الجمهورية الجزائرية الديمقراطية الشعبية
République algérienne démocratique et populaire
وزارة التعليم العالي والبحث العلمي
Ministère de l'enseignement supérieur et de la recherche scientifique

UNIVERSITÉ BADJI MOKHTAR
- ANNABA-

جامعة باجي مختار- عنابة

Année / 2019-2020

Faculté des sciences
Département de chimie

THÈSE
Présentée en vue de l'obtention du diplôme de Doctorat

Option : *Chimie de l'environnement et QSAR*

THÈME

Méthodes assistées par ordinateur pour la prédiction de la solubilité et la température d'ébullition des phénols et des indices de rétention des pyrazines

Présentée par : Mme Kherouf Soumaya

Devant le jury :

Présidente	Mme.BIDJOU-HAIOUR Chahra	Pr	Université Badji Mokhtar Annaba
Rapporteur	Mr.MESSADI Djelloul	Pr	Université Badji Mokhtar Annaba
Examinateus	Mr. MERDES Rachid	Pr	Université 08 Mai 1945 Guelma
	Mr. GHEID Abd Elhak	Pr	Université Med chérif Messadia Souk Ahras
	Mme. SEDRATI Nassima	MCA	Université Badji Mokhtar Annaba

À celle qui m'a donné la vie, le symbole de tendresse, qui s'est sacrifiée pour mon bonheur et ma réussite,

À ma mère *…*

A mon père, *qui a veillé tout au long de ma vie pour m'encourager.*

Que Dieu les garde et les protège.

A mon mari *pour le soutien, l'encouragement et le support qu'il m'a apportés tout au long de la réalisation de ce travail.*

À mon prince Mohammed Joud *et mes frères* ***Achref, Djallel****, et* ***Anis****.*

Je dédie ce travail.

Remerciements

Ce travail a été réalisé au ***Laboratoire de Sécurité Environnementale et Alimentaire (LASEA)*** *de l'université Badji Mokhtar-Annaba*

J'exprime ma gratitude au directeur du LASEA Monsieur le professeur ***MESSADI Djelloul*** *qui m'a encadrée tout au long de cette thèse et m'avoir fait bénéficier de l'étendue de ses connaissances. Je le remercie très chaleureusement pour son engagement, sa disponibilité permanente et les encouragements qu'il a su me prodiguer jusqu'au dernier jour ;*

J'exprime également ma profonde gratitude à Mme BIDJOU HAIUOR Chahra professeur à l'université Badji Mokhtar-Annaba, pour nous avoir fait l'honneur de présider le jury de cette thèse.

Mes vifs remerciements vont également aux membres de jury : à Mr MERDES Rachid, professeur à l'université de Guelma, Mr GHEID Abd Elhak, professeur à l'université de Souk Ahras, et Mme Sedrati Nassima Maître de conférences à l'université Badji Mokhtar-Annaba, pour avoir accepté d'examiner et de juger ce travail.

Résumé

Les relations quantitatives structure-propriété (QSPR) sont très utiles pour comprendre comment la structure chimique se corrèle avec la propriété des produits chimiques naturels et synthétiques. Dans le présent travail, des modèles QSPR ont été développés pour la prédiction de trois caractéristiques environnementales importantes d'un ensemble hétérogène de composés.

Des techniques de régression linéaire multiple (MLR) et de réseau neural artificiel (ANN) ont été utilisées pour les études quantitatives de la relation structure-solubilité (QSSR) de 68 phénols basés sur des descripteurs moléculaires calculés à partir des structures 3D optimisées.

Ainsi qu'une approche hybride algorithme génétique/ régression multilinéaire a été appliquée pour modéliser la température d'ébullition de 56 phénols, et l'indice de rétention de 113 pyrazines.

Il s'avère que les meilleurs modèles dont la stabilité est confirmée par la validation interne Q^2_{LOO} « leave-one-out » est capable de décrire environ 89,33 % de la variance de la solubilité aqueuse expérimentale, 89,33% de la variance de la température d'ébullition de 56 phénols et 99,51% pour l'indice de rétention des pyrazines.

Mots clés : QSPR- Solubilité- Phénols- MLR- RNA- Température d'ébullition – Indice de rétention- Pyrazines.

Abstract

Quantitative Structure-Property Relationships (QSPRs) are very useful for understanding how the chemical structure correlates with the property of natural and synthetic chemicals. In this work, QSPR models have been developed for the prediction of three important environmental features of an heterogeneous set of compounds.

Multiple Linear Regression (MLR) and Artificial Neural Network (ANN) techniques were used for quantitative studies of the structure-solubility (QSSR) of 68 phenols based on molecular descriptors calculated from optimized 3D structures.

As well as an hybrid approach genetic algorithm / multilinear regression was applied to model the boiling temperature of 56 phenols, and the retention index of 113 pyrazines.

It turns out that the best models whose stability is confirmed by internal validation Q^2_{LOO} "leave-one-out" is able to describe about 89.33% of the variance in experimental aqueous solubility, 89.33% of variance of the boiling temperature of phenols and 99.51% for the retention index of pyrazines.

Key words: QSPR- Solubility- Phenols- MLR- RNA- Boiling temperature - Retention index- Pyrazines.

ملخص

العلاقات الكمية بين الهيكل والخاصية (QSPR) مفيدة للغاية لفهم كيفية ارتباط التركيب الكيميائي بخاصية المواد الكيميائية الطبيعية والاصطناعية. في هذا العمل ، تم تطوير نماذج QSPR للتنبؤ بثلاث خصائص بيئية مهمة لمجموعة غير متجانسة من المركبات.

تم استخدام تقنيات الانحدار الخطي المتعدد (MLR) والشبكة العصبية الاصطناعية (ANN) في الدراسات الكمية لعلاقة 68 فينول بالذوبانية (QSSR) على أساس الواصفات الجزيئية المثلى المحسوبة من الهياكل ثلاثية الأبعاد.

وكذلك تم تطبيق النهج المختلط للخوارزمية الوراثية / الانحدار متعدد الخطوط على نموذج درجة حرارة الغليان لـ 56 فينول ، ومؤشر الاحتفاظ بـ 113 بيرازين.

اتضح أن أفضل النماذج التي تم تأكيد ثباتها من خلال التحقق من فاعليتها الداخلية "Q^2_{LOO}" قادرة على وصف حوالي 89.33٪ من تباين الذوبانية المائية التجريبية، و89.33٪ من تباين درجة حرارة الغليان من 56 فينول و99.51 ٪ لمؤشر الاحتفاظ للبيرزينات.

الكلمات المفتاحية -QSPR :الذوبانية- الفينولات -MLR- RNA -درجة حرارة الغليان - مؤشر الاستبقاءPyrazines -

SOMMAIRE

Chapitre 3 : Modélisation moléculaire

Chapitre 4 : QSAR /QSPR

LISTE DES TABLEAUX

LISTES DES FIGURES

SYMBOLES ET ABREVIATIONS

AG: Algorithme génétique.

AQUAFAC Aqueous functional group activity coefficients.

AG-SSV Algorithme génétique-Sélection des sous ensembles de variables.

ATSDR: Agence des substances toxiques et du registre des maladies(Agence for Toxic Substances and Disease Registry).

CAS : Chemical Abstracts Service.

CCC : Coefficient de corrélation de concordance.

CGA : Contribution de groupe additive.

CGL : Chromatographie gaz-liquide.

CGS : Chromatographie gaz-solide.

COV : Composés organiques volatils.

CPG : Chromatographie en phase gazeuse.

DA : Domaine d'application.

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.

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

ei_std: Résidu standardisé.

F : Statistique de Fisher.

FEMA: Flavor and extract manufacturers.

FIT: Fonction de KUBINYI.

FIV: Facteur d'inflation de la variance.

GRAS Generally recognized as safe.

HF : Hartree –Fock.

hii : Eléments diagonaux de la matrice chapeau.

IR : Indice de rétention.

IUPAC International union of pure and applied chemistry.

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

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

MLR : Régression linéaire multiple.

MM+: Mécanique Moléculaire (+).

MM2 Mécanique moléculaire 2.

n: Dimension de la population (échantillon).

n-p : Nombre de degrés de liberté.

OCDE : **O**rganisation de **C**oopération et de **D**éveloppement **E**conomiques.

OLS : Ordinary least square (moindres carrés ordinaires).

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

PEG : Polyethylene glycols.

PM3 : Parametrization Method 3.

PRESS : Somme des carrés des erreurs de prédiction.

PS : Polysiloxane.

Q^2_{LOO} Coefficient de prédiction.

QSAR: Relations Quantitatives Structure/ Activité.

QSPR : Relations Quantitatives Structure/ Propriété.

Q^2_{Yscr} Coefficient de prédiction des modèles où les Y sont randomisés.

R^2 : Coefficient de détermination.

REACH : en**R**egistrement, **E**valuation et **A**utorisation des produits **CH**imiques.

RLM (MLR): Régression linéaire multiple.

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

RNA: Réseaux de neurones artificiels.

R^2_{Yscr}	Coefficient de détermination des modèles où les Y sont randomisés.
S :	Erreur standard.
SCE :	Somme des carrés des écarts.
SCT :	Somme des carrés totale.
Sw:	Solubilité dans l'eau (solubility in water).
t :	t de Student.
t_i :	Résidu studentisé externe.
UNIFAC :	UNIversal Function Activity Coefficient.
y_i:	Valeur observée.
$\hat{y}_i$:	Valeur estimée.
$\hat{y}_{(i)}$:	Valeur prédite.

İNTRODUCTİON GÉNÉRALE

Introduction générale

Les produits chimiques constituent certainement l'une des grandes évolutions industrielles du XXe siècle. Qu'il s'agisse de détergents, de fibres synthétiques, ou de matériaux, ils sont devenus incontournables dans notre vie. On en trouve partout, sans même nous en rendre compte. Parfois, être en contact avec certaines de ces substances chimiques à des moments particuliers de notre vie peut avoir des effets sur notre santé. Face à l'omniprésence des substances chimiques dans notre quotidien et devant les inquiétudes liées aux effets secondaires et leur devenir dans l'environnement, l'enjeu est de déterminer l'équilibre acceptable entre les bénéfices apportés et la prise de risque pour la santé humaine et l'environnement.

La détermination expérimentale des propriétés de composés chimiques représente une étape initiale importante dans la maîtrise des risques, ce processus est particulièrement difficile et contraignant pour des raisons de temps, de coûts, d'éthique (essais sur animaux) et de faisabilité au niveau recherche et développement.

Pour cette raison le recours aux méthodes alternatives moins coûteuses et plus rapides est indispensable. Ainsi l'utilisation des **R**elations **Q**uantitatives **S**tructure / **P**ropriété (QSPR) qui s'avère d'un grand intérêt, est désormais recommandée dans les nouvelles règlementations afin d'obtenir les données nécessaires à l'enregistrement des substances.
L'élaboration des modèles mathématiques QSPR reliant les propriétés physicochimiques à la structure moléculaire permet, d'une part, d'expliquer l'origine de ces propriétés et, d'autre part, de les prédire pour des molécules pour lesquelles les données expérimentales ne sont pas disponibles.

Des principes ont même été mis en place par l'OCDE [1] en 2009 afin de valider scientifiquement et réglementairement les modèles QSAR/QSPR. Cela dans le but d'augmenter la confiance en ces modèles mathématiques prédictifs et de favoriser ainsi leur utilisation par les industriels.

Nous nous proposons, de développer et d'évaluer le potentiel de tels modèles QSPR/ QSRR pour la prédiction : de la solubilité et la température d'ébullition des phénols ainsi que l'indice de rétention des pyrazines séparées, tour à tour, sur deux colonnes de polarités différentes.

Introduction générale

Le présent manuscrit de thèse comporte deux parties importantes:

La première partie présente une étude bibliographique et comporte 4 chapitres.
Le premier est un aperçu sur les composés étudiés; le deuxième porte sur les propriétés étudiées, et les deux derniers décrivent les méthodes de modélisation moléculaire et les principes de la technique QSAR/QSPR

La deuxième partie est consacrée aux applications QSPR et à l'analyse des résultats : elle comporte 2 chapitres de Modélisation : la modélisation de la solubilité aqueuse et la température d'ébullition des phénols. Ainsi que la prédiction des indices de rétentions des pyrazines

CHAPİTRE 1 ÉTAT DE L'ART DES COMPOSÉS ÉTUDİÉS

İntroduction

- Phénols
- Pyrazines

Conclusion

Introduction

Le développement continu de la chimie se traduit par la disponibilité d'une grande variété de composés d'intérêt chimique, biologique, pharmacologique et industriel. Les phénols et les pyrazines sont parmi les éléments les plus présents dans le domaine industriel en particulier dans l'agrochimie, la pétrochimie et la pharmacochimie.

La caractérisation de ces composés, est cruciale pour la recherche de nouveaux composés, répond à une demande toujours croissante de molécules originales.

Dans ce chapitre, nous nous intéressant à l'état de l'art sur le phénol et ses dérivés en faisant ressortir leurs propriétés, puis nous présenterons les propriétés des Pyrazines.

1. Phénols

1.1 Historique :

En 1650, Johann Rudolf Glauber, un scientifique allemand découvrit le phénol à l'état impur à partir de la distillation du goudron de houille. Il le décrivit comme "une huile vive et rouge sang qui assèche et guérit tous les ulcères humides". Environ deux siècles après, en 1834 son concitoyen Friedrich Ferdinand Runge parvint à isoler pour la première fois le phénol et il le nomma "acide carbolique" (Karbolsäure). Ensuite, En 1841 Auguste Laurent, un chimiste français fut le premier à préparer le phénol pur. Il le nomme acide phénolique. En 1843, le chimiste français, Charles Fréderic Gerhard inventa le nom de «phénol», nom qui a détrôné celui d'acide phénique. Le mot « phénol» est tout à fait conforme à la nomenclature officielle [1].

Le phénol a été produit, durant la première guerre mondiale, pour des applications militaires. Il est utilisé dans la synthèse des résines, plus tard il est devenu un objet de nombreuses études et plusieurs applications. Il intervient dans de nombreux procédés de fabrication de divers composés: sous-produits de raffinage de pétrole, produits pharmaceutiques, colorants [2]…

1.2 Définition

Le phénol est un composé organique aromatique qui se compose d'un noyau benzénique relié à un groupement hydroxyle (–OH). Bien qu'il ait une fonction alcool, le phénol a des propriétés uniques et n'est pas classé comme un alcool. Sa structure est relativement simple, le groupement hydroxyle est lié à un atome de carbone du cycle benzénique (figure 1)

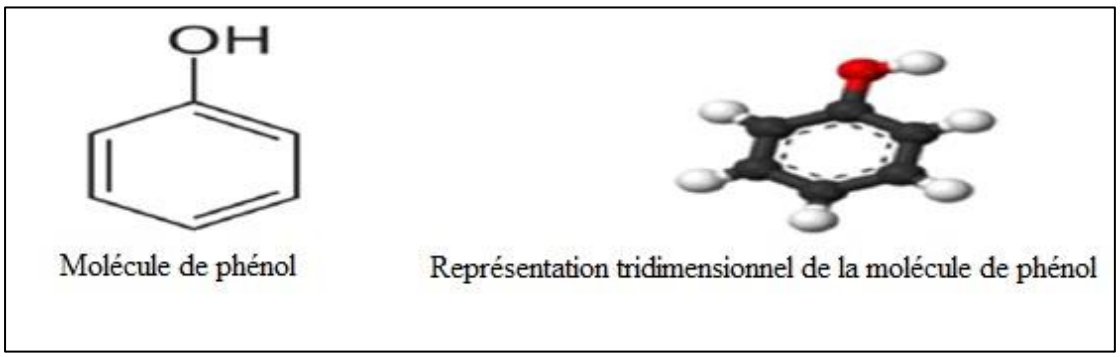


Figure.1 Phénol

1.3 Classification des phénols

Le terme composé phénolique englobe une grande variété de substances possédant un ou plusieurs noyaux aromatiques, substitués par un ou plusieurs groupements hydroxyles et / ou méthoxyles [3]. Cette définition présente une dizaine de familles de phénols. Nous limitons notre classification à 2 groupes principaux.

1.3.1. Les phénols simples

Les phénols simples englobent toutes les molécules hydroxylées diversement substituées du noyau de phénol. Selon cette définition plusieurs phénols appartiennent à ce groupe, par exemple : le phénol, les alkylphénols, les nitrophénols et les chlorophénols.

Les chlorophénols ont un noyau phénolique substitué de différentes manières par un ou plusieurs atomes de chlore remplaçant l'hydrogène du cycle aromatique. Par conséquent nous comptons 19 congénères de chlorophénols.

Les chlorophénols servent pour produire des intermédiaires dans la synthèse des phénols plus chlorés ou de dérivés des chlorophénols comme les herbicides, les colorants, les pigments et les résines phénoliques. Dans les stations d'épuration, les phénols peuvent se retrouver dans les eaux de surface destinées à la préparation d'eau potable. Après désinfection de cette eau par le chlore ou ses dérivés, les composés phénoliques peuvent générer des chlorophénols. Aussi, certains chlorophénols sont utilisés comme fongicides, antiseptiques, désinfectants et agents anti-gommage pour l'essence. D'autre part, comme la plupart des composés phénoliques, les chlorophénols sont toxiques pour les milieux aquatiques. A certains seuils ils sont capables d'entraîner des changements histopathologiques mutagènes et cancérigènes [4].

1.3.2 Les polyphénols

Chimiquement, les polyphénols sont des composés phénoliques à hauts poids moléculaires. Ils se composent d'un ou plusieurs cycles benzéniques portant un ou plusieurs groupements hydroxyles et autres constituants [5].

Les polyphénols regroupent un vaste ensemble de substances chimiques. Ils peuvent exister dans diverses substances naturelles: sous forme d'anthocyanine dans les fruits rouges, de proanthocyanidines dans le chocolat, d'acides cafeoylquinique et feruloylquinique dans le café, de quercitrine dans les pommes, etc.… [6].

Le polyphénol se présente comme un composé de base, constituant de plusieurs produits synthétiques simples possédant une fonction alcool supplémentaire comme l'hydroquinone, catéchol et sous forme de polyphénols polymériques comme les colorants, les plastiques et les résines à base de bisphénol.

Concernant les polyphénols simples, dans une perspective écologique, le pyrocatéchol et l'hydroquinone sont dangereux pour les écosystèmes et notamment pour les eaux car ils sont peu biodégradables et en partie toxiques pour les poissons. Le contact cutané avec le catéchol ou l'hydroquinone provoque une dermatite eczémateuse chez l'homme. Dans le cas de polyphénols polymériques, par exemple le bisphénol nuit beaucoup au système hormonal de l'organisme. Cette molécule peut, en effet, agir en tant qu'analogue d'œstrogène dans les systèmes biologiques. Les œstrogènes sont des molécules organiques dérivées du cholestérol.

I.4 Origine, source et application

À l'état naturel les phénols sont présents dans le bois, l'aiguille de pin, l'urine des herbivores (sulfate phénolique) et le goudron de houille. Les composés phénoliques définissent un ensemble de substances appelées (tannin).Ce sont des alcools aromatiques qui proviennent de végétaux, ils sont souvent des constituants très odoriférants [7], tels que le thymol, l'eugénol, zingivérone, la vanilline....

- Le thymol présent dans le thym
- L'eugénol présent dans le clou de girofle
- zingivérone dans le gingembre
- La vanilline rencontrée dans la gousse de vanille et le benjoin de Siam.

Dans les laboratoires, à l'heure actuelle le phénol est généralement préparé par le procédé Hock qui consiste à oxyder l'isopropylbenzène par le dioxygène de l'air. Le sous-produit de la réaction est la propanone qui est également un produit important utilisé notamment comme solvant. Par la suite, le phénol est séparé de l'acétone [8].
Le phénol constitue un des grands produits intermédiaires de l'industrie chimique [9]. Environ deux millions de tonnes de phénol sont utilisées annuellement dans la communauté européenne [10]. Le phénol et ses dérivés sont couramment utilisés dans des fabrications très diverses mais principalement en synthèse organique tels que :

- Le raffinage des pétroles ;
- L'industrie des matières plastiques (phénoplastes, poly-époxydes, polycarbonates) ;
- L'industrie pharmaceutique comme produits désinfectants, antiseptiques, antifongiques, anesthésiques locaux, aspirine, paracétamol….
- L'industrie chimique comme agents de tannage, révélateurs photographiques, additifs des lubrifiants et des essences.
- Les matières explosives (mélinite à base d'acide picrique) ;
- La fabrication de détergents, pesticides, colorants, peintures, produits cosmétiques [9,11] etc…

Les trois principales applications des phénols résident dans la fabrication des résines phénoliques, du bisphénol A et du caprolactame.

