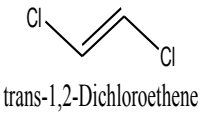
4-Bromo-m-xylene

583-70-0
(92)



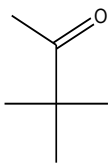
Methyl propionate

554-12-1
(93)



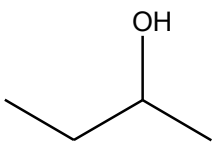
trans-1,2-Dichloroethene

156-60-5
(94)



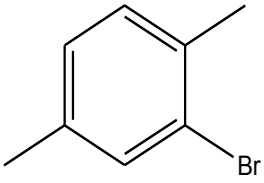
3,3-Dimethyl-2-butanone

75-97-8
(95)



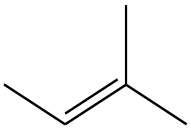
sec.-Butanol

78-92-2
(96)



2-Bromo-p-xylene

553-94-6
(97)

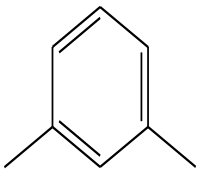


2-Methyl-2-butene

513-35-9
(98)

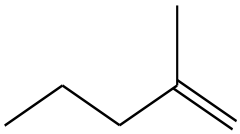
1,1-Dimethylcyclohexane

590-66-9
(99)



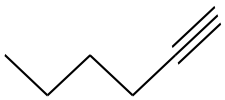
m-Xylene

108-38-3
(100)



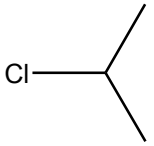
2-Methyl-1-pentene

736-29-1
(101)



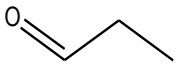
1-Hexyne

693-02-7
(102)



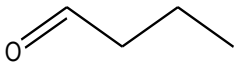
2-Chloropropane

75-29-6
(103)



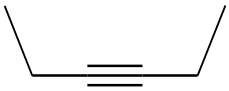
Propionaldehyde

123-38-6
(104)



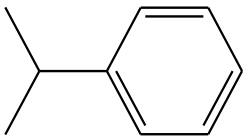
Butyraldehyde

123-72-8
(105)



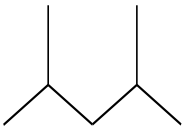
3-Hexyne

592-76-7
(106)



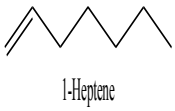
Cumene

108-08-7
(107)



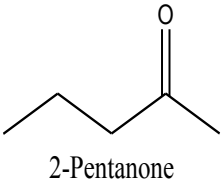
2,4-Dimethylpentane

98-82-8
(108)



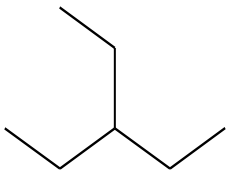
1-Heptene

928-49-4
(109)



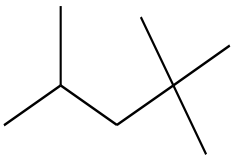
2-Pentanone

123-72-8
(110)



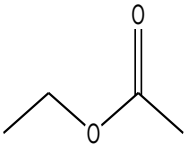
3-Ethylpentane

617-78-7
(111)



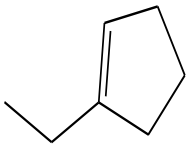
2,2,4-Trimethylpentane

540-84-1
(112)



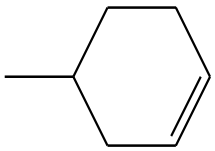
Ethyl acetate

141-78-6
(113)



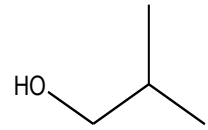
1-Ethylcyclopentene

2146-38-5
(114)



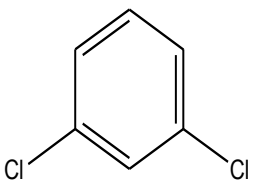
4-Methylcyclohexene

591-47-9
(115)



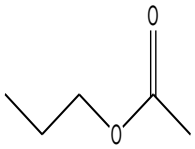
2-Methyl-1-propanol

78-83-1
(116)



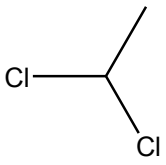
1,3-Dichlorobenzene

541-73-1
(117)



Propyl acetate

109-60-4
(118)

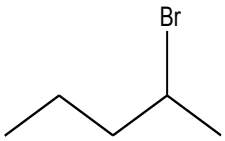


dichloroethane

1300-21-6
(119)

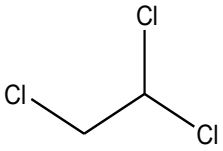
1-Butanol

71-36-3
(120)



2-Bromopentane

107-81-3
(121)



1,1,2-Trichloro-ethane

79-00-5
(122)