1. 5. Propriétés physicochimiques

1.5.1 Propriétés physiques

Le phénol est un solide incolore cristallisé sous forme d'aiguilles dans les conditions ambiantes habituelles, il est plus ou moins facilement fusible, moins volatil que les composés correspondants ne contenant pas de groupement hydroxyle (benzène et homologues), et l'élévation du point d'ébullition est attribuable à l'association par pont hydrogène entre les groupes -OH des molécules. En plus, le phénol est hygroscopique et a une odeur âcre et douceâtre. Au contact de l'air ou sous l'influence de l'humidité le phénol s'oxyde légèrement pour donner des traces de quinone. Il prend alors une couleur rose, puis rouge

Il est miscible dans l'eau. Sa solubilité est limitée à 97g/l à 20°C. Il est aussi très soluble dans plusieurs solvants organiques tels que l'acétone, l'éthanol, l'oxyde de diéthyle. Certaines propriétés du phénol sont regroupées dans le tableau 1.

Tableau 1 : Propriétés physiques du phénol [12]

T° fusion	43°C
T° ébullition	182°C
Solubilité à 20°C	97g l^{-1}
Masse volumique	1,073 g cm^{-3}
T° d'auto-inflammation	715°C
Point d'éclair	79°C
Limites d'explosivité dans l'air	1,36-10% vol
Pression de vapeur saturante à20°C	47 Pa
Point critique	61,3bar à 421,05°C
Temps de demi-vie dans l'air	env.20 h
Temps de demi-vie dans l'eau	env.55 h

1.5.2. Propriétés chimiques

Le phénol est incolore, cependant il se colore rapidement à l'air par oxydation. Sa formule chimique est la suivante **:** C_6H_5OH [13].

Le phénol réagit vivement avec des oxydants puissants comme les peroxydes. Vers 800°C et en présence de zinc, la molécule de phénol se réduit en benzène. A haute température, le phénol pur se décompose entièrement en oxyde de carbone, carbone et hydrogène.

A chaud, le phénol liquide attaque certains métaux, tels que le plomb, le zinc, l'aluminium… et aussi certains plastiques, comme le polyéthylène [14].

En solution, le phénol forme un acide très faible. Par contre son acidité est plus forte que celle des alcools (le pKa à 25°C du couple phénol/phénolate est de l'ordre de 9,9). Le phénol peut perdre un ion hydrogène et l'ion phénolate (phénoxyde) se stabilise dans la solution.

Tableau 2 : Propriétés chimiques du phénol

Formule brute	C_6H_6O
Masse molaire	94,1112 ± 0,0055 g mol^{-1} C: 76,57%, H : 6,43% et O : 17%
pKa	9 ,99 (phénol\ phénolate)
Moment dipolaire	1,22 ± 0,008 D
Diamètre moléculaire	0,55 nm

1.6 - Toxicité des phénols

Le phénol est classé par l'Union Européenne comme mutagène catégorie III [15]. La majorité des phénols et de leurs dérivés sont des substances toxiques. Beaucoup d'entre eux sont classés comme des déchets dangereux, et certains d'entre eux sont connus ou soupçonnés d'être cancérogènes. Aussi le phénol est répertorié sur la liste prioritaire des substances dangereuses identifiées par ATSDR (Agence for Toxic Substance and Disease Registry) [16].

Le phénol est probablement le composé organique le plus souvent associé à des problèmes organoleptiques [17].

1.6.1. Impacts sur l'homme

Le phénol est rapidement absorbé (70 à 80% en 6 heures) par toutes les voies sensuelles, puis rapidement distribué dans les tissus. Le phénol dénature les protéines et détruit les parois cellulaires. Les organes ciblés sont le cerveau, les reins, le foie, les poumons et la muqueuse gastro-intestinale [12,18]. On distingue deux types d'intoxication qui sont : l'intoxication aiguë et l'intoxication chronique.

Les phénols chlorés aussi sont très toxiques pour les êtres humains et les animaux. En présence de chlore, le phénol forme des chlorophénols qui sont facilement absorbés par tractus gastro-intestinal provoquant une toxicité aiguë. Même pour des concentrations aussi faibles de l'ordre de 0,1 mg /L, les chlorophénols produisent un goût désagréable lorsqu'ils sont mélangés avec de l'eau potable [2].

La toxicité augmente avec le degré de chloration qui pourrait générer des composés de chlorophénols mutagènes et cancérigènes [8].
Le phénol et ses solutions concentrées exercent une action caustique sur la peau ; c'est d'abord une sensation de brûlure qui est ressentie, suivie d'une perte de sensibilité locale de la peau [19].

Le phénol peut provoquer d'autres problèmes :

- Maux de tête et vomissements.
- Faiblesse musculaire et étourdissement.
- Troubles de la vision et de l'audition.
- Respiration rapide et irrégulière et la mort pouvant survenir par défaillance respiratoire.
- Perte de conscience [2 ,20].

1.6.2. Impacts sur l'environnement

Le phénol entre comme intermédiaire dans plusieurs procédés de synthèse, de fabrication, et de transformation. Par conséquent, on le trouve dans les rejets de nombreuses industries. Le rejet du phénol dans la nature, sans traitement et sans contrôle peut modifier les écosystèmes aquatiques et causer des dommages aux ressources précieuses. La faune et la flore sont les principales cibles de ces effluents. Le phénol est un produit répandu et nuisible à la vie aquatique. Il est très toxique dans l'eau, polluant du sol et conduit à de nombreux effets indésirables sur l'environnement et sur la santé [21].

Dans l'atmosphère, les vapeurs du phénol sont plus lourdes que l'air et forment des mélanges explosifs sous l'effet de la chaleur. Le phénol s'oxyde à l'air et ce processus d'oxydation est accéléré par la lumière [22]. Une partie des vapeurs du phénol est lessivée par la pluie [23]. Une exposition excessive au phénol par inhalation ou par contact cutané peut causer des effets sur la santé, de cerveau, de système digestif, les yeux, le cœur, les reins, le foie, les poumons, les nerfs périphériques, la peau et l'enfant à naître. Le phénol est classé « Composé Organique Volatil »(COV) qui peut potentiellement contribuer à la formation d'ozone troposphérique et du smog photochimique.

Dans le sol, le phénol subit une dégradation microbienne aérobie ou anaérobie, de sorte que l'effet d'accumulation reste limité. Comme le phénol est soluble dans l'eau et modérément volatil, il est très mobile dans les sols. Par conséquent, le phénol peut être lessivé facilement des sols et ainsi contaminer la nappe phréatique.
Le phénol a tendance à se biodégrader rapidement dans le sol et dans les sédiments.

Dans l'eau Le phénol est plus lourd que l'eau et tend à se déposer. Il se dissout lentement et « même dilué » il continue de former des solutions toxiques [24]. Il a été détecté dans les eaux de surface, les eaux souterraines [25,26], eaux de pluie, les effluents industriels et les eaux de ruissellement urbaines. Le phénol est susceptible d'atteindre les sources d'eau potable en aval des rejets. Le phénol donne un goût désagréable même à faibles concentrations et des odeurs dans l'eau potable. La concentration maximale en phénol admissible est de 35 mg/kg [27].

Les phénols sont relativement persistants et récalcitrants à la biodégradation et beaucoup plus difficiles à dégrader que de nombreux autres contaminants organiques. La toxicité des eaux phénoliques est due à leur pouvoir réducteur élevé et à leur demande biochimique en oxygène qui confère à ces eaux les éléments de pollution [9].

.

2. Pyrazines

2.1 Historique

La première synthèse enregistrée d'une pyrazine est celle de la tétraphénylpyrazine réalisée par Laurent [28] en 1855.

Après une série d'expériences et d'articles scientifiques, en 1882 Wleugel [29], a proposé pour la pyrazine une structure cyclique à six chaînons analogue à la pyridine. Ensuite, deux chercheurs, Mason [30] et Wolff [31], suggèrent indépendamment le nom «pyrazine» pour désigner le noyau afin de rappeler l'analogie avec la pyridine

En 1888 la pyrazine, a été préparée pour la première fois en trace par Wolff [32] en chauffant de l'aminoacétaldéhyde diéthylacétal avec de l'acide oxalique anhydre à 110-190 ° C. Depuis ce temps-là plusieurs chercheurs ont préparé cette molécule mère à partir d'une série de réactions, et ils ont publié d'importantes revues de la chimie des pyrazines. Leur structure a été établie en 1893 par Wolff [33].

En 1978 la presse a annoncé la production de plus de 30 composés de pyrazines destinés pour la vente aux industries cosmétiques et de parfumeries. Aujourd'hui, la fabrication des pyrazines offre plus de 150 composés pyraziniques avec leurs différents dérivés [34].

2.2 Définition

Les pyrazines, sont des substances aromatiques actives très largement distribuées dans le règne animal et végétal, et très présentes dans l'arôme des aliments. Principalement dans beaucoup de légumes, insectes, crustacés, et organismes marins ; elles peuvent être synthétisées biologiquement par des micro-organismes pendant leur métabolisme primaire ou secondaire [35-38], ou chimiquement lors de la réaction de Maillard [39]. Cette réaction permet, grâce à divers réactions complexes, de former des pyrazines à partir d'acides aminés et de sucres: les acides aminés servent alors de source d'azotes et les sucres (glucides), fournissent les carbones qui viendront former la molécule de pyrazine.

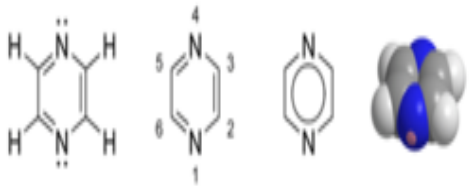
L'application des pyrazines synthétiques est toujours croissante dans les saveurs et les parfums industriels, en raison de leurs grands intérêts avec des concentrations très basses mais très efficaces.

Des milliers d'articles de recherches ont été édités, et de plus en plus (dans l'ordre de 100) des pyrazines ont été identifiées. Beaucoup de ces dernières sont intégrées maintenant dans la liste GRAS (Generally Recognized As Safe) et dans FEMA (Flavor and Extract Manufacturers) et sont destinées à l'usage alimentaire ; parmi les 60 pyrazines les plus connues pour cet usage, 30 sont préconisées dans les applications cosmétiques (FEMA). En plus, 20 dérivés de pyrazine ont un intérêt scientifique dans la biologie et la recherche pharmaceutique [40].

2.3. Structure

La pyrazine ou 1,4-diazine, de formule brute $C_4H_4N_2$, est un hétérocycle azoté et fondamental qui se rapproche de la structure du benzène où deux des groupements CH sont remplacés par des atomes d'azote.

Tableau 3 : La pyrazine

Nom IUPAC	Pyrazine	 **Structure de la pyrazine**
Synonymes	1,4-Diazabenzène 1,4-Diazine Paradiazine Piazine	

La structure de la pyrazine a été déterminée par diffraction des rayons X et par diffraction des électrons. La présence de deux atomes d'azote en position para qui sont plus électronégatifs que le carbone, rend la fonctionnalisation de ce composé plus difficile, ainsi la substitution électrophile n'est pas possible sur le cycle [34].

2.4. Propriétés physiques et chimiques :

- En raison de leur faible pression de vapeur, les pyrazines s'évaporent facilement [41].
- Les pyrazines sont formées par chauffage des aliments dans La réaction de Maillard [39].
- Au cours de la formation biologique, les pyrazines sont également formées de la réaction entre les acides aminés et les sucres [41].

- Les plus grandes quantités de pyrazines se forment à des températures comprises entre 120 et 150 ° C [41],
- Les pyrazines ne sont pas seulement formées dans des aliments chauffés, elles sont également produites dans les aliments fermentés pendant le processus de fermentation [41].
- La pyrazine fond à 55°C et donne la pipérazine $C_4H_8N_2$ par une hydrogénation complète [42].
- La présence et l'augmentation du nombre d'azotes dans un hétérocycle (comme la pyrazine) construit un composé idéal, en raison de ses nouvelles propriétés basées sur une haute densité, une chaleur de formation positive élevée et une haute stabilité thermique [43].
- La pyrazine est une dibase faible (pK1 = 0,57; pK2 = -5,51), comparativement à la pyridine (pyridine pKa = 5,2), ceci est dû à l'effet de l'introduction du deuxième azote [44].
- Les pyrazines sont des substances aromatiques très actives et très volatiles. Elles confèrent aux aliments des propriétés gustatives, olfactives mais également une couleur particulière.

D'autres propriétés sont illustrées dans le tableau 4

Tableau 4 *:* Propriétés physiques et chimiques de la pyrazine

Propriétés physiques		**Propriétés chimiques**	
T° fusion	55 °C	Formule brute	$C_4H_4N_2$ [Isomères]
T° ébullition	115 °C	Masse molaire	80,088 ± 0,004 g·mol^{-1} C 59,99 %, H 5,03 %, N 34,98 %,

2.5. Propriétés biologiques et aromatiques

L'intérêt de ces molécules vient de leur pouvoir odorant. Ce caractère sera lié d'une part à leur fort potentiel de volatilité et d'autre part, aux électrons de valence situés sur les atomes d'azote qui interagissent de façon ionique avec les récepteurs olfactifs lorsque la molécule est orientée de façon appropriée. Cette interaction est renforcée par les groupements alkyls reliés aux atomes de carbones situés sur l'hétérocycle des alkylpyrazines [45].

Les pyrazines d'alkyle et d'acyle sont trouvées dans beaucoup de nourritures cuites, y compris les pommes de terre frites, les viandes cuites, les noix et le café grillés [46].

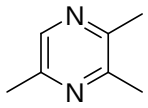
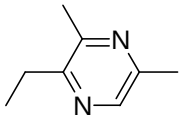
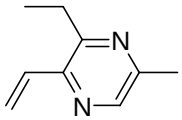
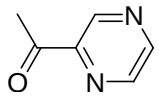
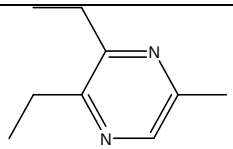
Des pyrazines ont été identifiées chez les animaux, les plantes et dans des cultures bactériennes. Chez les animaux et les plantes, les pyrazines sont considérées comme des signaux d'alerte, ayant un effet dissuasif ou attractant en fonction des circonstances sans avoir un effet nocif ou bénéfique [38].

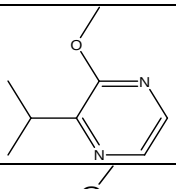
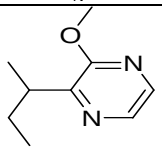
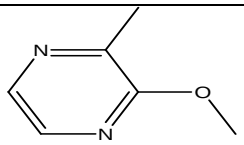
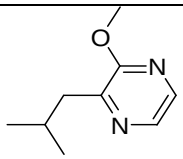
Pour l'homme, la nourriture constitue la principale source d'absorption de pyrazines.

Des pyrazines ont été détectées dans des aliments chauffés tels que les produits de bœuf, l'orge grillée, le cacao, le café, les cacahuètes, les maïs soufflés, les chips de pomme de terre, et l'avoine grillée ; ainsi que dans les aliments frais comme les tomates, les pois, les poivrons verts, l'asperge, chou-rave, et les produits laitiers [47].

La concentration des pyrazines dans les aliments varie entre 0,001 et 40 ppm [47]. En général, on peut dire que les pyrazines sont largement distribuées dans la nature et exceptionnellement utilisées par les bactéries, les plantes, les animaux vertébrés et invertébrés et par l'homme.

Tableau 5 : Propriétés de quelques dérivés de pyrazine [48]

	Structure	Substituant	Qualité d'arome	Valeur du seuil d'odeur (µg /l, dans l'eau)
1		Triméthyl	Terreux	90
2		2-éthyl-3,5-diméthyl	Pomme de terre	0,04
3		2-éthenyl-3-éthyl-5-méthyl	Café grillé	0,1
4		Acétyl	Pop-corn	62
5		2,3-Diéthyl-5-méthyl	Fromage parmesan	0,5

6		2-isopropyl-3-méthoxy	Pommes de terre	0,002
7		2sec-butyl-3-méthoxy	Pois vert, poivron	0,001
8		2- méthyl-3-méthoxy	Cacahuète grillée	0,004
10		2-isobutyl-3-méthoxy	Paprika chaud (Poivron rouge)	0,002

2.6. Toxicité des pyrazines

Certaines pyrazines présentent des effets pharmacologiques, comme diurétique [49], hepatoprotective [50] **,** antithrombogénique [51]**,** et des effets tuberculostatiques [52] **;** alors que d'autres affectent la fluidité des membranes des muscles lisses vasculaires. Les dérivés synthétiques de pyrazine sont également utiles comme drogues (antiviral, anticancéreux, antimycobactérien, etc.). Néanmoins, la toxicité aiguë des pyrazines qui sont utilisées comme additifs aromatisants a été jugée très faible.

Les études chroniques de toxicité subaiguë à court et long terme, affirment qu'aucun effet indésirable n'a été constaté, et pour le test de cancérogénicité, mutagénicité, et de génotoxicité in vitro et in vivo., l'influence des pyrazines a également été négative. Sur la base de toutes ces découvertes, les pyrazines ont reçu le statut GRAS par l'Association des fabricants d'arômes et d'extraits aromatisants dans les aliments.

Conclusion

Dans ce chapitre nous avons présenté deux familles de composés chimiques incontournables dans notre vie quotidienne : Les phénols et les pyrazines. La prévision de leurs profils moléculaires est indispensable pour l'évaluation des risques environnementaux et humains. Ainsi que la caractérisation expérimentale de leurs

propriétés utilisant des méthodes de prédiction in silico pour obtenir les données nécessaires à l'enregistrement de ces substances.

Références

[1] Rappoport Z. "The Chemistry of Phenols", John Wiley & Sons Ltd, Chichester, England 2003.

[2] Ehtash M. Purification des eaux polluées par du phénol dans per-tracteur à disques tournants, Thèse de doctorat, Institut national des sciences appliquées de Rouen, France 2011.

[3] Charriere B. Les composés phénoliques marqueurs de la matière organique terrestre dans deux écosystème marins : le delta du Rhône et le pro delta de la Têt, Thèse de doctorat, Institut national polytechnique de Toulouse, Université de Perpignan, France 1991.

[4] Gimeno O, Carbajo M, Beltrán F. J, Javier Rivas F. Phenol and substituted phenols AOP remediation, Journal of Hazardous Materials. 2005. B 119. 99.

[5] Pacheco Palencia L.A. Chemical characterization, bioactive properties and pigments stability of polyphenolics in ACAI (Euterpe Oleracea Mart)", PhD Thesis, Texas A&M University, U.S.A. 2009.

[6] Layton L. Reversing itself, FDA expresses concerns over health risks from BPA",Washington Post, U.S.A. 2010.

[7] Guivarch E.Z. Traitement des polluants organiques en milieux aqueux par procédé électrochimique d'oxydation avancée « Electro-fenton », Application à la minéralisation des colorants synthétiques, Thèse de Doctorat, Université.de Manne-lavalleé. 2004.

[8] INERIS, " Phénol Fiche de données toxicologiques et environnementales des substances chimiques. Institut National de l'EnviRonnement Industriel et des riSques INERIS–DRC-01-25590-01DR021. N°2-1. 2005. 1.

[9] Technique d'ingénieur, « phénol », 1997, J6020-5000 (*https://www.techniques-ingenieur.fr*)

[10] Arnaud P. Cours de chimie organique, 16éme édition, Dunod. 1997. 550

[11] Phenol, http://en.wikipedia.org/wiki/Phenol mars 2015 .

[12] Pichard A. phénol , version N°=2, INERIS 2005. *https://substances.ineris.fr › substance › getDocument*

[13] Ouahes C. chimie organique, office des publications universitaires : 10 92, codification :1.03.2687. ISBN: 978.9961.0.0165.3. 2014.

[14] Moussaoui S. Valorisation des palmes sèches du palmier dattier dans le traitement des eaux contaminées par phénol, mémoire de master, université Kasdi Merbah Ouargla. 2012. 30.

[15] SFC. Données industrielles, économiques, géographiques sur les principaux produits chimiques, métaux et matériaux. Société Française de Chimie, 8ème Edition.2009. http://www.societechimiquedefrance.fr/extras/Donnees/acc.htm

[16] ATSDR. Notice of the revised priority list of hazardous substances that will be the subject of toxicological profiles. Agency for Toxic Substances and Disease Registry. 2007.

[17] SPF Emploi, travail et concentration social, " phénol",CRC/CL/0011-F Version1 (2004) .

[18] Environnement Canada, santé Canada, loi canadienne sur la protection de l'environnement-phénol, ISBN 0-662-84220-0 NO de cat. EN 40-215/45F

[19] SPF Emploi, travail et concentration social, « phénol »,CRC/CL/0011-F Version 1(2004) .

[20] INRS, Base de données FICHES TOXICOLOGIQUES " fiche toxicologique N°15-phénol ", édition 1997. 9

[21] INERIS, " Phénol Fiche de données toxicologiques et environnementales des substances chimiques. Institut National de l'EnviRonnement Industriel et des riSques INERIS–DRC-01-25590-01DR021. N°2-1 (2005) 1-47. ref 8

[22] Nuhoglu A., Yalcin B. Modeling of phenol removal in a batch reactor. Process Biochemistry . 2005. 40. 1233

[23] " Health and Safety Guide". Phenol health and safety guide. IPCS-INCHEM. ISBN 924 151088 9 -ISSN 0259-7268. No 88.1994.

[24] Gianfreda L, Sannino F, Rao M.A, Bollag J-M. Oxidative transformation of phenols in aqueous mixtures. Water Research. 2003. 37. 3205

[25] " TOXICOLOGICAL PROFILE FOR PHENOL". ATSDR. Atlanta, Georgia 30333.2008. 269.

[26] Sheldon R. A. Heterogeneous Catalytic Oxidation and Fine Chemicals ,Study. Surf. Science Catalysts. 1991. 66. 33.

[27] Balaska A. Thèse de doctorat « Etude de la dégradation du phénol en milieu aqueux en présence des HPA de type DAWSON » 2015.UBMA- Annaba.

[28] Laurent, A. Neue Benzoylverbindungen. Ueber die Destillations producte des Benzensulfürs und Benzenazotürs [sulfure et azoture de benzène]. Justus Liebigs Ann. Chem., 1844. 52. 348.

[29] S. Wleiigel, Ber., 1882, 15, 1050

[30] A. T. Mason, Ber., 1887, 20, 261. 6.

[31] L. Wolff, Ber., 1887, 20,425.

[32] L. Wolff, Ber., 1888,21,1481

[33] L. Wolff, Ber., 1893,26,721

[34] www.pyrazinespecialties.com/products.htm

[35] Adams, T. B, Doull, J, Feron, V. J, Goodman, J. I, Marnett, L. J, Munro, I. C, Newberne, P.M, Portoghese, P. S, Smith, R. L, Waddell, W. J , Wagner, B. M. The FEMA GRAS assessment of pyrazine derivatives used as flavor ingredients. Food Chem. Toxicol., Vol. 40, No 4. ISSN 0278-6915. 2002. pp.429-451.

[36] Beck, H. C, Hansen, A. M, Lauritsen, F. R. Novel pyrazine metabolites found in polymyxin biosynthesis by Paenibacillus polymyxa. FEMS Microbiol. Lett., ISSN 0378-1097.2003. 220.pp.67-73.

[37] Wagner, R, Czerny, M, Bielohradsky, J, Grosch, W. Structure-odour-activity relationships of alkylpyrazines. Z. Lebensm. Unters. Forsch. A, , ISSN 1431-4630. 1999. Vol. 208, No. 5-6, pp.308-316.

[38] Woolfson, A. & Rothschild, M. Speculating about pyrazines. Proc. R. Soc. Lond. B, ISSN 0962-8452. 1990. Vol. 242, pp.113-119

[39] Wong J. W, Shibamoto T. Genotoxicity of Maillard reaction products. In The Maillard reaction: consequences for the chemical and life sciences. R.IKAN, Ed., John WILEY & Sons Ltd, 1996. pp.129-159.

[40] Maga J. A., Pyrazines in Foods: an Update: Critical Reviews in Food Technology, C R C Critical Reviews in Food Science and Nutrition. 1982.16. 1.

[41] Müller R, Rappert S. Pyrazines: occurrence, formation and biodegradation. Appl Microbiol Biotechnol. 2010.85, 1315.

[42] Stich H. F, Stich W, Rosin M. P. Powrie W. D. Food Cosmetics Toxicol. 1980,18, 581.

[43] Arnaud P. Cours de chimie organique, 18e édition, Dunod. 2009.

[44] www.intechopen.com

[45] Shimazaki, K., Inoue, T., Shikata, H., Sakakibara, K., 2005. Evaluation of the odor activity of pyrazine derivatives using structural and electronic parameters derived from conformational study by molecular mechanics (MM3) and ab initio calculations. Journal of molecular structure 749, 169–176

[46] Belitz H. D, roxh W. G. Schieberle P. Food chemistry. 3rd revised edition, 2004 page 374, SE/12691/01.

[47] Maga JA, Sizer CF. Pyrazines in foods. A review. J Agric Food Chem. 1973. 21. 22.

[48] H. D. Belitz, W. G. roxh & P. Schieberle, 2004, Food chemistry, 3rd revised edition, page 374, SE/12691/01

[49] Jones J.H, Bicking J.B, Cragoe E.J. Pyrazine diuretics. IV. N-Amidino-3-amino-6-substituted pyrazinecarboxamides. J Med Chem 196710. 899

[50] Kim N.D, Kwak MK, Kim SG (1997) Inhibition of cytochrome P450 2EI expression by 2-(allylthio)pyrazine, a potential chemoprotective agent: hepatoprotective effects. Biochem Pharmacol 51:261–269

[51] Lian X, Wang S, Xu G, Lin N, Li Q, Zhu H (2008) The application with tetramethyl pyrazine for antithrombogenicity improvement on silk fibroin surface. Appl Surf Sci 255:480–482

[52] Milczarska B, Foks H, Sokołowska J, Janowiec M, Zwolska Z,Andrzejczyk Z (1999) Studies on pyrazine derivatives. XXXIII. Synthesis and tuberculostatic activity of 1-[1-(2pyrazinyl)- ethyl]-4-N-substituted thiosemicarbazide derivatives. Acta PolPharm 56:121–126 zide derivatives. Acta PolPharm 56:121–126

CHAPİTRE 2
LES PROPRİETÉS ÉTUDİÉES

- Température d’ébullition
- Solubilité aqueuse
- Indice de rétention

1. La température d'ébullition (Teb)

La température d'ébullition est la température à laquelle les phases liquide et gazeuse d'une substance pure sont en équilibre à une pression donnée, c'est la température à laquelle la substance change d'état, du liquide au gaz à une pression donnée. Le point d'ébullition normal est le point d'ébullition à la pression atmosphérique normale (1,013.105 kPa). En termes d'interactions intermoléculaires, le point d'ébullition représente la température à laquelle les molécules possèdent l'énergie thermique suffisante pour surmonter les attractions intermoléculaires liant les molécules dans le liquide. La température d'ébullition d'un composé pur augmente avec la taille, la ramification de la molécule, et avec la présence des liaisons hydrogènes et des interactions dipôle-dipôle. La température d'ébullition est importante pour la caractérisation et l'identification du composé. Elle fournit également une indication de la volatilité d'un composé. D'autres propriétés physiques, telles que la température critique [1], le point d'éclair [2], et l'enthalpie de vaporisation [3], peuvent être prédits ou estimés à partir des points d'ébullition. Le besoin de données fiables pour l'optimisation des processus industriels, le développement des modèles QSPR fiables pour l'estimation des points d'ébullition normaux pour les composés qui ne sont pas encore synthétisés est devenu important. De nombreuses méthodes ont été développées pour l'estimation des points d'ébullition normaux des composés, et de nombreuses corrélations QSPR ont été rapportées. Des tentatives préliminaires ont été faites pour corréler les points d'ébullition des hydrocarbures homologues avec le nombre d'atomes de carbone ou le poids moléculaire [4]. Des méthodes ultérieures ont employé des paramètres physiques tels que le parachor et la réfractivité molaire [5]. Des méthodes pour l'estimation des points d'ébullition ont été résumées par Rechsteiner [3] et Horvath [6]. Des efforts ont été faits pour estimer les points d'ébullition par contribution de groupe additive (CGA) [3,7] basée sur l'hypothèse que les forces de cohésion dans les liquides sont de courte portée [8] et procède de la division d'une molécule en groupes structuraux prédéfinis, dont chacun ajoute un incrément constant à la valeur de la propriété [9]. Les méthodes de contribution de groupe fournissent une bonne prédiction des points d'ébullition [10,11], avec une erreur absolue moyenne de 15,5 K, pour les petites molécules non polaires. Cependant, les méthodes CGA sont limitées aux molécules contenant des groupes présents dans l'ensemble des molécules d'étalonnage, et

certains schémas de contribution de groupe ne sont pas suffisamment complets pour couvrir plusieurs substitutions de groupes fonctionnels.

Mise à part les simples corrélations des points d'ébullition avec le nombre d'atomes de carbone ou le poids moléculaire pour des séries homologues de composés, Wiener a été le premier à corréler les points d'ébullition avec des descripteurs topologiques [12]. Wiener a introduit deux paramètres structurels, appelés l'indice de Wiener (W), défini comme la somme des distances entre deux atomes de carbone dans la molécule [12], et l'indice de polarité de Wiener (P), défini comme le nombre de paires non ordonnées de sommets dont la distance entre deux sommets est égale à 3. Sur la base de ces indices, il a prédit les points d'ébullition des paraffines avec une erreur moyenne de 1°C [12]. D'autres indices topologiques, y compris les indices de connectivité moléculaire de Randić [13], et de Kier & Hall [14], ont permis de corréler les points d'ébullition des alcanes et des amines. Pendant plus de quatre décennies, la corrélation des points d'ébullition des hydrocarbures avec la structure chimique a suscité un intérêt considérable. Cependant, pour une meilleure prévisibilité d'une propriété sous la forme d'un modèle général, la recherche de meilleurs descripteurs a été un point focal de la recherche QSPR.

À l'heure actuelle, un grand nombre de modèles QSPR ont été développés pour la corrélation et la prédiction des points d'ébullition de diverses classes de composés organiques tels que les hydrocarbures, les hydrocarbures halogénés, les alcools, les composés carbonylés, les amines, les nitriles, les pyrènes, les furannes, les thiophènes, les sulfures, les éthers et les peroxydes

2. La solubilité aqueuse (Sw)

La solubilité aqueuse (Sw) des composés organiques est l'un des facteurs clés à prendre en considération lors du classement des produits chimiques organiques significatifs pour l'environnement par rapport à leur mobilité dans le sol et leur volatilité à la surface de l'eau.

C'est aussi un paramètre particulier important dans les études sur l'absorption, la distribution, les métabolismes et l'excrétion xénobiotique chez les êtres humains. Cependant, la mesure expérimentale de la solubilité est difficile car elle peut être très longue pour atteindre l'équilibre de solubilité dans le cas des composés apolaires ou nécessiter une grande quantité de produits chimiques dans le cas de molécules hautement hydrophiles. En outre, les valeurs de la solubilité de la majorité des composés organiques restent inconnues [15].

La prédiction de la solubilité dans l'eau est importante dans les sciences de l'environnement. Les approches les plus utiles pour l'estimation de la solubilité dans l'eau et dans des solutions non aqueuses sont basées sur une contribution de groupes fonctionnels. Le coefficient UNIversel d'ACtivité Fonctionnelle (en anglais UNIFAC) est une approche fiable et rapide pour prédire les coefficients de solubilité / activité aqueuse des non-électrolytes dans le mélange liquide [16,17]. Ce modèle a été étendu à l'eau comme solvant. Le principe fondamental est qu'un composé comporte des groupes fonctionnels, et chaque groupe apporte une contribution unique à la solubilité aqueuse. Un grand nombre de composés contiennent simplement un nombre limité de groupes fonctionnels, et il est donc possible de prévoir la solubilité de nombreux composés en utilisant un nombre minimal de groupes fonctionnels organiques. Cependant, l'exactitude du modèle proposé par la Fédération internationale des Nations Unies reste controversée [18]. Une autre approche similaire axée sur la prédiction de la solubilité aqueuse est bien connue sous le nom de coefficients d'activité de groupe fonctionnel aqueux (AQUAFAC) « Aqueous functional group activity coefficients » [19]. L'idée fondamentale est d'exprimer les contributions enthalpiques et entropiques à l'énergie excédentaire en additionnant les parties interactives du soluté, les molécules organiques dissoutes et leurs groupes fonctionnels. Certaines méthodes prometteuses pour prédire la solubilité aqueuse sont basées sur des descripteurs moléculaires topologiques, géométriques et électroniques

[20]. Duchowicz et al. [21] ont développé un modèle QSPR linéaire, généralement applicable, basé sur 147 composés pharmaceutiques contenant trois descripteurs moléculaires. Kim *et al*. [22] ont corrélé la solubilité aqueuse de médicaments peu solubles, tels que l'acide ursodésoxycholique ($C_{24}H_{40}O_4$), la diphénylhydrantoïne ($C_{15}H_{12}N_2O_2$;)et le diméthylbiphényldicarboxylate (C_{16} $H_{14}O_4$).

Trois ensembles de données de 50 composés ont été extraits des données de la littérature en fonction de leur similitude structurelle avec chaque médicament. Des modèles QSPR rapides et prédictifs (R^2> 0,90) ont été développés et validés (R^2> 0,85). Huuskonen *et al*. [23] , ont extrait un ensemble de calibrage de 191 composés antidrogue de la base de données AQUASOL pour corréler la solubilité aqueuse par un modèle à cinq paramètres (C log P, poids moléculaire, variable indicatrice pour les groupes amines aliphatiques, nombre de liaisons rotatives et un nombre d'anneaux aromatiques) avec les statistiques: R^2 = 0,87 et s = 0,51. Le modèle a été appliqué à un ensemble d'essai de 174 composés antidrogue avec R^2 = 0,80 et s = 0,68. Les résultats de cette étude suggèrent que l'augmentation de la taille moléculaire, la rigidité et la lipophilie diminuent la solubilité alors que l'augmentation de la flexibilité conformationnelle et la présence d‘une amine non conjuguée augmente la solubilité des composés pharmaceutiques. Du-Cuny *et al*. [24] visant à modéliser la solubilité aqueuse de composés apparentés aux médicaments dans des séries congénères, la lipophilie (C log P) combinée à l'information sur les fragments structurels, les facteurs de correction basés sur des fragments et les indices de séries de congénères ont été utilisés comme descripteurs pour une ACP suivie d'une régression PLS multivariée. Le modèle général résultant (R^2 = 0,84 et rms = 0,51) était basé sur un ensemble de données internes de 1515 composés pharmaceutiques, et la solubilité de l'ensemble d'essai de 958 composés était prédite avec un haut degré de précision, R^2 = 0,81 et s = 0,42. Au cours du développement du modèle, des règles ont été dérivées qui peuvent être utilisées par les médecins chimistes ou les scientifiques intéressés en tant que directive approximative sur la contribution des fragments structurels à la solubilité

3. L'indice de rétention chromatographique (IR)

La chromatographie est la méthode de séparation la plus générale, la plus puissante, et la plus simple qui soit actuellement disponible. Elle représente un des procédés physico-chimiques de séparation, au même titre que la distillation, la cristallisation ou l'extraction fractionnée, des constituants d'un mélange homogène liquide ou gazeux. Ce procédé hydrodynamique a donné naissance à une méthode analytique instrumentale présentant un très grand domaine d'applicabilité et par suite se trouve très répandu [25].

Le principe de base repose sur les équilibres de concentration qui apparaissent lorsqu'un composé est mis en présence de deux phases non miscibles.

L'une des phases, dite stationnaire, est emprisonnée dans une colonne ou fixée sur un support et l'autre, dite mobile, se déplace au contact de la première. Si plusieurs composés sont présents, ils se trouvent entraînés à des vitesses différentes, provoquant leur séparation.

En chromatographie gazeuse la phase mobile est un gaz, et la phase stationnaire peut être un liquide fixé par imbibition d'un support inerte (CGL), ou un solide adsorbant (CGS).

3.1. Phases stationnaires :

Les phases stationnaires garnissent des tubes métalliques ou en verre de faible diamètres (colonne à garnissage) ou sont déposées sur les parois internes d'un tube de très faible diamètre (colonnes capillaires).

Les phases actuelles correspondent à deux principaux types de composés: les polysiloxanes et les polyéthylèneglycols « PEG », chaque catégorie pouvant faire l'objet de modifications structurales mineures [25].

3.1.1-La phase stationnaire OV-101:

Les phases les plus répandues sont les polymères siliconés. L'OV-101 appartient à cette famille dérivée du diméthyl polysiloxane qui se présente comme suit:

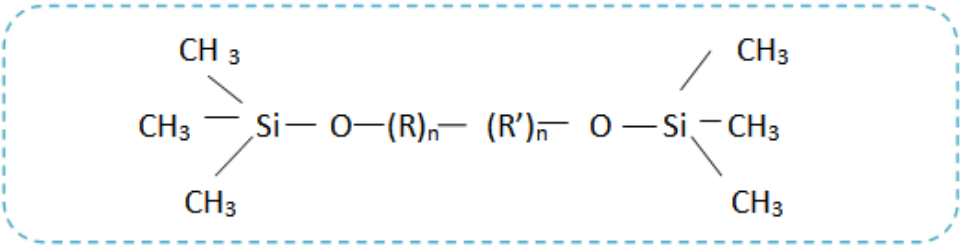


Figure 1 : Formule générale du diméthyl polysiloxane

Cette phase est greffée sur la colonne en silice. R et R' sont des groupements silylés pouvant posséder des groupements polaires ou polarisables qui déterminent les propriétés de ces phases. L'OV-101 diméthylsiloxane est très peu polaire, à une viscosité faible, présente une bonne inertie chimique, et peut être chauffé sans dommage jusqu'à 300°C [26].

3.1.2. La phase stationnaire CW-20M :

Le Carbowax « CW -20M» appartient à la famille des polyéthylèneglycols (PEG), qui sont greffés sur les parois en silice de la colonne. Leur formule est la suivante:

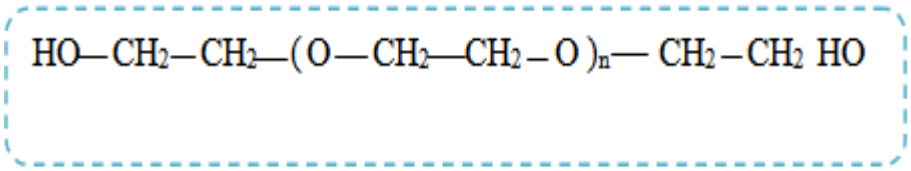


Figure 2*:* Formule générale des PEG

Ces phases sont utilisables entre 20° et 250°C, sont moins inertes que les phases siliconées et sont particuliérement très sensibles à l'oxygène.

Ces phases stationnaires possèdent de nombreux oxygènes sont classées parmi les phases stationnaires les plus polaires et sont utilisées pour séparer les molécules de forte polarité comme celles possédant des fonctions alcool, aldéhyde ou cétone [26]**.**

3.2. Indice de rétention (Ir)

La chromatographie en phase gazeuse peut permettre l'identification d'un produit dans un mélange complexe. A cette fin on utilise le temps de rétention réduit (**t'_R**), ou plus communément l'indice de rétention (Ir) qui dépend de la structure du composé [27]**.** L'introduction de ces paramètres a au moins trois objectifs:

- Caractériser tout composé par une grandeur plus générale que son temps de rétention dans des conditions définies. Il en résulte le système des indices de rétention qui est un moyen efficace et peu coûteux pour éviter certaines erreurs d'identification.
- Suivre l'évolution dans le temps des colonnes et par suite leurs performances.
- Classer entre elles les phases stationnaires connues pour faciliter le choix d'une colonne bien adaptée pour tout problème nouveau de séparation, sachant que la polarité ou la nature chimique d'une phase stationnaire ne permettent pas, seules, de prévoir sa réelle aptitude séparatrice [25]**.**

Le calcul peut se faire pour une expérimentation à température constante par interpolation logarithmique: indices de Kováts [28]**,** ou en programmation de température par interpolation linéaire, indices de rétention, ou indices de van den Dool et Kratz [29]**.** Bien que dans la grande majorité des cas, chaque molécule possède des indices de rétention sur colonne apolaire et polaire qui lui sont propres, deux molécules peuvent fortuitement être co-éluées et présenter des indices de rétention identiques [30].

3.3. Indice de rétention de Kováts :

L'indice de rétention de Kováts (1958) est une mesure de la rétention d'un composé par rapport à la rétention des alcanes normaux (hydrocarbures à chaine droite) à une température constante. Ils sont utilisés comme référence car ils sont non polaires, chimiquement inertes et solubles dans la plupart de phases stationnaires. L'indice de rétention de Kováts d'un composé (x) sur la colonne considérée se calcule comme suit :

$$I_x = 100 \times n + 100 \frac{\log t'_{R(x)} - \log t'_{R(n)}}{\log t'_{R(n+1)} - \log t'_{R(n)}} \qquad (1)$$

I_x : Indice de rétention d'un composé x.

$t'_{R(x)}$: Temps de rétention réduit du composé x.

$t'_{R(n)}$: Temps de rétention réduit de l'alcane élué avant x.

$t'_{R(n+1)}$: Temps de rétention réduit de l'alcane élué après x.

Les deux temps de rétention se rapportent à deux alcanes successifs (n et n +1atomes de carbone), ou à deux composés de même type.

-Le temps de rétention réduit du composé t'_R est la différence entre son temps de rétention et le temps de rétention nulle t_0, comme le montre l'équation :

$$t'_R = t_R - t_0 \qquad (2)$$

Le temps de rétention nulle ou le temps de rétention d'un composé non retenu t_0 : correspond au temps que met un constituant pour traverser l'ensemble du système chromatographique sans interaction avec la phase stationnaire.

Un des désavantages de l'indice de Kováts est qu'il n'est pas utilisable en chromatographie gazeuse à température programmée (CGTP).Pour combler cette lacune van den Dool et Kratz proposèrent de calculer une grandeur(I_p), semblable à l'indice de Kováts , en remplaçant dans l'expression de ce dernier le logarithme du temps de rétention réduit directement par le temps (ou la température) de rétention [31].

3.4. Indice de rétention de van den Dool et Kratz :

$$\frac{I_p}{100} = n + \frac{T_{R(x)} - T_{R(n)}}{T_{R(n+1)} - T_{R(n)}} \qquad (3)$$

Avec TR(n) < TR(x) < TR (n+1) ;

TR(x) est la température de rétention du soluté x, TR(n) et TR (n+1) celles des n-alcanes de référence à n et (n+1) atomes de carbone l'encadrant sur le chromatogramme [32].

L'indice de rétention est une donnée facilement accessible avec précision, qui est indépendante des caractéristiques de l'appareil (paramètres de colonne) et n'est fonction que du soluté, de la température et de la phase stationnaire. L'utilisation simultanée des indices de rétention sur 2 colonnes de polarités différentes conduit à la notion d'incrément d'indice.

$$\Delta Ir_i = Ir_i^p - Ir_i^a \qquad (4)$$

D'après Kováts le ΔIr d'un composé résulte des différences d'interactions entre les molécules des 2 phases stationnaires et les groupements fonctionnels que possède le soluté i.

Ir_i^p : Indice de rétention d'un composé i sur la phase polaire ;

Ir_i^a : Indice de rétention d'un composé i sur la phase apolaire.

Références

1. Fisher C H. Boiling-point gives critical-temperature. Chem. Eng. 1989. 96, 157.
2. Satyanarayana K, Kakati M C. Note: Correlation of flash points. Fire. Mater. 1991. 15, 97.
3. Rechsteiner C E. In Handbook of Chemical Property Estimation Methods; Lyman, W. J., Reehl, W. F., Rosenblatt, D. H., Eds.; McGraw-Hill: New York, 1982; Chapter 12.
4. Walker J. The boiling points of homologous compounds. Part I. Simple and mixed ethers .J. Chem. Soc. 1894. 65, 193.
5. Meissner H P. Critical constants from parachor and molar refraction. Chem. Eng. Progr. 1949. 45, 149.
6. Horvath A L. Molecular Design: Chemical Structure Generation from the Properties of Pure Organic Compounds; Elseiver: Amster-dam, 1992; Chapter 2.
7. Reid R C, Prausnitz, J. M.; Poling, B. E. The Properties of Gases and Liquids, 4th ed.; McGraw-Hill: New York, 1987.
8. Benson S W, Buss J H. Additivity rules for the estimation of molecular properties. Thermodynamic properties. J. Chem. Phys. 1958. 29, 546.
9. Copeman T W, Mathias P M, Klotz H C. In Physical Property Prediction in Organic Chemistry; Jochum, C., Hicks, M. G., Sunkel, J., Eds.; Springer-Verlag: New York, 1988; pp 351.
10. Joback K G, Reid R C. Estimation of pure-component properties from group-contributions. Chem. Eng. Commun. 1987. 57, 233.
11. Stein S E, Brown R L. Estimation of normal boiling points from group contributions. J. Chem. Inf. Comput. Sci. 1994. 34, 581.
12. Wiener H. Influence of interatomic forces on paraffin properties, J. Am. Chem. Soc. 1947. 69, 17
13. Randić M. Characterization of molecular branching. J. Am. Chem. Soc. 1975, 97, 6609.
14. Kier L B, Hall L H. Molecular Connectivity in Chemistry and Drug Research; Academic: New York, 1976.
15. Muir D C G, Howard P H, Are there other persistent organic pollutants? A challenge for environmental chemists. Envi. Sci & Tech. 2006. 40 , 7157.

16. Fredenslund A, Jones R L, Prausnitz J M, Group-Contribution Estimation of Activity Coefficients in Nonideal Liquid Mixtures. Aiche J., 1975. 21, 1086
17. Hansen H K , Rasmussen P, Fredenslund A, Schiller M, Gmehling J, Vapor-Liquid Equilibria by UNIFAC Group Contribution. 5. Revision and Extension. Ind. Eng. Chem. Res.,1991. 30, 2352.
18. Kan A T, Tomson M B, UNIFAC Prediction of Aqueous and Nonaqueous Solubilities of Chemicals with Environmental Interest. Environ. Sci. Technol. 1996. 30, 1369.
19. Myrdal P, Ward G H, Simamora P, Yalkowsky S H , AQUAFAC: Aqueous Functional Group Activity Coefficients. SAR QSAR Environ. Res. 1993. 1, 53.
20. Mitchell B E, Jurs P C, Prediction of Infinite Dilution Activity Coefficients of Organic Compounds in Aqueous Solution from Molecular Structure. J. Chem. Inf. Comput. Sci. 1998. 38, 200.
21. Duchowicz P R, Talevi A., Bruno-Blanch L E, Castro E A., New QSPR study for the prediction of aqueous solubility of drug-like compounds. Bioorg. Med. Chem. 2008. 16, 7944.
22. Kim J, D Jung H, Rhee H, Choi S H, Sung M J., Choi W S, Aqueous solubility of poorly water-soluble drugs: Prediction using similarity and quantitative structure-property relationship models. Korean J. Chem. Eng. 2008, 25, 865.
23. Huuskonen J, D. Livingstone J, Manallack D T, Prediction of drug solubility from molecular structure using a drug-like training set. SAR QSAR Environ. Res. 2008, 19, 191.
24. Du-Cuny L, Huwyler J, Wiese M, Eur. Kansy M, Computational aqueous solubility prediction for drug-like compounds in congeneric series. J. Med. Chem. 2008, 43, 501.
25. Rouessac F. Rouessac A .Analyse chimique : Méthodes et techniques instrumentales modernes .6e édition, Paris, Dunod, 2004.
26. Mahuzier.G Hamon.M Ferrier.D Prognon.P « Chimie analytique : Méthode de séparation »Tome 2,3 édition, paris, Masson, 1999
27. Messadi-Lourici.L, 2004. Thèse de doctorat, Université Badji Mokhtar –Annaba. Contribution à l'étude des indices de rétention en chromatographie gazeuse. p169

28. Kováts E. 1958. Chromatographische charakterisierung organischer verbidugen. Teil 1: Retention indices aliphatischer halogenide, alkohole, aldehyde und ketones. Helvetica Chimica Acta, vol 41(7): 1915–1932.
29. Van den Dool H. Kratz P. 1963. A generalization of the retention index system including linear temperature programmed gas–liquid partition chromatography.Journal of Chromatography, vol 11: 463-471
30. Sutour S. 2010.Thèse de doctorat. Université de corse pascal Paoli. Étude de la composition chimique d'huiles essentielles et d'extraits de menthes de corse et de kumquats. p 22
31. Messadi D, Souici L et Lourici L.2013 « Interpolation par fonction B-splines pour le calcul des indices de rétention en programmation de température : application à un mélange de phénols tests séparés des colonnes garnies de Tenax-GC modifié » Rev.Sci.Technol. , Synthése 27 :89-98
32. Lourici L et Messadi D. 2009 méthodes non linéaires pour le calcul des indices de rétention en chromatographie gazeuse à programmation linéaire de température. Lebanese Science Journal, Vol. 10(10) :77-86

CHAPİTRE 3
MODÉLİSATION MOLÉCULAİRE

Introduction

- **La mécanique quantique**
- **La mécanique moléculaire**
- **La dynamique moléculaire**

Conclusion

Introduction

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 soit pour analyser les molécules et les systèmes moléculaires ou prédire les propriétés moléculaires, chimiques et biochimiques [1]. Elle sert de pont entre la théorie et l'expérience pour:

1. Extraire des résultats pour un modèle particulier.

2. Comparer les résultats expérimentaux du système.

3. Comparer les prédictions théoriques du modèle.

4. Aider à comprendre et interpréter les observations expérimentales.

5. Etablir une corrélation entre détails microscopiques au niveau atomique et moléculaire et les propriétés macroscopiques.

6. Fournir de l'information inaccessible à partir d'expériences réelles.

Grace au développement informatique de ces dernières années et à l'essor du calcul parallèle intensif en particulier, la modélisation moléculaire est devenue un véritable enjeu. En effet les systèmes moléculaires qui sont amenés à être étudiés tendent à devenir de plus en plus complexes. Cette complexité est bien sûr liée à la taille des molécules envisagées (plusieurs centaines de milliers d'atomes pour les molécules biologiques par exemple) ainsi qu'à la structure intrinsèque des atomes eux-mêmes, mais aussi au degré de précision exigé pour le calcul de certaines grandeurs physiques [2].

La modélisation moléculaire est une application des méthodes théoriques et des méthodes de calcul pour résoudre des problèmes impliquant la structure moléculaire et la réactivité chimique [3]. Ces méthodes peuvent être relativement simples et utilisables rapidement ou au contraire elles peuvent être extrêmement complexes et demander des centaines d'heures de temps d'ordinateur, même sur un superordinateur. En plus, ces méthodes utilisent souvent des moyens infographiques très sophistiqués qui facilitent grandement la transformation de quantités impressionnantes de nombres en quelques représentations graphiques facilement interprétables [4].

Les méthodes de la modélisation moléculaire peuvent être rangées en trois catégories:

- Les méthodes quantiques.
- La mécanique moléculaire.
- La dynamique moléculaire.

1. La mécanique quantique

Les méthodes de chimie quantique permettent le calcul de la structure électronique de systèmes tels que les atomes, les molécules neutres, les espèces radicalaires, les ions, les clusters d'atomes, les surfaces de solides, etc. Des algorithmes de calculs très précis sont utilisés pour minimiser l'énergie totale en fonction des paramètres structuraux et pour prédire la structure la plus stable des composés étudiés.

Les méthodes de modélisation basées sur la mécanique quantique [5] visent à décrire le système étudié par une fonction d'onde qui peut théoriquement être déterminée par résolution de l'équation de Schrödinger [6]. Cette équation relie les états stationnaires d'un système moléculaire et les énergies qui leur sont associées à un opérateur hamiltonien et à leur fonction d'onde.

L'équation de Schrödinger non relativiste, indépendante du temps, se présente sous la forme suivante :

$$\hat{H}\,\Psi = E\,\Psi \qquad (1)$$

Où Ĥ est l'opérateur hamiltonien et E l'énergie du système.

On peut résoudre l'équation de Schrödinger analytiquement seulement pour des problèmes très simples, tels que la particule unique dans une boîte, l'oscillateur harmonique ou l'atome d'hydrogène isolé. Afin d'effectuer des résolutions numériques complexes et obtenir des résultats dans un temps raisonnable, il est nécessaire d'introduire plusieurs approximations. Trois principales approximations sont adoptées et employées : l'approximation de Born-Oppenheimer, l'approximation d'orbitales moléculaires « Combinaisons linéaire d'orbitales atomiques » : MO-LCAO pour (Molecular Orbitals-Linear Combination of Atomic Orbitals).

Nous allons décrire brièvement le principe des deux approximations. La troisième sera présentée ensuite avec la méthode Hartree-Fock.

L'approximation de Born-Oppenheimer [7] est la première et la plus fondamentale utilisée dans toutes les méthodes de MQ [8]. Selon cette approximation, le mouvement des électrons est séparé de celui des noyaux en prenant en compte le fait que les électrons sont beaucoup plus légers, et donc peuvent réagir à chaque changement de positions des noyaux presque instantanément. On considère les noyaux comme fixes, donc la fonction d'onde électronique dépend seulement de leurs positions (et non de leurs mouvements). D'autres approximations généralement utilisées concernent la forme de la fonction d'onde, comme l'approximation MO-LCAO [9] qui représente l'orbitale moléculaire comme une combinaison linéaire d'orbitales atomiques. En fonction des autres approximations utilisées, les méthodes de MQ sont divisées en quatre groupes principaux:

1.1 Méthodes empiriques

Les méthodes empiriques, comme, par exemple, les Orbitales Moléculaires de Hückel (HMO) pour « Hückel Molecular Orbital » ou la méthode de Hückel étendue (EHT) pour « Extended Hückel Theory », réduisent considérablement les temps de calcul en ne considérant que les parties "nécessaires" ou "intéressantes" [10]. Ces méthodes utilisent des approximations très grossières qui produisent de grandes erreurs de calcul, et donc ne sont employées que très rarement.

1.2 Méthodes semi-empiriques

Les calculs semi-empiriques sont eux développés sur la même structure générale que les calculs Hartree-Fock, mais certaines parties de l'information sont sujettes à approximation ou même complètement omises, afin de les rendre moins exigeants en termes de temps de calcul.

Les termes énergétiques les plus difficiles à calculer sont estimés à partir des données expérimentales. Les temps de calculs sont considérablement raccourcis, mais la méthode est tributaire des composés qui ont servi à l'étalonner. Selon la nature des approximations utilisées [3], on distingue plusieurs variantes :

- CNDO : (*Complete Neglect of Differential Overlep*) ou « l'approximation au recouvrement différentiel nul », la première méthode semi empirique, elle a été proposée par Pople, Segal et Santry en 1965. Méthode présentant certains défauts entre autres : elle ne tient pas compte de la règle de Hund.

- INDO: (*Intermediate Neglect of Differential Overlap*) Proposée par Pople, Beveridge et Dobosh en 1967. Elle permet de distinguer entre les états singulets et les états triplets d'un système en conservant les intégrales d'échange.
- MINDO/3 : Proposée par Bingham, Dewar et Lo en 1975. Paramétisation effectuée en se référant aux résultats expérimentaux et non pas aux résultats ab-initio, de plus l'algorithme d'optimisation utilisé est très efficace (Davidon-Fletcher-Powel). Cependant, elle surestime la chaleur de la formation des systèmes insaturés et sous-estime celle des molécules contenant des atomes voisins ayant des paires libres.
- MNDO : (*Modified Neglect of Diatomic Overlap*) Proposée par Dewar et Thiel en 1977. Méthode basée sur l'approximation NDDO (*Neglect of Diatomic Differential Overlap*) qui consiste à négliger le recouvrement différentiel entre orbitales atomiques sur des atomes différents. Cette méthode ne traite pas les métaux de transition et présente des difficultés pour les systèmes conjugués.
- **AM1** : (*Austin Model 1*) Proposée par Dewar en 1985. Il a tenté de corriger les défauts de MNDO.
- **PM 3** : (Parametric Method 3) Proposée par Stewart en 1989. Présente beaucoup de points en commun avec AM1. D'ailleurs il existe toujours un débat concernant les mérites relatifs de paramétrisation de chacune d'elles.
- **SAM1** : (*Semi-ab-intio Model 1*) La méthode la plus récente proposée par Dewar en 1993. Elle inclut la corrélation électronique.

Toutes ces méthodes, CNDO, INDO, NNDO, MNDO ou les plus utilisées Austin Model 1 (AM1) [11] et Parametrization Method 3 (PM3) [12a-b] négligent généralement le calcul de certaines intégrales et les remplacent par des paramètres expérimentaux. Ces paramètres peuvent être obtenus, par exemple, à partir des caractéristiques spectrales des atomes, de calcul ab initio de haut niveau ou d'autres méthodes expérimentales [13]. Une approximation supplémentaire des méthodes semi-empiriques est de considérer uniquement les électrons de valence dans le calcul, le reste des électrons étant inclus dans le "coeur" (avec le noyau), qui n'est pas pris en compte dans le calcul. De cette manière, tout en appartenant toujours aux méthodes MQ, ces approches réduisent considérablement la puissance de calcul requis par rapport aux méthodes ab initio classiques et elles peuvent être utilisées dans l'étude des réactions chimiques.

1.3 Méthodes ab initio

Les méthodes ab initio résolvent l'équation de Schrödinger en utilisant un nombre minimal d'approximations, telles que celles de Born-Oppenheimer et de MO-LCAO. Elles ne comportent aucun paramètre expérimental ou empirique. Le problème principal des méthodes ab initio est la résolution des interactions électrostatiques entre les électrons. La méthode Hartree-Fock (HF) [14] était la première théorie introduite pour donner une réponse à ce problème.

1.4 Au-delà de Hatree-Fock

L'ensemble des électrons est défini de manière à ce que chaque électron se déplace dans le champ électrostatique moyen des autres électrons.

Le système entier est donc décrit par une série d'équations HF, qui sont résolues par un processus itératif en utilisant la méthode du champ autocohérent (SCF) pour self-consistent field. Dans la procédure SCF, les fonctions « Ψ » de toutes les orbitales moléculaires (OM) sont initialement estimées et sont utilisées pour construire les opérateurs hamiltoniens de chaque électron.

Ces opérateurs hamiltoniens sont nécessaires pour générer une nouvelle série de «Ψ», qui sera plus précise. Ce cycle est répété jusqu'à ce que la convergence soit atteinte. La qualité des résultats de méthodes HF dépend de la qualité de l'expansion de la fonction d'onde dans l'ensemble des bases.

L'énergie calculée par la méthode HF convergera vers la meilleure énergie accessible (limite HF) avec l'ensemble complet des bases. Mais toutefois, la méthode de Hatree-Fock souffre d'un inconvénient majeur : dès lors que la répulsion électronique est moyennée, une partie de la corrélation électronique est négligée. Afin de pallier ces désavantages, des méthodes, dites post Hartree-Fock, ont été développées.

- **Post- Hatree-Fock**

Malheureusement, la limite HF n'est pas encore l'énergie réelle à cause de quelques approximations supplémentaires dans la théorie HF. Cette théorie, par exemple, ne prend pas en compte les possibles effets relativistes, par exemple que la masse des électrons n'est pas forcément constante et qu'elle peut changer en fonction de leur vitesse. Une autre approximation, selon laquelle l'électron est en mouvement dans le potentiel moyen des autres électrons, et donc que sa position n'est pas affectée par la position des électrons voisins, ne permet pas de représenter les effets de corrélation

électronique. Cette énergie de corrélation est définie comme la différence entre l'énergie exacte et l'énergie limite de Hartree Fock.

De nombreuses méthodes qui traitent la corrélation électronique, appelées aussi méthodes post- Hatree-Fock, ont été développées pour inclure l'énergie de corrélation dans le calcul [15]. Les méthodes les plus populaires et les plus fréquemment utilisées sont, par exemple,

- La méthode de perturbation Møller-Plesset [16] dans les versions MP2, MP4 et MP6.
- Les méthodes d'interaction de configuration « configuration-interaction (CI) » [8, 17,18] utilisées dans les niveaux d'excitation Simple « CIS », Double « CID », Simple et Double « CISD » ou Quadratique « QCISD ».
- Les méthodes des groupes couplés « coupled clusters (CC) » [19] utilisées le plus fréquemment dans les variantes CCSD ou CCSD(T).

2. Mécanique Moléculaire (MM)

Malgré le grand succès qu'ont reçu les méthodes de la mécanique quantique, leurs utilisations restent restreintes. En effet, les méthodes de MQ sont classées parmi les méthodes les plus rigoureuses mais elles sont très coûteuses en temps de calculs et parfois même il sera impossible de faire un calcul en utilisant les méthodes MQ pour des systèmes de grosse taille moléculaire même en utilisant les méthodes abrégées (semi empiriques). Si l'on désire donc modéliser une grosse molécule de taille supérieure à celles accessibles par la méthode semi-empirique alors il est possible d'éviter la mécanique quantique en utilisant la méthode de mécanique moléculaire (MM) [13].

La Mécanique Moléculaire permet le calcul de l'énergie d'interaction d'un système en fonction des seules positions des noyaux, en ignorant ainsi le mouvement des électrons. L'approximation de Born-Oppenheimer est aussi utilisée, cependant, le mouvement des noyaux n'est plus décrit par un Hamiltonien quantique comme en Mécanique Quantique, mais par les descriptions de la mécanique classique ou les atomes sont assimilés à des masses ponctuelles (éventuellement chargés) et les liaisons chimiques à des ressorts mécaniques.

L'expression "Mécanique moléculaire" désigne actuellement une méthode de calcul largement utilisée qui permet, a priori, d'obtenir des résultats de géométries et d'énergies moléculaires en se basant sur la mécanique classique.

La mécanique moléculaire (MM), appelée parfois "calcul par champ de force empirique" [20], qui est un outil informatique mis à la disposition du chimiste pour étudier la structure 3D des molécules et les propriétés physico -chimiques associées. C'est une méthode non quantique qui résulte de l'ajustement de données expérimentales sur des fonctions mathématiques simples [21]. En particulier, la mécanique moléculaire permet l'étude d'une gamme étendue de propriétés en décrivant l'énergie d'une somme d'une série de contributions rendant compte des interactions intra et intermoléculaires. Pour chacune des contributions, des pénalités énergétiques sont appliquées lorsqu'une variable (par exemple, une longueur de liaison ou un angle de valence), s'écarte de sa valeur de référence. Ces variables du calcul sont alors les coordonnées internes du système.

La MM est une méthode empirique où les atomes (noyaux) sont représentés par des masses ou des sphères, et les liaisons par des ressorts de différentes forces (figure 1).

La MM n'utilise pas de fonction d'onde ni de densité électroniques. Les constantes des équations sont obtenues à partir des données expérimentales spectroscopiques et à partir des calculs ab initio.

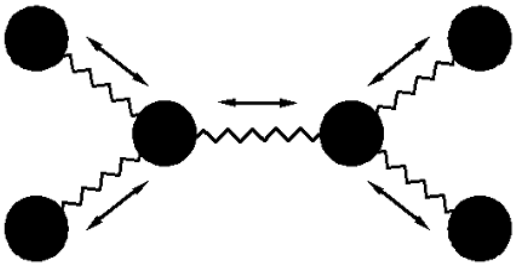


Figure 1 : Représentation mécanique d'une structure moléculaire

L'avantage principal de la MM par rapport aux autres méthodes est la rapidité des calculs. Ceci est dû à une simplification des équations du champ de force ; il est donc possible de traiter des systèmes moléculaires très grands. En revanche, cette méthode ne prend pas en compte la structure électronique moléculaire. Il est donc impossible d'étudier des systèmes dans lesquels les effets électroniques sont prédominants.

L'idée directrice de cette méthode est d'établir, par le choix des fonctions énergétiques et des paramètres qu'elles contiennent, un modèle mathématique, le «champ de force », qui représente aussi bien que possible les variations de l'énergie potentielle avec la géométrie moléculaire.

2.1 Champ de force :

Le champ de force est une expression que la mécanique moléculaire a empruntée à la spectroscopie vibrationnelle, en modifiant légèrement sa signification [22]. Sous ce terme sont en fait regroupés deux éléments : d'une part l'expression des différentes fonctions contribuant au calcul énergétique et d'autre part les valeurs des différentes constantes paramétrant ces fonctions. Ces paramètres sont identifiés à partir de données expérimentales (IR, micro -onde…) ou évalués théoriquement.

L'énergie totale du système est donc une énergie potentielle multidimensionnelle décrivant les interactions intramoléculaires ou interactions liantes (impliquant des atomes reliés par des liaisons explicites) et les interactions intermoléculaires ou interactions non liantes (impliquant des atomes non liés directement par des liaisons explicites) : $\mathrm{E}_{Totale} = E_{Intermoléculaire} + E_{Intramoléculaire}$

Par conséquence, le champ de force est un modèle du système réel que l'on veut étudier. De sa qualité descriptive dépend la qualité du résultat des simulations numériques par comparaison avec les valeurs expérimentales. L'amélioration des champs de forces, modèles de système réel, représente un enjeu majeur pour le succès des techniques de la simulation moléculaire à prédire avec précision les propriétés et le comportement de systèmes toujours plus nombreux et plus complexes.

- **Quelques champs de force**

Différents champs de force sont proposés dans la littérature, ils se distinguent les uns des autres par les termes dans le développement de l'expression de l'énergie de la molécule. Chacun a un domaine d'application spécifique de sorte que le choix d'un champ de force dépend des propriétés et de l'application du système que l'on veut étudier. De tels champs de force sont apparus en début des années 1970 et continent à évoluer aujourd'hui.

- MM2 / MM3 /MM4 : est le premier champ de force développé par Allinger et col. [23]. Il a été conçu au début pour les molécules simples (alcanes, alcènes, alcynes non conjugués, amines…), mais ses versions améliorées MM3 (1989) [24] et MM4 (1996) [25] permettent de traiter des molécules organiques de plus en plus complexes.

- OPLS : Le programme OPLS (Optimized Potentials for Liquid Simulations), comme l'indique son nom, est conçu pour optimiser le potentiel qui permet la description des propriétés de solvatation. Il est écrit par W. L Jorgensen et J. Tirado Rives [26].

- GROMOS : (Groningen Molecular Simulation Program Package), est écrit par Van Gusteren [27] et conçu spécialement pour les biomolécules en milieu aqueux en vue de l'étude des interactions entre les molécules d'eau et les groupements polaires des protéines.

- CHARMM (Bio+) : Développé par Karplus et col. [28], pour le calcul de biomolécules. Son concept est semblable à celui d'AMBER. Bien qu'au début, ce champ de force est conçu pour les aminoacides et les protéines, maintenant il traite d'autre biomolécules.

-SPASIBA: (Spectroscopic Potentiel Algorithm for Simulating biomolecular Conforamtional Adaptability), élaboré par Gérard Vergoten et col. (1995). Il combine le champ de force spectroscopique modifié de Urey-Bradly-Shimanouchi [29] et le champ de force AMBER [30]. Il permet de trouver les structures, les énergies de conformation et les fréquences vibrationnelles au minimum énergétique d'une molécule [31]

-AMBER: (Assisted Model Building with Energy Refinement), a été écrit par Kollman [32]. Le champ est paramétré pour les protéines et les acides nucléiques (UCSF, 1994). Il a été utilisé pour les polymères et pour d'autres petites molécules.

2.2 But de la mécanique moléculaire : Principes de la minimisation

La mécanique moléculaire a pour but de trouver le minimum de la fonction énergie E. Pour avoir un minimum global de l'énergie, il serait nécessaire de parcourir tout l'espace des variables indépendantes, ce qui est impossible vu leur nombre important. Toutes les méthodes de minimisation ne permettent de trouver que des minimums locaux et la surface d'énergie pour un tel nombre de variables est très accidentée. Les structures trouvées par minimisation d'énergie sont donc toujours relativement proches de la structure de départ. A partir d'une

géométrie très approximative, il faut chercher le jeu de coordonnées qui réduit au minimum la somme de toutes les contributions énergétiques dues aux déformations 3N-2 coordonnées internes et aux interactions entre atomes non liés.

La minimisation de la fonction énergie s'effectue par une dérivation de l'équation de l'énergie de la molécule par rapport à chacun des degrés de liberté de la molécule, et en cherchant le lieu où les dérivées s'annulent simultanément.

Il est à noter que le minimum global d'énergie est très difficile à trouver car les paramètres conformationnels à partir desquels, la minimisation va être effectuée sont primordiaux : les méthodes de minimisation ne font en général que proposer le minimum d'énergie le plus proche.

Les différentes méthodes de minimisation les plus utilisées sont :

- La méthode de la plus grande pente appelée «Steepest descent».
- La méthode du gradient conjuguée.
- La méthode de Newton Raphson.

La conjugaison des différentes méthodes de minimisation compte tenu des avantages et limitations de chacune des méthodes de minimisation de l'énergie, celles-ci sont le plus souvent conjuguées afin d'utiliser leur vitesse de convergence et leur précision dans leur domaine d'application.

Ainsi, en mécanique moléculaire, la minimisation d'une structure moléculaire met souvent en œuvre un calcul de type "steepest descent", permettant une convergence rapide vers le minimum énergétique le plus proche, suivi d'un calcul de type Newton-Raphson/Gradient conjugué, permettant une convergence précise vers ce minimum.

3. Dynamique moléculaire

La dynamique moléculaire est la méthode la plus fréquemment utilisée pour la simulation de systèmes réels. Le principe de base de cette méthode est l'échantillonnage de l'espace conformationnel du système étudié par l'intégration des équations du mouvement de Newton pour tous les atomes présents dans le système.

Les simulations de DM consistent à calculer les positions et les vitesses d'un système d'atomes [33]. Elles sont très importantes pour la recherche du comportement structural des biomolécules en fonction du temps. En utilisant la DM, on peut étudier la flexibilité ou la rigidité des biomolécules, mesurer les interactions intermoléculaires

entre la protéine et les ligands ou d'autres biomolécules, calculer l'énergie libre ou bien étudier l'effet du solvant sur la structure des biomolécules. Ainsi, au contraire de la mécanique moléculaire, la dynamique moléculaire produit des conformations qui dépendent moins de la structure initiale.

Une simulation de dynamique moléculaire se réalise généralement en quatre étapes:

1- Une étape de minimisation de la structure initiale destinée à éliminer les contacts stériques.

2- Une étape dite de thermalisation au cours de laquelle les vitesses des atomes sont augmentées progressivement afin d'atteindre la température finale choisie.

3- Une étape d'équilibre pendant laquelle les vitesses ne sont plus modifiées. L'énergie cinétique se répartit sur toute la molécule afin d'atteindre son état d'équilibre.

4- Enfin, durant la dernière étape, appelée dynamique productive, les coordonnées et les vitesses sont sauvegardées pour une analyse de la dynamique.

Conclusion

Ce chapitre est un aperçu sur les bases de la chimie théorique, les méthodes quantiques de l'équation de Schrödinger, la mécanique moléculaire, ainsi que la dynamique moléculaire. Ces méthodes nous servirons pour l'optimisation de la géométrie des molécules avant d'utiliser leur structure pour le développement de modèles QSPR.

.

Références

1. Höltje H.D. Folkeis G. Molecular Modeling: Basic Principles and Applications. VCH, New-York, 1997

2. Audouze C. Vers une parallelisation par bandes en chimie quantique,Laboratoire de Mathematique, UMR CNRS 8628, Universite Paris-Sud, 2003.

3. Liotta D. Advances in Molecular Modeling, 1, JAI Press, Greenwich, 1988

4. Tsai C.S. An Introduction to Computational Biochemistry, Wiley-Liss, New York, 2002.

5. Leach A.R. Quantum Mechanical Models, in Molecular modelling: Principles and applications. Addison Wesley Longman Ltd., Harlow, 1996.

6. Schrodinger E. Ann. phys. Leipzig.1926.76. 361.

7. Born M. Oppenheimer R. Quantum theory of the molecules, Annalen der Physik. 1927. 84. 457.

8. Hehre W.J. Radom L. P. Schleyer V.R. Pople J.A. Ab Initio Molecular Orbital Theory. JohnWiley and Sons, New York. 1986.

9. Mulliken R.S. Electronic population analysis on LCAO-MO molecular wave functions. The Journal of Chemical Physics., 1955, 23, 1833.

10. Leach A.R. Molecular Modelling: Principles and Applications. Pearson: Prentice Hall, Harlow, 2001

11. Dewar M.J.S. Zoebisch E.G. Healy E.F., Stewart J.J.P. AM1: A New General Purpose Quantum Mechanical Molecular Model. J. Am. Chem. Soc. 1985. 107. 3902.

12. a) Stewart J.P.P. Optimization of parameters for semi-empirical methods I. Method", The Journal of Computational Chemistry. 1989, 10, 209

    b) J.P.P. Stewart. Optimization of parameters for semi-empirical methods II. Applications. Journal of Computational Chemistry. 1989, 10, 221.

13. Young D.C. Computational Chemistry: A Practical Guide for Applying Techniques to Real World Problems", Éditions Wiley Inter-Science, Chapitre 6 - Molecular Mechanics. 2001 pp.49-59

14. Fock V. Näherungsmethode zur losung des quanten-mechanischen mehrkörperprobleme. Zeitschrift für Physik . 1930. 61. 126.

15. Cramer C.J. Essentials of Computational Chemistry: Theories and Models. John Wiley and Sons, New York, 2002.

16. Møller C. Plesset M.S. Note on an Approximation Treatment for Many Electron Systems", Phys. Rev. 1934. 46. 618.

17. Shavitt I. Shaefer H.F. Methods of Electronic Structure Theory", Ed., Plenum Press, New-York, 1977.189.

18. Jugl A. Chimie Quantique Structurale et Éléments de Spectroscopie Théorique. 1978.

19. Cizek J. On the Correlation Problem in Atomic and Molecular Systems. Calculation of Wavefunction Components in Ursell-Type Expansion Using Quantum-Field Theoretical Methods", Journal of Chemical Physics. 1966, 45, 4256.

20. Williams. J.E, P. Strang. J, P. von R. Schleyer, Ann. Rev. Phys. Chem., 1968, 19, 531

21. Allinger. N.L, Am. J Chem. Soc. , 1977, 99, 8127

22. Lomas. J.S, L'actualité chimique. La mécanique moléculaire, une méthode non quantique pour le calcul de la structure et de l'énergie d'entité moléculaire, Mai 1986.

23. Frankie Tristram, Viviane Durier, Gérard Vergoten, Journal of Molecular Structure., 1996 ; 378, 249-256.

24. Rahman. A, Stillinger. F.H, J. Chem. Phys., 1971, 5, 3336.

25. Allinger L.N, K. Chen, J-H. Lii, J. Comp. Chem., 1996, 17, 642

26. Jorgensen. W.L, Rives. J.T, J. Am. Chem. Soc., 1988, 110, 1657

27. Van Gusten. W.F, Karplus. M, Macromolecules, 1982, 15, 1528

28. Brooks .B.R, Bruccoleri. R.E, B. D. Ofalson, D. J. States, S. Swaminathan & M. Karplus, J. Comp. Chem. , 1983, 4, 187.

29. Shimanouchi.T, Pure Appl. Chem. , 1963, 7,131.

30. Weiner S.J, Kollman. P.A, Nguyent. T, D. A. Case, J. Comput .Chem. , 1986, 7, 230.

31. Tristram. F, Durier. V, Vergoton. G, J. Mol. Struct. , 1996, 378, 249-256.

32. J.A. McCammon, S.C. Harvey: "Dynamics of Proteins and Nucleic Acids", Cambridge, 1987.

CHAPİTRE 4
QSAR/QSPR

İntroduction

- **Historique**
- **Définition**
- **Principe et théorie**
- **Méthodologie des études QSPR**
- **Validation d'un modèle QSPR par les principes OCDE**
- **İnterprétation des modèles**

Conclusion

İntroduction

La connaissance des propriétés et des activités est d'une importance capitale pour pouvoir classer et utiliser les composés chimiques. La caractérisation expérimentale complète est difficile, voire impossible, pour des raisons de temps, de coût, de dangerosité de certains essais ou d'éthique (limitations des essais sur les animaux). L'utilisation des méthodes alternatives à l'expérience est devenue plus qu'indispensable. Parmi ces méthodes, on trouve les méthodes de modélisation moléculaire qui permettent de justifier les données expérimentales disponibles et prédire les propriétés pour des composés nouveaux ou des composés pour lesquels les données expérimentales ne sont pas disponibles. Ces méthodes s'appuient sur le principe que les propriétés physico-chimiques et les activités biologiques des molécules dépendent fortement de leurs structures chimiques.

L'acronyme QS/XR [X= A (activité), P (propriété), R (indice de rétention), T (toxicité)…] est utilisé lorsqu'une propriété est modélisée.

1. Historique

En fait, les premiers travaux QSPR remontent au 19ème siècle. En effet, dès 1868, Crum-Brown et Fraser [1] ont postulé l'existence de relations entre les activités physiologiques et les structures chimiques en reliant les variations de l'activité biologique à des modifications structurales. Cependant, à cette époque, les structures moléculaires n'étaient pas encore connues.

Une avancée importante vers les modèles QSPR proprement dits a été réalisée grâce au développement des équations de Hammett [2] dans lesquelles les constantes σ caractérisent de manière quantitative les vitesses de réactions pour les composés organiques [2,3].

Les premiers travaux utilisant la méthodologie QSPR/QSAR telle qu'employée actuellement sont dus à Hansch [4] ainsi qu'à Free et Wilson [5]. D'un côté, Hansch a proposé des modèles reliant directement l'activité biologique des composés avec les propriétés hydrophobes, électroniques et stériques à l'échelle moléculaire. D'un autre côté, Free et Wilson ont développé des modèles empiriques, dits de contributions de groupes, pour l'étude de l'activité biologique.

Au cours de ces dernières décennies, l'utilisation des méthodes QSPR/QSAR n'a pas cessé de progresser. Elle est même devenue indispensable en chimie pharmaceutique, en toxicologie et pour la conception de médicaments [6-8].

2. Définition et principe

La relation quantitative structure-propriété (QSPR) est une alternative au processus expérimental qui prédit les diverses propriétés physico-chimiques.

Les études (QSPR) sont des techniques appliquées pour l'estimation des propriétés physico-chimiques pour les substances qui n'ont pas été examinées expérimentalement. Malgré l'amélioration des équipements de laboratoire, l'analyse expérimentale de tous les produits récemment synthétisés est impossible. Ainsi, une approche QSPR fournit un compromis nécessaire qui permet l'estimation des paramètres physicochimiques de grandes classes de composés qui sont importants du point de vue théorique ou pour les applications industrielles [9].

Un modèle QSPR a la forme d'une équation mathématique

$$\textbf{Propriété} = f(x_1, x_2 \ldots, x_n) \qquad (1)$$

Cette équation concerne notamment les propriétés physicochimiques QSPR, et l'indice de rétention QSRR.

$x_1, \ldots, x_n$: descripteurs moléculaires d'une certaine fonction (f) de n variables. La fonction (f) peut être inconnue, complexe ou non linéaire.

Les études QSRR (Quantitative Structure-Retention Relationships) sont des dérivés d'études QSPR (Quantitative Structure-Property Relationships) Le principe de ces études est d'établir une corrélation entre des données structurales de la molécule (appelée descripteurs) et leur indice de rétention.

En QSRR, la rétention des composés dépend des conditions expérimentales notamment de la phase stationnaire (les colonnes) utilisée dans la chromatographie (CPG). L'objectif principal des études QSRR est de prévoir l'indice rétention des composés ciblés à partir des descripteurs de la structure moléculaire.

Le point de départ de telles méthodes se construit sur la définition des descripteurs moléculaires empiriques ou théoriques. Ces dernières prennent en compte des informations sur la structure et les caractéristiques physico-chimiques des molécules.

Le choix de la base de données expérimentale de référence est décisif dans une étude QSPR. Enfin, le lien entre les descripteurs et la base de données est déterminé grâce à des outils d'analyse comme les régressions multilinéaires (MLR), les arbres de décisions, les réseaux de neurones, et les algorithmes génétiques.

Les informations extraites à partir des résultats d'études QSPR peuvent être utilisées pour obtenir une meilleure connaissance des structures moléculaires et probablement le mode d'action au niveau moléculaire. Ces informations peuvent alors être utilisées pour prévoir les propriétés physicochimiques de nouveaux composés ainsi que pour concevoir de nouvelles structures [10].

3. Méthodologie générale des études QSPR

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

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

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

c- Diviser cette base de données, en une série d'apprentissage (training set) qui contient généralement les 2/3 de la base de données et une série de test (test set) constituée par le 1/3 restant.

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

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

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

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

h- Explorer et exploiter les modèles validés pour comprendre les mécanismes et les modes d'action [11].

Le schéma fonctionnel de la figure 1 représente la stratégie générale d'une étude QSPR

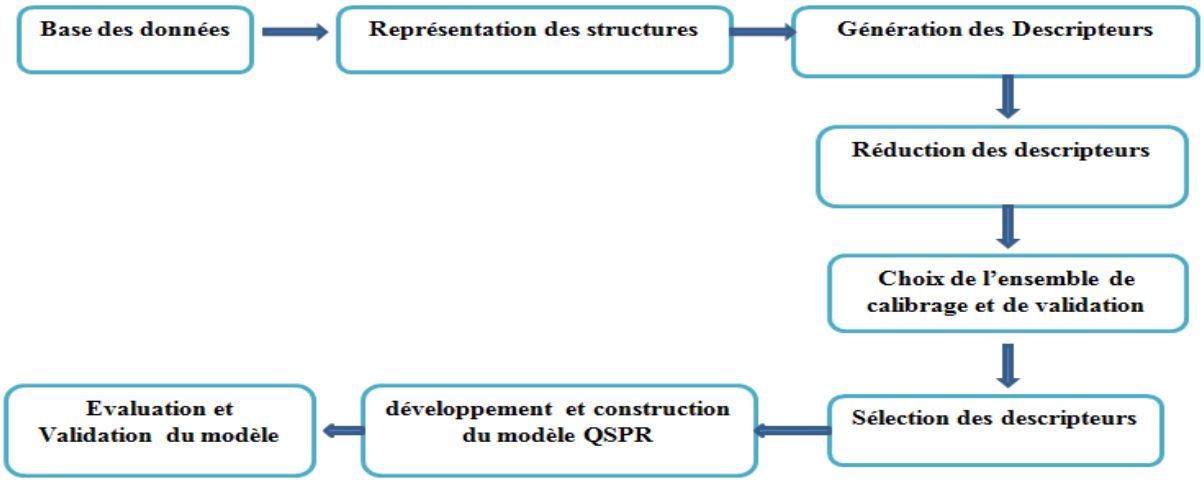


Figure 1 : La stratégie générale d'une étude QSPR

3.1. Base des données

Une partie préliminaire primordiale pour le développement ainsi que pour la validation des modèles QSPR/QSAR : est la sélection de la base de données. Ces données devraient, idéalement, être de grande qualité, ce qui signifie qu'elles devraient être fiables et cohérentes. Il est donc important de les choisir parmi celles présentant des incertitudes faibles afin de limiter les barres d'erreurs expérimentales. De plus, le modélisateur doit s'assurer que les données expérimentales utilisées ont été obtenues selon le même protocole dans les mêmes conditions expérimentales car ils ont une forte influence sur les valeurs obtenues et donc sur la qualité des modèles [12]. Il faut également que la distribution des données soit la plus homogène et normale, car la plupart des méthodes statistiques sont basées sur ce type de distribution [13].

En effet la taille de la base de données est aussi importante : une base de données trop petite rend difficile l'obtention de modèles prédictifs et parfois impossible leur validation. Il faut aussi veiller aux erreurs dans les structures des molécules, qui sont la base des modèles QSPR, dont la bonne représentation est cruciale. Il est indispensable de ne pas avoir d'erreur à ce niveau. Il faut être très attentif aux structures et corriger les éventuelles erreurs comme : les incohérences entre numéro CAS et nom de la molécule, la présence des valeurs de propriétés différentes pour un même composé. Une seule erreur à ce niveau peut imposer de redévelopper complètement les modèles et avoir une grande influence sur leur prédictivité si l'erreur n'est pas repérée. Il faut aussi vérifier l'absence de redondance [14] et les données aberrantes (outliers), qui sont des molécules ayant une structure trop différente des autres

molécules de la base de données ou une valeur de propriété complètement différente, ces données doivent être supprimées afin de ne pas perturber le développement des modèles.

Les données utilisées dans les évaluations QSAR sont obtenues soit à partir de la littérature, soit générées spécifiquement pour les analyses de type QSAR.

3.2 Représentation des structures.

Les structures peuvent être représentées sous forme de graphique 2D avec suppression de l'hydrogène. Des logiciels sont disponibles pour dessiner cette structure 2D. On peut citer Chem Draw qui utilise le nom IUPAC (International Union of Pure and Applied Chemistry) [15] pour la représentation des structures. Et le logiciel de modélisation moléculaire HyperChem qui peut générer une structure tridimensionnelle (3-D) obtenue à partir de structures 2-D qui sont nécessaires à la génération de descripteurs géométriques.

3.3 Génération des descripteurs

Une étape majeure dans le développement des modèles QSPR est la génération de descripteurs moléculaires, qui servent à quantifier des caractéristiques physicochimiques ou structurelles de molécules. Les descripteurs réalisent un codage numérique de l'information chimique en vecteurs réels, une fois que les nombres sont disponibles, il est possible d'établir une relation entre ceux-ci et une propriété moléculaire, à l'aide d'outils de modélisation classiques.

Beaucoup de logiciels calculent de larges ensembles de descripteurs théoriques différents, des SMILES, des graphiques 2D aux coordonnées 3Dx,y,z, parmi lesquels nous mentionnerons: ADAPT [16], OASIS [17], CODESSA[18], MolConnZ [19] et DRAGON [20] (logiciel utilisé dans ce travail pour la génération de plus de 1600 descripteurs).

On estime que plus de 10 000 descripteurs moléculaires sont maintenant disponibles, et la plupart d'entr'eux ont été résumés et expliqués [21- 23]. Le grand avantage des descripteurs théoriques est qu'ils peuvent être calculés de manière homogène par un logiciel défini pour tous les produits chimiques, même ceux qui n'ont pas encore été synthétisés, le seul besoin étant une structure chimique présumée, ils sont donc reproductibles.

Les descripteurs moléculaires sont fréquemment classés par rapport à la dimensionnalité de la représentation moléculaire sur laquelle ils sont calculés : on parlera alors de descripteurs 0D, 1D, 2D, ou 3D [24].

3.3.1. Les descripteurs 0D

Tous les descripteurs moléculaires pour lesquels aucune information sur la structure moléculaire et les connectivités atomiques n'est nécessaire appartiennent à la classe des descripteurs 0D. Le nombre d'atome(s) et de liaison(s), ainsi que la somme ou la moyenne des propriétés atomiques sont typiques de cette classe de descripteurs. Ces descripteurs peuvent toujours être facilement calculés, interprétés naturellement, ne nécessitent pas d'optimisation de la structure moléculaire et sont indépendants de tout problème conformationnel. Ils montrent généralement une très forte dégénérescence, c'est-à-dire qu'ils ont des valeurs égales pour plusieurs molécules, telles que les isomères [25].

3.3.2. Les descripteurs 1D :

Ils sont accessibles à partir de la formule brute de la molécule et décrivent des propriétés globales du composé comme le nombre d'atomes et la masse moléculaire etc…. Ces descripteurs sont couramment utilisés du fait de leur extrême simplicité. Cependant, ils peuvent poser problème pour une bonne interprétation des mécanismes d'interaction du fait qu'ils ne permettent pas de tenir compte des effets stériques et d'isomérie.

3.3.3. Les descripteurs 2D

Les descripteurs moléculaires utilisant la représentation des molécules sous forme de graphes sont dits « descripteurs 2D » et contiennent des informations relatives à la connectivité ou à certains fragments moléculaires, mais aussi des estimations des propriétés physico-chimiques. C'est à partir de ce niveau que l'on peut espérer la capture d'informations chimiques pertinentes pour la prédiction de la majorité des propriétés moléculaires. On trouvera dans cette catégorie les descripteurs suivants :

- **Les indices topologiques**, qui considèrent la structure du composé comme un graphe, les atomes étant les sommets et les liaisons les arêtes. De nombreux indices quantifiant la connectivité moléculaire ont été développés en se basant sur cette approche, comme par exemple l'indice de Wiener [26], qui compte le nombre total de liaisons dans les chemins les plus courts entre toutes les paires d'atomes (en excluant les hydrogènes). D'autres indices basés sur les chemins ont été développés [27-29].
- **Les indices constitutionnels** caractérisent les différents composants de la molécule. Il s'agit par exemple du nombre de liaisons simples ou multiples, du nombre de cycles.

3.3.4. Les descripteurs 3D

Sont évalués à partir des positions relatives des atomes dans l'espace, et décrivent des caractéristiques plus complexes ; leurs calculs nécessitent donc de connaître, le plus souvent par modélisation moléculaire empirique ou ab-initio, la géométrie 3D de la molécule. On distingue plusieurs familles importantes de descripteurs 3D :

- **Les descripteurs géométriques** : Les plus importants sont le volume moléculaire, la surface accessible au solvant, le moment principal d'inertie.
- **Les descripteurs électroniques** : Permettent de quantifier différents types d'interactions inter- et intramoléculaires, de grande influence sur l'activité biologique des molécules. Le calcul de la plupart de ces descripteurs nécessite la recherche de la géométrie pour laquelle l'énergie stérique est minimale, et fait souvent appel à la chimie quantique. Par exemple, les énergies de la plus haute orbitale moléculaire occupée (HOMO) et de la plus basse vacante (LUMO) (orbitales frontières) sont des descripteurs fréquemment sélectionnés.
- **Descripteurs spectroscopiques.** Les molécules peuvent être caractérisées par des mesures spectroscopiques, par exemple par leurs fonctions d'onde vibrationnelles. En effet, les vibrations d'une molécule dépendent de la masse des atomes et des forces d'interaction entre ceux-ci ; ces vibrations fournissent donc des informations sur la structure de la molécule et sur sa conformation. Le descripteur (EVA) [30] est ainsi obtenu à partir des fréquences de vibration de chaque molécule. Les descripteurs de type MoRSE [31] (Molecule Representation of Structures based on Electron diffraction) sont calculés à partir d'une simulation du spectre infrarouge ; ils font appel au calcul des intensités théoriques de diffraction d'électrons.

3.4 Réduction des descripteurs

Ainsi, plus d'un millier de descripteurs peuvent être obtenus mais tous ne sont pas nécessaires au développement du modèle. Des méthodes de sélection de variables sont disponibles pour réduire ce nombre, notamment afin de ne pas obtenir des équations sur-paramétrées. De manière générale, la réduction des descripteurs commence par la suppression des données redondantes c'est-à-dire très corrélées entre elles. De plus, les descripteurs considérés comme pertinents sont ceux ayant une grande corrélation avec la propriété et ayant une variance significative sans laquelle le descripteur ne permet pas la distinction des différentes données entre elles.

L'ensemble de descripteurs doit donc être le plus petit possible mais le plus riche en informations possible. Pour cette raison, les méthodes utilisées doivent choisir les descripteurs les plus informatifs.

Finalement, le sens chimique des descripteurs utilisés doit, bien entendu, être pris en considération puisque, plus les descripteurs sont reliés chimiquement au phénomène, plus la probabilité de faire intervenir des descripteurs par le biais du hasard est réduite [32].

3.5 Choix de l'ensemble de calibrage et de validation

Avant de commencer le développement des modèles, nous cherchons à diviser les données en deux sous-ensembles : un pour le calibrage et un pour la validation externe du modèle. Le sous-ensemble de calibrage doit être représentatif des données initiales et le sous-ensemble de la validation doit être choisi pour évaluer la qualité du modèle. Il existe de nombreux algorithmes de sélection de ses deux sous-ensembles qui se différencient principalement par leurs techniques de base. Nous citerons les plus utilisées par les praticiens du QSAR [33] :

- Sélection aléatoire : L'ensemble de données peut être divisé par un simple processus de sélection aléatoire de l'ensemble de calibrage et de test (pour la validation externe).

- Basé sur la réponse Y : Cette approche est basée sur l'échantillonnage de l'activité (Y-response). La gamme complète de la réponse est divisée en bacs et les composés appartenant à chaque bac sont affectés aux ensembles de calibrage ou de test de façon aléatoire ou personnalisée.

- Basé sur la réponse X : Les propriétés et la similarité structurelle des molécules sont considérées pour le groupement de composés similaires. Ensuite, une fraction pré-décidée des composés est affectée au calibrage ou au jeu de tests manuellement ou de façon régulière. Parmi les outils les plus couramment utilisés pour la division rationnelle des ensembles de données on peut citer :

• K-Means clustering [34],
• La sélection par carte auto-organisée de Kohonen [35],
• La conception moléculaire statistique [36],
• Les sphères d'exclusion [37],
• Sélection du jeu de test orientée vers l'extrapolation [38],
• Algorithme DUPLEX [39],et
• Algorithme de Kennard-Stone (CADEX) [40],

- **L'Algorithme de répartition « Kennard et Stone » (CADEX)**

Une alternative à la sélection aléatoire est l'utilisation de l'algorithme de Kennard et Stone [40 ,41]; l'algorithme maximise la distance euclidienne minimale entre les échantillons déjà sélectionnés et les échantillons restants.

Cette procédure illustrée par la figure 2, est rappelée ci-après :

a) sélection des échantillons les plus éloignés. Il s'agit ici des échantillons 1 et 2 qui sont entourés sur la Figure 2a ;

b) pour chaque échantillon restant, calcul de la distance euclidienne par rapport à l'échantillon le plus proche déjà sélectionné (Figure 2b) ;

c) sélection de l'échantillon ayant la plus grande distance avec l'échantillon déjà sélectionné. Le troisième échantillon sélectionné est l'échantillon n°4.

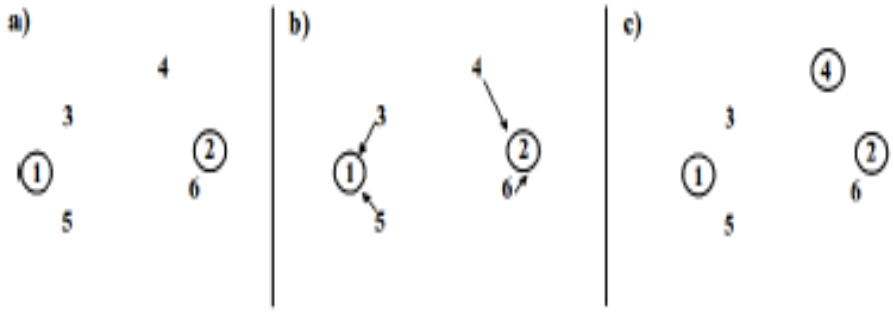


Figure 2 : Répartition des échantillons avec l'algorithme de Kennard et Stone

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. L'algorithme CADEX de Kennard et Stone est considéré comme l'un des meilleurs moyens pour la construction des ensembles de calibrage et de validation (test) [42].

- **Algorithme DUPLEX**

Une version améliorée appelée DUPLEX a été proposée par Snee [43] ; elle est largement utilisée dans le domaine de la chimiométrie. 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 :

$$z_{ij} = \frac{x_{ij} - \overline{x}_j}{S_{jj}^{1/2}} \qquad i = 1, \ldots\ldots, n \;\; ; \;\; j = 1, \ldots, \ell \qquad \textbf{(2)}$$

où

$$S_{jj} = \sum_{i=1}^{n} \left(x_{ij} - \overline{x}_j \right)^2 \; ; \; \overline{x}_j \text{ étant la moyenne du j-ème régresseur.} \qquad \textbf{(3)}$$

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 :

$$\boldsymbol{z' \cdot z = t' \cdot t} \qquad \textbf{(4)}$$

t est une matrice ($\ell \times \ell$) triangulaire supérieure unique, dont les éléments peuvent être obtenus par la méthode de Cholesky. On opère alors la transformation :

$$\boldsymbol{w = z \cdot t^{-1}} \qquad \textbf{(5)}$$

Qui conduit à un nouvel ensemble de variables w orthogonales et de variance unité. Celles-ci sont utilisées pour calculer la distance euclidienne, entre les C_n^2 paires de points. Les 2 points les plus éloignés sont sélectionnés pour l'ensemble de calibrage, puis parmi les points restants, les 2 plus éloignés sont sélectionnés pour la validation (ensemble de test). Puis parmi les points restants, le plus éloigné des points de calibrage précédemment sélectionnés est sélectionné pour le calibrage. Puis parmi les points restants, le plus éloigné des points de validation précédemment sélectionnés est sélectionné pour la validation. Puis l'algorithme continue à placer les points restants, alternativement dans l'ensemble de calibrage et dans l'ensemble de validation, jusqu'à ce que les n points soient affectés.

Les ensembles de calibrage et de validation n'étant pas forcément de même taille, l'algorithme DUPLEX peut séparer les données dans n'importe quel rapport souhaité.
De telles séparations sont réalisées en utilisant l'algorithme jusqu'à ce que l'ensemble de validation contienne le nombre de points requis, puis en versant les points non assignés dans l'ensemble de calibrage. L'utilisation de l'algorithme DUPLEX suppose que le nombre

d'observations, n, est tel que : $n \geq 2\ \ell + 25$, ℓ désignant le nombre de régresseurs ; l'ensemble de validation devant contenir 15 éléments au minimum [39 ,43].

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.

3.6 Sélection des descripteurs

Les meilleurs descripteurs sont sélectionnés en explorant la qualité statistique de toutes les combinaisons possibles des descripteurs disponibles, en utilisant la régression par les moindres carrés ordinaires (OLS) et la sélection de sous-ensembles de variables par algorithme génétique (AG-SSV). Ces programmes utilisent beaucoup d'outils pour l'analyse exploratoire des données : division d'ensembles de données en ensembles d'entraînement et de prédiction, détection de valeurs aberrantes et prédictions interpolées ou extrapolées, validations interne et externe par différents paramètres, modélisation consensuelle et divers graphes pour les visualisations sont implémentées.

Cette procédure de « sélection de variables » génère une « population » de modèles, classés selon les valeurs décroissantes de R^2. Les meilleurs modèles ont été choisis en utilisant le Q^2_{LOO} comme valeur d'optimisation et en tenant compte du principe de parcimonie concernant la dimension des modèles qui devrait être aussi petite que possible.

En outre, la corrélation entre les descripteurs et la réponse modélisée est vérifiée par la règle QUIK (Q Under Influence of K) pour exclure les modèles à forte colinéarité entre les descripteurs [44].

3.7 Développement et méthodes de modélisation QSPR

Pour élaborer un modèle QSPR nous avons besoin d'une méthode d'analyse de données, cette méthode permet de quantifier la relation qui existe entre la propriété et la structure (descripteurs). Il existe plusieurs méthodes pour construire un modèle et analyser les données statistiques de ce dernier, certaines sont linéaires telles que la régression linéaire multiple (MLR), d'autres sont non linéaires comme les arbres de décisions, les réseaux de neurones…

Ces méthodes sont disponibles dans des logiciels tels que, Excel, Minitab,Statistica, Matlab etc...

Dans l'ensemble de notre travail, nous avons utilisé la régression linéaire multiple MLR, et les réseaux de neurones artificiels RNA dans MATLAB pour la construction des modèles QSPR.

3.7.1 La Régression Linéaire Multiple (RLM)

La régression linéaire multiple RLM est l'une des méthodes de modélisation les plus en vogue grâce à sa simplicité d'utilisation et son interprétation facile. L'avantage important de la régression linéaire multiple est qu'elle est très transparente, puisque l'algorithme est disponible, et que les prédictions peuvent être réalisées facilement [45].

La méthode MLR se base sur l'hypothèse que la propriété dépend linéairement des différentes variables (les descripteurs), selon la relation :

$$y = a_0 + \sum_{i=1}^{n} a_i x_i \qquad \textbf{(6)}$$

Où a_i sont les coefficients de la régression.

Pour déterminer la valeur des coefficients a_i, la méthode des moindres carrés est utilisée. Elle a pour but de minimiser le carré des résidus SCR ou encore RSS représenté sur la Figure 3 c'est-à-dire la somme des carrées des écarts SCR entre les valeurs prédites et les valeurs réelles sur toute la base de données de p molécules.

$$SCR = \sum_{i=1}^{n} (y_{excp,i} - y_{calc,i})^2 \qquad \textbf{(7)}$$

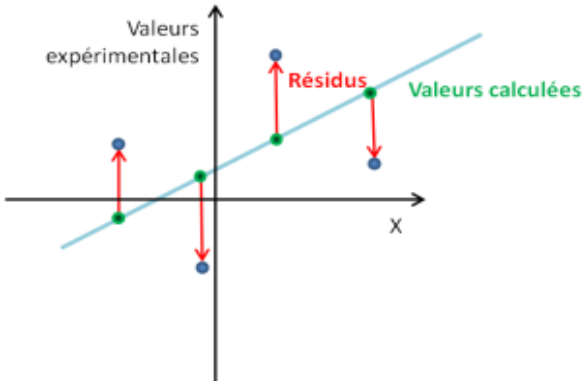


Figure 3 : Représentation graphique des résidus

En pratique, il s'agit de résoudre un système à p équations (correspondant au nombre de molécules) pour n variables (nombre de descripteurs) avec n<p en minimisant le SCR.

Ce système peut être résolu en utilisant une notation matricielle : $\boldsymbol{A} = (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{Y}$ **(8)**

Où A est la matrice des coefficients ai, X celle des variables xi et Y le vecteur contenant les valeurs de la propriété [46].

La régression linéaire multiple présente certains désavantages. Le principal découle de sa linéarité. Elle est 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.

3.7.2 Les réseaux de neurones artificiels RNA

Les réseaux de neurones formels [47,48] sont devenus en quelques années des outils précieux dans plusieurs domaines. Néanmoins, ils n'ont pas encore atteint leur plein développement, pour des raisons plus psychologiques que techniques, liées aux connotations biologiques du terme, un "neurone formel" (ou simplement "neurone") est une fonction algébrique non linéaire et bornée, dont la valeur dépend de paramètres appelés coefficients ou poids (w_i).
Les variables de cette fonction sont habituellement appelées "entrées" du neurone, et la valeur de la fonction est appelée sa "sortie". Un neurone est donc avant tout un opérateur mathématique, dont on peut calculer la valeur numérique par quelques lignes de logiciel. On a pris l'habitude de représenter graphiquement un neurone comme indiqué sur la figure suivante.

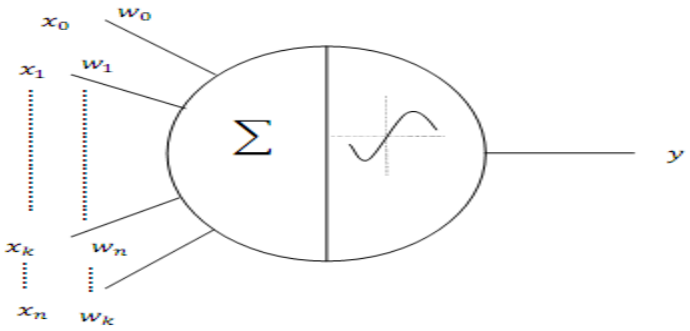


Figure 4 : Représentation d'un neurone formel.

X_k (où k = 1, 2, …, n) sont les neurones de l'entrée (descripteurs) et W_k (où k = 0, 1, 2, …, n) sont les poids. W_0 est le poids associé à une entrée fixée à 1, appelée biais. L'équation du neurone est de la forme suivante :

$$y = f\left(w_0 + \sum_{k=1}^{n} w_k x_k\right) \qquad \textbf{(9)}$$

Les neurones seuls réalisent des fonctions assez simples, et ce sont leurs combinaisons, qu'on appelle réseaux de neurones, qui permettent de construire des fonctions particulièrement intéressantes (figure 5).

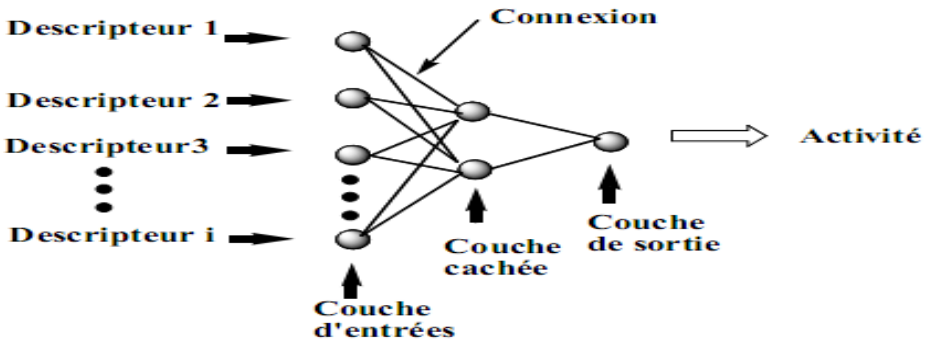


Figure 5 : Architecture des réseaux de neurones

Comme la figure 5 l'indique : un réseau de neurones est constitué de multiples couches : une couche d'entrée représentée par les descripteurs moléculaires, une ou plusieurs couches achées et une couche de sortie représentée par les propriétés à modéliser. Les neurones d'une couche sont interconnectés avec les neurones d'une couche voisine.

La couche de sortie compte autant de neurones que de propriétés modélisées ; dans notre cas une seule sortie.

Pendant la phase d'apprentissage du modèle par un réseau de neurones, les molécules sont présentées une à une aux neurones de la couche d'entrée. Les poids (wi) associés aux neurones d'entrée sont ajustés itérativement, afin de minimiser l'erreur entre la propriété calculée et la propriété expérimentale.

- **Apprentissage des réseaux de neurones artificiels**

Dans le domaine des réseaux de neurones, l'apprentissage est une phase très importante qui désigne la procédure ou la façon qui consiste à déterminer l'architecture et les paramètres du réseau [49].

L'apprentissage des réseaux de neurones artificiels se fait grâce à des algorithmes d'apprentissage qui permettent de faire varier la complexité du réseau en augmentant ou en diminuant le nombre de neurones cachés. Nous avons utilisé dans ce travail l'algorithme Levenberg-Marquardt de rétropropagation (fonction TRAINLM de la boîte à outils du logiciel MATLAB 7.0) [50] pour l'apprentissage du réseau, dont le principe consiste à minimiser la

différence entre la sortie calculée (propriété calculée) et les valeurs expérimentales de la propriété.

Défini en deux étapes, l'algorithme propage dans une première étape les entrées (descripteurs) vers l'avant jusqu'à obtenir une sortie (propriété) calculée par le réseau. Dans une seconde étape, la sortie calculée est comparée avec la valeur expérimentale. L'erreur obtenue par cette comparaison est ensuite rétro-propagée vers la couche d'entrée en modifiant les poids des neurones. Ce processus est réitéré jusqu'à l'obtention d'une erreur négligeable [51].

En effet, si les poids sont ajustés sur toutes les données de l'ensemble d'apprentissage (≈ 70% de la base de données globale), on risque d'avoir le « sur-apprentissage » ou l'apprentissage par cœur, dans ce cas le réseau apprend très bien les données présentées dans la phase d'apprentissage sans pour autant être capable de généraliser le modèle à des données nouvelles. Pour éviter le « sur-apprentissage » on introduit un nouvel ensemble de données appelé l'ensemble de validation (≈ 30% de la base de données globale). Comme pour l'ensemble de test (≈ 30% de la base de données globale), les éléments de cet ensemble ne participent pas à l'apprentissage. De plus, cet ensemble doit bien sûr avoir les mêmes contraintes que l'ensemble de test quant à sa représentativité et sa taille. L'ensemble de validation est utilisé de la façon suivante : dès que l'on s'aperçoit que l'erreur sur l'ensemble de validation stagne ou augmente, alors on arrête la procédure d'apprentissage [49].

Avant tout, il faut calculer les poids du réseau c'est-à-dire estimer les paramètres essentiels. Pour cela, il faut construire un réseau reliant directement les neurones représentant les descripteurs moléculaires choisis avec les neurones de sortie. Chaque descripteur est alors affecté d'un poids en fonction de l'importance de chacun d'entre eux dans la propriété étudiée. Ensuite, il faut choisir l'architecture du réseau d'apprentissage c'est-à- dire choisir les entrées externes, le nombre de neurones dans la couche cachée et l'arrangement des neurones entre eux. Le nombre d'unités cachées joue un rôle important dans la qualité du réseau. Si le nombre est trop petit, le réseau possède trop peu de paramètres et ne peut interpréter les dépendances servant à modéliser et prévoir. Si le nombre de neurones dans la couche cachée est trop grand, le réseau risque de s'ajuster au bruit présent dans les données de l'ensemble d'apprentissage [52].

4. Validation du modèle selon les principes de l'OCDE

Une fois développé, le modèle doit être interprété en analysant tous ses paramètres statistiques, sa qualité doit être aussi étudiée, cette qualité est vérifiée par ce que l'on appelle

validation, la validation des modèles QSAR a été reconnue ces dernières années, par des groupes d'experts spécifiques de l'OCDE [53] comme un point crucial et urgent, ce qui a conduit au développement, pour des raisons réglementaires, des « principes de l'OCDE pour la validation des modèles (Q)SAR ». La nécessité de cette action importante est principalement due à la nouvelle politique des produits chimiques de la Commission Européenne (REACH : enregistrement, évaluation et autorisation des produits chimiques), qui énonce explicitement la nécessité d'utiliser des modèles QSAR pour réduire les tests expérimentaux (y compris les essais sur les animaux). De toute évidence, pour satisfaire aux exigences de la législation REACH, il est essentiel d'utiliser des modèles QSAR qui produisent des estimations fiables, c'est-à-dire des modèles validés. Ainsi, un modèle QSAR fiable doit être associé aux informations suivantes :

1) Une définition précise de la propriété prédite par le modèle, incluant le protocole et les conditions expérimentales ;
2) Une équation mathématique (ou un algorithme) sans équivoque (reproductible), incluant la définition des différents paramètres employés ainsi que les méthodes de calculs éventuellement utilisées pour les obtenir ;
3) Un domaine d'applicabilité défini, permettant de déterminer pour quelles molécules les prédictions sont fiables ;
4) Des mesures appropriées des performances du modèle en termes de corrélation et de prédiction, incluant donc la mesure de son pouvoir prédictif pour un jeu de molécules de validation ;
5) Si possible, une interprétation des mécanismes moléculaires mis en jeu au travers des descripteurs employés et de la structure du modèle.

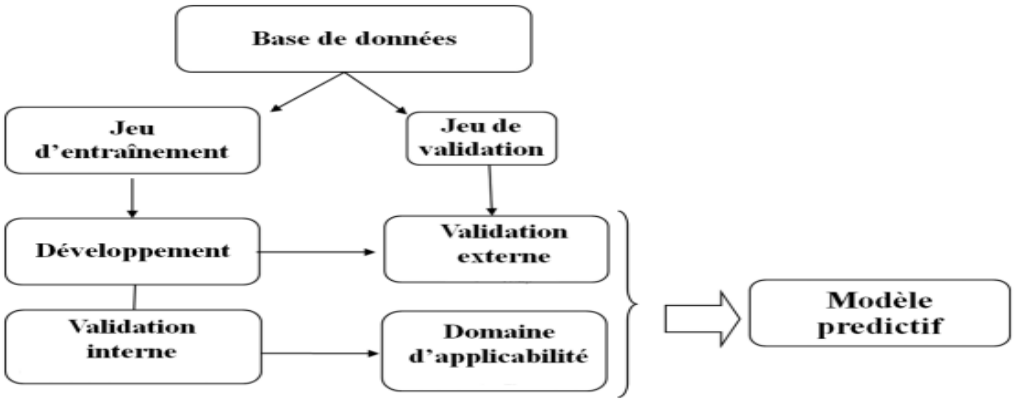


Figure 6 : Partage des données expérimentales pour le développement d'un modèle

La robustesse du modèle fourni par la modélisation c'est-à-dire l'influence des composés de la série d'apprentissage sur le modèle, est estimée par des méthodes de validation interne. Afin d'estimer son pouvoir prédictif, des données expérimentales supplémentaires sont nécessaires pour déterminer la capacité du modèle à prédire ces valeurs : c'est ce que l'on appelle validation externe. Enfin, il est important de savoir quel type de molécule utilisée avec quel modèle. On parle alors de domaine d'applicabilité.

- Afin de déterminer la qualité d'un modèle, différents indicateurs statistiques sont employés.

Le plus répandu d'entre eux est le coefficient de corrélation R^2 qui évalue la part de la variance expliquée par le modèle ; il est défini par la relation suivante, où $\bar{y}$ est la valeur moyenne des valeurs prédites, yi et y_i sont respectivement les valeurs observées et estimées:

$$R^2 = 1 - \frac{\sum(yi-\hat{y}_i)2}{\sum(yi-\bar{y}i)2} \quad \textbf{(10)}$$

L'erreur absolue moyenne (*MAE,* pour *mean absolute error*), est un autre indicateur utilisé défini par la relation

$$MAE = \frac{\sum|\hat{y}_i-y_i|}{P} \quad \textbf{(11)}$$

La déviation standard ***s*** définie par la relation s;

$$s = \sqrt{\frac{\sum(\hat{y}_i-y_i)^2}{P-n-1}} \quad \textbf{(12)}$$

L'indice de Fisher ***F*** est couramment employé pour mesurer le niveau de signification statistique du modèle, c'est-à-dire la qualité du choix du jeu de paramètres.

$$F = \frac{\sum(\hat{y}_i-\bar{y})^2}{\sum(y_i-\hat{y}_i)^2}\ \frac{P-n-1}{n} \quad \textbf{(13)}$$

La pertinence des descripteurs dans le modèle, est également évaluée par le test-*t* de Student. Il s'agit de tester l'hypothèse considérant le descripteur comme non significatif. Pour une régression multi-linéaire, cela revient à supposer le coefficient ai qui lui est associé comme nul.

Cette hypothèse est rejetée (avec un intervalle de confiance ***α***) si le ratio ti entre ai et son erreur type **s(ai)** atteint la valeur du fractile d'ordre **(1- *α/2)*** de la loi de Student à $\mathbf{(p_{n-2})}$ degrés de liberté.

$$|t_i| = \left|\frac{a_i}{s(a_i)}\right| > t_{1-\frac{\alpha}{2}}^{P-n-2} \quad \textbf{(14)}$$

Un facteur d'inflation de la variance (VIF) est employé pour tester l'existence de multicolinéarités parmi les descripteurs, qui est défini par la relation :

$$VIF = 1/(1-R^2) \quad \textbf{(15)}$$

Où R^2 est le coefficient de corrélation de l'équation de régression multiple entre les descripteurs du modèle. Si le VIF est égal à un, il n'existe pas d'intercorrélation pour chaque descripteur ; si le VIF se maintient dans la plage de 1,0 à 5,0, le modèle correspondant est acceptable ; si le VIF est supérieur à 10,0, le modèle correspondant est instable [54].

4.1. Validation interne

La validation interne est souvent la technique la plus employée, dans les études QSAR pour déterminer la stabilité du modèle et de tester l'influence de chaque échantillon de l'ensemble d'apprentissage sur le modèle final [55]. Pour ce faire, on emploie les techniques de la validation croisée (cross validation CV) leave-one-out et leave-*m*-out.

4.1.1 Validation croisée LOO (leave- one-out)

Pour déterminer la validation croisée LOO, l'ensemble de calibrage est principalement modifié en éliminant un composé de l'ensemble. Le modèle QSAR est ensuite reconstruit en fonction des molécules restantes de l'ensemble de calibrage en utilisant les descripteurs choisis, et la propriété du composé supprimé est calculée en fonction de l'équation QSPR Ce cycle est répété jusqu'à ce que toutes les molécules de l'ensemble de calibrage aient été supprimées une fois, et les données de la propriété prévue pour tous les composés de calibrage sont utilisées pour le calcul de divers paramètres de validation internes. Enfin, la prédiction du modèle est jugée à l'aide de la somme des carrés des erreurs de prédiction (PRESS) et du Q^2 pour le modèle tandis que la valeur de l'écart quadratique moyen de prédiction (SDEP) est calculée à partir du PRESS.

$$PRESS = \sum\left(Y_{obs} - Y_{pred}\right)^2 \quad \textbf{(16)}$$

$$SDEP = \sqrt{\frac{PRESS}{n}} \quad \textbf{(17)}$$

$$Q^2 = 1 - \frac{\sum\left(Y_{obs(train)} - Y_{pred(train)}\right)^2}{\sum\left(Y_{obs(train)} - \bar{Y}_{training}\right)^2} = 1 - \frac{PRESS}{\sum\left(Y_{obs(train)} - \bar{Y}_{training}\right)^2} \quad \textbf{(18)}$$

Dans les équations (16) et (18), Y obs et Y pred correspondent aux valeurs de la propriété bservée et prédite par LOO, n est le nombre d'observations, $Y_{obs\ (train)}$ est la propriété observée, $Y_{pred\ (train)}$ est la propriété prédite des molécules du jeu d'entraînement basée sur la technique LOO. La valeur seuil de Q^2 est de 0,5.

4.1.2 Validation croisée LMO (Leave-Many-Out)

Le principe de base de la technique LMO est qu'une partie définie de l'ensemble d'entraînement est supprimée et éliminée dans chaque cycle. Pour chaque cycle, le modèle est construit en fonction des molécules restantes (et en utilisant les descripteurs sélectionnés à l'origine), puis la propriété des composés supprimés est prévue en utilisant le modèle développé. Une fois tous les cycles achevés, les valeurs des propriétés prédites des composés sont utilisées pour calculer Q^2_{LMO}.

4.1.3 Validation par le test de randomisation

Afin de s'assurer qu'un modèle QSPR est fiable, les tests de randomisation de Y [56] sont une des techniques les plus employées. En effet, il n'est pas rare d'obtenir des corrélations fortuites (ou « chance correlation »), c'est-à-dire un modèle affichant de bons résultats statistiques (R^2, Q^2) pour l'apprentissage, mais impliquant des descripteurs qui dans la réalité ne sont pas reliés à la propriété modélisée. Ces modèles aléatoires peuvent être détectés par la procédure randomisation de Y. Elle consiste à mélanger aléatoirement les propriétés expérimentales pour le jeu d'apprentissage et, en utilisant les mêmes descripteurs (figure 7), à entraîner à nouveau l'algorithme d'apprentissage pour tenter d'obtenir un modèle. Normalement, les modèles obtenus doivent avoir des performances très faibles.

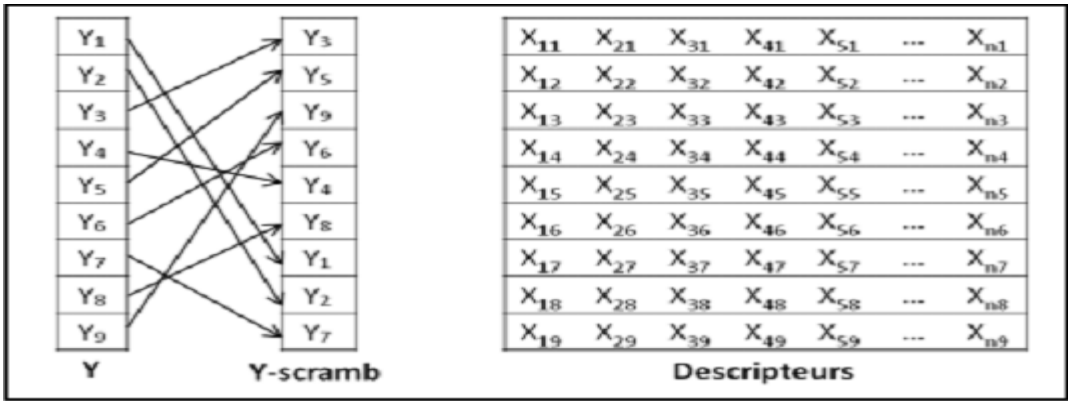


Figure 7 : Illustration de la méthode « Y-scrambling » (randomisation de Y)

La distribution des modèles obtenus permet de fixer un seuil heuristique de signification des modèles.

Cependant, la validation interne est insuffisante pour étudier le pouvoir prédictif d'un modèle. Pour cette raison la validation externe du modèle est devenue une norme et une partie obligatoire dans la modélisation QSPR [57,58].

4.2 Validation externe

Afin de tester de manière fiable le pouvoir prédictif du modèle QSPR, il est nécessaire d'employer un ensemble de validation externe, non employé pour le développement du modèle. Donc l'ensemble de molécules sera divisé en deux : un ensemble d'apprentissage sur lequel le modèle est développé et un ensemble de validation utilisé pour déterminer sa validité externe. Une fois l'ensemble de validation mis en place, il suffit alors d'appliquer le modèle QSPR aux molécules qui le composent et de déterminer la corrélation existante entre les propriétés calculées et celles expérimentales. Plus cette corrélation est importante, plus le modèle est capable de prédire les propriétés pour des molécules hors l'ensemble d'apprentissage [59,60].

Pour ce faire, une série de coefficients doit être vérifiée : R^2_{EXT}, RMSEP, les coefficients Q^2_{F1} [61], Q^2_{F2} [62], Q^2_{F3} [63, 64] et CCC [65-68].

- La capacité prédictive externe d'un modèle QSAR/QSPR peut aussi être déterminée par l'erreur quadratique moyenne dans la prédiction (RMSEP) donnée par l'équation 20 où n_{ext} désignant le nombre de composés de test (prédiction).

$$RMSEp = \sqrt{\frac{\sum(y_{obs\,(test)} - y_{pred(test)})^2}{n_{ext}}} \quad \textbf{(20)}$$

- L'article de Chirico et Gramatica [60] répertorie différents coefficients de calcul de la prédictivité Q^2_{F1}, Q^2_{F2}, Q^2_{F3}

$$Q^2_{F1} = 1 - \sum_{i=1}^{n_{ext}}(y_i - \hat{y}_i)^2 / \sum_{i=1}^{n_{ext}}(y_i - \bar{y}_{Tr})^2 \quad \textbf{(21)}$$

$$Q^2_{F2} = 1 - [\sum_{i=1}^{n_{ext}}(y_i - \hat{y}_i)^2 / \sum_{i=1}^{n_{ext}}(y_i - \bar{y}_{ext})^2] \quad \textbf{(22)}$$

$$Q^2_{F3} = 1 - [(\sum_{i=1}^{n_{ext}}(y_i - \hat{y}_i)^2 / n_{ext}) / (\sum_{i=1}^{n_{tr}}(y_i - \bar{y}_{tr})^2 / n_{tr})] \quad \textbf{(23)}$$

y_i La valeur expérimentale de la propriété, y_i la valeur prédite/calculée de la propriété et $\overline{y}_{Tr}$ La moyenne des valeurs y_i du jeu d'entraînement.

Le coefficient Q^2_{F1} proposé par Tropsha [57, 69,70] n'est pas une vraie validation externe car il fait intervenir des informations sur le jeu d'entraînement (training).

En 2008, une autre mesure de la prédictivité, proposée par Schüürmann [62], est le Q^2_{F2} qui se différencie de Q^2_{F1} 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^2_{F1} est plus optimiste [60] car $Q^2_{F1} \geq Q^2_{F2}$ et par conséquent accepte plus facilement les modèles. Le risque d'avoir un modèle non prédictif accepté est moins grand avec Q^2_{F2} .

En 2009, Le coefficient Q^2_{F3} a été proposé par Consonni [63] 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 [56]. En effet, il a été observé que la valeur de Q^2_{F3} est identique quel que soit le jeu de validation. Il semble également être insensible au nombre de molécules. En effet, la valeur de Q^2_{F3} ne change pas avec la taille du jeu de validation, contrairement à Q^2_{F2} dont la valeur augmente avec le nombre de molécules. Cependant, tout comme Q^2_{F1}, ce coefficient n'est pas une vraie validation externe car il fait intervenir des informations sur le jeu d'entraînement.

- Le coefficient de corrélation de concordance (CCC) proposé par Lin [71] légèrement réarrangé par rapport à l'original pour une lisibilité plus facile, parce qu' il est bien ajusté pour mesurer l'accord entre les données expérimentales et prédites, ce qui devrait être le véritable objectif de toute prédiction des modèles QSAR/QSPR :

$$CCC = 2\sum_{i=1}^{n_{ext}}(y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}}) / \sum_{i=1}^{n_{ext}}(y_i - \bar{y}_{ext})^2 + \sum_{i=1}^{n_{ext}}(\hat{y}_i - \bar{\hat{y}})^2 + n_{ext}(\bar{y}_{ext} - \bar{\hat{y}})^2 \quad \textbf{(24)}$$

Ce coefficient mesure à la fois la précision (à quelle distance des observations) et la justesse (c'est-à-dire 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.

4.3 Autres Critères de validation

Parmi les critères de validation les plus utilisés, Tropsha [57] propose d'accéder à la prédictivité du modèle en mesurant les coefficients de détermination lorsque la ligne de

régression passe par zéro, R_0^2 (valeurs prédites vs valeurs expérimentales) et R'^2_0 (valeurs expérimentales vs valeurs prédites) ainsi que les pentes k et k' de ces lignes de régressions :

- $Q^2>0,5$
- $R^2>0,6$
- $\frac{(R^2-R_0^2)}{R^2}<0,1$ ou $\frac{(R^2-R_0'^2)}{R^2}<0,1$
- $0,85<k$ ou $k'<1,15$

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 et MAE). Mais d'autres paramètres sont pris en considération pour le choix du modèle comme la possibilité d'interprétation des descripteurs.

4.4 Domaine d'applicabilité

Le 3ème principe de l'OCDE pour la validation des modèles QSAR/QSPR est la définition d'un domaine d'applicabilité. En effet, un modèle QSPR s'applique uniquement à des composés similaires à ceux avec lesquels le modèle a été développé. Le domaine d'applicabilité [72 -78] du modèle (AD) est l'espace chimique dans lequel le modèle est fiable et peut être interpolé. Il est déterminé sur les données du jeu d'entraînement à partir des descripteurs du modèle (voir Fig. 6) permet de déterminer si le modèle peut être utilisé pour prédire la propriété pour une nouvelle molécule donnée.

Il existe plusieurs méthodes pour la détermination du domaine d'applicabilité d'un modèle QSPR/QSAR parmi lesquelles on trouve la méthode du levier « leverage ». Cette méthode est basée sur la variation des résidus standardisés de la variable dépendante avec le levier (la distance entre les valeurs des descripteurs et leurs moyennes). L'effet de levier d'un composé dans l'espace est défini par [22]:

$$h_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \qquad (25)$$

Où xi est la valeur de descripteur du composé considéré, $\overline{x}$ la valeur moyenne du descripteur dans l'ensemble d'apprentissage alors que n est le nombre de substances dans l'ensemble d'apprentissage (calibrage) .

Si un composé a un résidu et un levier qui dépasse le seuil [h*=3p/n] (ou p est le nombre de descripteurs plus 1 et n le nombre d'observations), ce composé est considéré en dehors du domaine d'applicabilité du modèle élaboré.

En effet, un modèle QSAR/QSPR n'est pas destiné à être employé en dehors de son domaine d'application, c'est-à-dire en dehors de l'espace chimique couvert par son jeu d'entraînement.

Mis à part l'importance de la détermination des DA d'autres paramètres sont pris en considération pour le choix du modèle comme la possibilité d'interprétation des descripteurs.

5. Interprétation des modèles

L'interprétation des modèles est un point important [79, 80] recommandé dans la démarche de validation des modèles pour la prédiction dans un cadre règlementaire. Outre l'aspect interprétation pure des modèles, qui peut permettre une meilleure compréhension des phénomènes chimiques mis en jeu, l'utilisation de paramètres interprétables dans les modèles prédictifs présente l'intérêt de limiter les risques d'avoir choisi ces derniers par chance.

Les descripteurs ne sont pas forcément faciles à interpréter chimiquement, par exemple les indices topologiques, notamment, sont des constructions mathématiques caractérisant la taille et la forme des systèmes moléculaires mais sans caractérisation explicite.

Si leur efficacité en termes de prédiction n'est plus à démontrer, une interprétation physico-chimique est en général très difficile. Un autre obstacle à l'interprétation peut provenir du type de modèle choisi. Un grand nombre de descripteurs, par exemple, rend une équation difficilement interprétable du fait d'un trop grand nombre d'informations. De même, certains modèles non linéaires rendent l'interprétation de descripteurs, pourtant significatifs chimiquement, totalement impossible. En termes de démarche, l'interprétation peut être prise en compte dès l'étape de sélection des données. En effet, au cours du processus, il peut être nécessaire de choisir entre deux descripteurs très proches statistiquement. Une technique automatisée peut, par exemple, mener au choix du moins significatif du point de vue chimique pour peu que sa corrélation avec la propriété expérimentale soit très légèrement supérieure. Inclure une phase de choix manuel des descripteurs peut alors permettre d'intégrer plus aisément des considérations physico-chimiques.

Conclusion

La méthodologie QSPR a parcouru un long chemin depuis son introduction sous la forme d'approches classiques de Hansch et de Free-Wilson. Il a progressivement évolué avec le temps grâce au raffinement des approches, à l'utilisation de nouveaux descripteurs, à l'application de divers outils chimiométriques et à l'emploi de tests de validation rigoureux. Cette méthode est maintenant devenue une discipline scientifique distincte à part entière. Une bonne pratique de modélisation QSPR, bien que l'utilisation des lignes directrices recommandées par l'OCDE, puisse développer de bons modèles prédictifs avec des applications pratiques démontrées dans divers domaines de la chimie biologique, ce qui pourrait renforcer son acceptabilité par la communauté scientifique.

Références

[1] Crum Brown, T.R. Fraser, On the connection between chemical constitution and physiological action. Part 1. On the physiological action of salts of the ammonium bases, derived from strychnia, brucia, thebia, codeia, morphia and nicotia.Trans. Roy.Soc. Edinburgh, 1868, 25, 151-203.

[2] Hammett L.P. The effect of structure upon the reactions of organic compounds.b Benzene derivatives. J. Am. Chem. Soc. 1937. 59. 96.

[3] Brown A.C. Fraser T.R. On the connection between chemical constitution and Physiological Action; with special reference to the physiological action of the salts of the ammonium bases derived from Strychnia, Brucia, Thebaia, Codeia, Morphia, and Nicotia, J. Anat. Physiol . 1868. 2. 224.

[4] Hansch C. Leo A. Hoekmann D. Exploring QSAR: hydrophobic, electronic and steric constants.Washington, DC: Am. Chem. Soc.1995.

[5] Free S.M. Wilson J.W. A mathematical contribution to structure–activity studies. J. Med. Chem.1964. 7. 395.

[6] Grover M. Singh B. Bakshi M. Singh S. Quantitative structure- Property relationships in pharmaceutical research Part2. Pharm. Sci. Tech. Today, 2000. 3. 50.

[7] Grover M. Singh B. Bakshi M. Singh S. Quantitative structure–property relationships in pharmaceutical research–Part 1Pharm. Sci. Tech. Today, 2000. 3. 28.

[8] Katritzky A.R.D, Fara C, Petrukhin R.O, Tatham D.B, U Maran. Lomaka A, Karelson M. The present utility and future Potential for medicinal chemistry of QSAR/QSPR with whole molecule descriptors, Curr. Top. Med. Chem. 2002, 2,1333.

[9] Alan R.Katritzky, Victor S.Lobanov and Mati Karelson, Normal Boiling points of Organic compounds: Correlation and prediction by a Quantitative Structure Property Relationship, J.Chem.Inf.Comput.Sci. 1998.38.28.

[10] Phuong H.T. Thèse de doctorat, Synthèse et étude des relations structure/activité quantitatives (QSAR/2D) d'analogues Benzo[c]phénanthridiniques, Université d'Angers, 2007

[11] Thése études des relations

[12] Saihi. S, Thèse de doctorat, Etude de la relation quantitative structure-activité inhibitrice des enzymes hydrolytiques : cas des alpha-glucosidases, Université Badji mokhtar – Annaba, 2015.

[13] Tropsha, Mol Inform. 29, 2010, 476

[14] Dearden, J. C.; Cronin, M. T. D.; Kaiser, K. L. E. How Not to Develop a Quantitative Structure–activity or Structure–property Relationship (QSAR/QSPR). SAR and QSAR in Environmental Research 2009, 20, 241–266.

[15] Advenced Chemistry Development, Inc. IUPAC Nomenclature of Organics Chemistry http://www.acdlabs.com/iupac/nomenclature

[16] Stuper A J, ADAPT: A computer system for automated data analysis using pattern recognition techniques. Jurs, J. Chem. Inf. Comput Sci. 1976. 16, 99.

[17] Mekenyan O, Bonchev D, Oasis method for predicting biological activity of chemical compounds. Acta Pharm Jugosl. 1986. 36, 225.

[18] Katritzky A R, Lobanov V S, CODESSA, Version 5.3, University of Florida, Gainesville,1994.

[19] MolConnZ, Ver. 4.05, 2003, Hall Ass. Consult., Quincy, MA

[20] Todeschini R, Consonni V, Mauri A, Pavan M, DRAGON Software for the calculation of molecular descriptors. Ver. 5.4 for Windows, 2006, Talete srl, Milan, Italy.

[21] Devillers J, Balaban A T, (Eds.) Topological Indices and Related Descriptors in QSAR and QSPR, Amsterdam: Gordon Breach Sci. Pub. 1999. p 811

[22] Karelson M, Molecular descriptors in QSAR/QSPR. New York: Wiley-InterScience, 2000.448

[23] Todeschini R, Consonni V, Handbook of Molecular Descriptors, Wiley-VCH, Weinheim (Germany). 2000. 667

[24] Dudek A Z, Arodz T, Gàlvez J. Computational methods in developing quantitative structureactivity relationships (QSAR): A review. Comb. Chem. High. T. Scr, 2006.9.213.

[25] Bonachera F, Les triplets pharmacophoriques flous : développement et applications [thèse en ligne]. PhD thesis, Lille : Université Lille1 sciences et technologies, 2011. 22-26, 35, 37.

[26] Wiener H. Influence of interatomic forces on paraffin properties, J. Am. Chem. Soc. 1947. 69.17.

[27] Schultz H P, Topological organic chemistry. 1. graph theory and topological indices ofalkanes. J. Chem. Inf. Compur. Sci. 1989. 29, 227–228, 86, 87.

[28] Balaban A T, Highly discriminating distance-based topological index. Chem Phys Lett,1982.89, 399.

[29] Randić M, Characterization of molecular branching. J. Am. Chem. Soc. 1975.97,6609.

[30] Trevor W. H, Allan M. F, David B. T,Peter.W, EVA: A Novel Theoretical Descriptor for QSAR Studies. Kluwer Academic Publishers. Printed in Great Britian, H. Kubinyi et al. (eds.) 3D QSAR in Drug Design, 1998 Volume 2. 381-398

[31] Schuur J H, Selzer P, Gasteiger J, The coding of the three-dimensional structure of molecules by molecular transforms and its application to structure-spectra correlations and studies ofbiological activity. J. Chem. Inf. Compur. Sci. 1996. 36. 334.

[32] Witten I H, Frank E, Data Mining: Practical Machine Learning Tools and Techniques, Morgan Kaufmann Publishers, San Francisco, 2005.

[33] Roy K, On some aspects of validation of predictive QSAR models. Expert Opin Drug Discov. 2007 2. 1567.

[34] Hartigan J A, Wong M A, ―A K-means clustering algorithm,‖ Journal of the Royal Statistical Society. Series C. (Applied Statistics). 1979.28. 100.

[35] Kohonen T, Self-Organization and Associative Memory, Series in Information Sciences, vol.8, SpringerV erlag, Heidelberg. 1984

[36] Linusson A, Gottfries J, Lindgren F, Wold S, Statistical molecular design of building blocks for combinatorial chemistry, Journal of Medicinal Chemistry. 2000. 43. 1320

[37] Hudson B D, Hyde R M, Rahr E, Wood J, Parameter based methods for compound selection from chemical databases. Quant. Struct.–Activity Relationships 1996. 15. 285

[38] Szàntai-kis C, Kövesdi I, Kèri G, Orfi L: Validation subset selections for extrapolation oriented QSPAR models. Mol. Divers. 2003. 7. 37.

[39] Daszykowski M, WalczakB, MassartDL, () "Representative subset selection", Analytica Chimica Acta, 2002. 468. pp. 91-103

[40] Kennard R W, Stone LA, Computer aided design of experiments. Technometrics. 1969. 11.137.

[41] Dantas Filho H A, Harrop Galvao R K, Ugulino Araujo M C, Da Silva E C, Bezerra Saldanha T C, Jose G E, Pasquini C, Raimundo I M, Rodrigues Rohwedder J J, A strategy for selectingcalibration samples for multivariate modelling, Chemometr Intell Lab.2004 72.83.

[42] Wold, S. and Eriksson, L. (1995)- Chemometric Methods in Molecular Design, VCH Publisher, Weinheim.

[43] Snee, R. D. (1977). Validation of Regression Models: Methods and Examples. Technometrics, 19(4), 415 428.doi:10.1080/00401706.1977.10489581

[44] Gramatica P, Chirico N, Papa E, Kovarich S, Cassani S, QSARINS: A new software for the development, analysis, and validation of QSAR MLR models, J. Comput. Chem. 2013, 34,2121.

[45] Roy K, Kar S, Narayan Das R., Understanding the Basics of QSAR for Applications in Pharmaceutical Sciences and Risk Assessment, Chapter 6-Selected Statistical Methods in QSAR, Academic Press, Boston, 2015, 191

[46] Thése vp

[47] François Blayo., Michel Verleysen, "Les réseaux de neurones artificiels", (1996), 1re édition.

[48] Léon Personnaz., Isabelle Rivals, Réseaux de neurones formels pour la modélisation, la commande et la classification, (2003), CNRS Éditions,

[49] Ammar M Y, Mise en oeuvre de réseaux de neurones pour la modélisation de cinétiques réactionnelles en vue de la transposition Batch/Continu, Thèse de Doctorat, Institut Nationale Polytechnique de Toulouse, 2007.

[50] MATLAB. (2004)-Version 7.0.0. 1992 0 (Release 14). The Language of Technical Computing. The Math Works, Inc. May 6..

[51] IDIOU G. Régression et modélisation par les réseaux de neurones. Mémoire de Magister, Université Mentouri-Algérie, 2009.

[52] Bouarra N, Thèse de doctorat, Université Badji Mokhtar –Annaba. Etudes QSPR des propriétés contrôlant l'évolution de quelques HAP dans l'environnement. p208.

[53] OCDE, Principles for the Validation, for Regulatory Purposes, of (Quantitative) Structure-Activity Relationship Models, Organisation de Coopération et de DéveloppementEconomique, Paris, 2009.

[54] Famini G. R. Penski C. A. Wilson L. Y., J Phys Org Chem 1992. 5. 395

[55] Tropsha A. f. A. J. Mol. Graph. Model. 2002. 20. 269.

[56] Tropsha A, Gramatica P, Gombar V. J, The importance of being earnest: validation is the absolute essential for successful application and interpretation of QSPR models. Gombar, QSAR. Comb. Sci. 2003. 22. 69.

[57] Tropsha A, Gramatica P. Gombar V.K., The importance of being Earnest: Validation is the absolute essential for successful application and interpretation of QSPR models. QSAR Comb.Sci. 2003. 22. 69.

[58] Roy P.P, Paul S. Mitra I, Roy K. On Two Novel Parameters for Validation of Predictive QSAR models, Molecules. 2009. 14.1660.

[59] Shi L.M., Fang H, Tong W., J. Chem. Inf. Comput. Sci. 41, 2001, 186.

[60] Chirico, N.; Gramatica, P. Real 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 2011, 51, 2320.

[61] Gramatica P. Mol Inf 2014; 33: 311–314.

[62] Schüürmann, G.; Ebert, R.-U.; Chen, J.; Wang, B.; Kühne, R. 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 2008, 48, 2140–2145.

[63] Consonni, V.; Ballabio, D.; Todeschini, R. Comments on the Definition of the Q2 Parameter for QSAR Validation. Journal of Chemical Information and Modeling 2009, 49, 1669–1678.

[64] Consonni V, Ballabio D, Todeschini R. J Chemom 2010 ; 24: 194−201

[65] Lin L I Biometrics 1989; 45: 255–268.

[66] Chirico N, Gramatica P. J Chem Inf Model 2011; 51: 2320−2335

[67] Netzeva T. I, Worth A.P, Aldenberg T, Benigni R, Cronin M T D, Gramatica P, Jaworska J S, Kahn S, Klopman G, Marchant C A, Myatt G, Nikolova-Jeliazkova N, Patlewicz GY, Perkins R, Roberts D W, Schultz T W, Stanton D T, van de Sandt J J M, Tong W, Veith G, Yang C. ATLA Altern Lab Anim. 2005. 33.155.

[68] Chirico N, Gramatica P. J Chem Inf Model 2012; 52: 2044–2058

[69] Golbraikh, A.; Shen, M.; Xiao, Z. Y.; Xiao, Y. D.; Lee, K. H.; Tropsha, A. Rational Selection of Training and Test Sets for the Development of Validated QSAR Models. Journal of Computer-Aided Molecular Design 2003, 17, 241.

[70] Gramatica, P. Principles of QSAR Models Validation: Internal and External. Qsar & Combinatorial Science 2007, 26, 694.

[71] Lin I, A Concordance Correlation Coeffcient to Evaluate Reproducibility. Biometrics 1989, 45, 255.

[72] Bhhatarai, B.; Gramatica, P. Prediction of Aqueous Solubility, Vapor Pressure and Critical Micelle Concentration for Aquatic Partitioning of Perfluorinated Chemicals. Environ. Sci. Technol. 2011. 45. 8120.

[73] Roy, P. P.; Kovarich, S.; Gramatica, P. 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 2011. 32. 2386.

[74] Jaworska, J.; Nikolova-Jeliazkova, N.; Aldenberg, T. QSAR Applicabilty Domain Estimation by Projection of the Training Set Descriptor Space: a Review. Altern Lab Anim 2005, 33, 445.

[75] Netzeva, T. I.; Worth, A.; Aldenberg, T.; Benigni, R.; Cronin, M. T. D.; Gramatica, P.; Jaworska, J. S.; Kahn, S.; Klopman, G.; Marchant, C. A.; et al. Current Status of Methods for Defining the Applicability Domain of (quantitative) Structure-activity Relationships. The Report and Recommendations of ECVAM Workshop 52. Altern Lab Anim 2005, 33, 155.

[76] Stanforth, R. W.; Kolossov, E.; Mirkin, B. A Measure of Domain of Applicability for QSAR Modelling Based on IntelligentK-Means Clustering. QSAR & Combinatorial Science 2007, 26, 837–844.

[77] Weaver, S.; Gleeson, M. P. The Importance of the Domain of Applicability in QSAR Modeling. Journal of Molecular Graphics and Modelling 2008, 26, 1315.

[78] Baskin I. I, Kireeva N,Varnek A. The One-Class Classification Approach to Data Description and to Models Applicability Domain. Molecular Informatics 2010. 29. 581.

[79] Stanton D T, On the physical interpretation of QSAR models. J. Chem. Inf. Comput. Sci.,2003. 43, 1423

[80] Katritzky A R, Petrukhin R, Tatham D, Basak S, Benfenati E, Karelson M, Maran U, Interpretation of quantitative structure- property and- activity relationships. J. Chem. Inf. Model. 2001. 41. 679.

CHAPİTRE 5

Modélisation QSPR de la solubilité aqueuse et la température d'ébullition des phénols

İntroduction

1. Modélisation linéaire et non linéaire des relations quantitatives structure-solubilité aqueuse des phénols

2. Modélisation quantitative pour la prévision du point d'ébullition des phénols

Conclusion

Introduction

Suite à la méthodologie de la construction des modèles QSPR, en vue de prédire les propriétés étudiées dans cette thèse on a utilisé les méthodes et les logiciels suivants **:**

- **Représentation des molécules et optimisation des géométries**

Le logiciel CHEMDRAW a été utilisé pour la représentation des molécules, dont les structures ont été pré-optimisées à l'aide du champ de force de la mécanique moléculaire MM+ (algorithme de Polack-Ribière) du logiciel de modélisation moléculaire HYPERCHEM 6.03 [1]. La géométrie finale de la conformation d'énergie minimale a été obtenue par la méthode semi-empirique PM3 [2] avec un niveau Hartree-Fock restreint sans interaction de configuration, en appliquant un gradient limite standard de 0,001 Å kcal.mol-1 comme critère d'arrêt.

- **Génération des descripteurs**

Les géométries optimisées sont transférées dans le logiciel DRAGON version 5.3 [3] pour le calcul de plus de 1600 descripteurs appartenant à différentes classes. Afin de réduire les informations redondantes et non utiles, les valeurs des descripteurs constantes ou quasi constantes (écarts types inférieurs à 0,001) avec une corrélation supérieure à 0,95 ont été exclus lors d'une étape de pré-réduction.

- **Sélection des ensembles de calibrage et validation**

Plusieurs procédures pouvant être adoptées pour la sélection des ensembles d'apprentissage et d'essai, ce dernier devant contenir entre 15 et 40% des composés dans le jeu de données complet [4].

Nous avons appliqué différentes techniques de fractionnement : (a) choix aléatoire, (b) sélection des réponses triées, (c) similarité structurelle ordonnée par le premier axe de l'analyse en composantes principales (ACP) [5], (d) utilisation de l'algorithme de Kennard and Stone (CADEX) [6] et de l'algorithme DUPLEX [7].

- **Sélection des descripteurs et développement du modèle**

L'analyse de régression linéaire multiple et la sélection de variables ont été effectuées à l'aide du logiciel QSARINS [8] (version 2.2) en exploitant la méthode des moindres carrés ordinaires (OLS) associée à la méthode GA / VSS (sélection de sous-ensembles de variables par algorithmes génétiques). Cette procédure de « sélection de variables » génère une « population » de modèles, classés selon les valeurs décroissantes de R^2.

Les meilleurs modèles ont été choisis sur la base de Q^2_{LOO} comme valeur d'optimisation et en prenant en compte le principe de parcimonie concernant la complexité des modèles, qui devrait être petite que possible.

1. Modélisation linéaire et non linéaire des relations quantitatives structure-solubilité aqueuse des phénols

Les phénols et leurs dérivés sont des polluants omniprésents dans tous les écosystèmes en raison de l'activité des industries chimique, pétrolière, ou pharmaceutique. Ils pénètrent dans les écosystèmes à la suite du drainage des eaux usées municipales ou industrielles vers les eaux de surface [9].
Le transport et le devenir des phénols dans l'environnement dépendent en partie de leurs propriétés physicochimiques et de leurs distributions relatives entre les différents compartiments de l'environnement. Par conséquent, il est essentiel de prévoir leur comportement en étudiant leurs propriétés.

La solubilité aqueuse est l'une des propriétés physicochimiques majeures à optimiser dans les études pharmaceutiques et environnementales ; elle est liée à l'absorption et à la distribution des composés "ADME-Tox" (Absorption, Distribution, Métabolisme, Excrétion et Toxicité). La solubilité aqueuse est la concentration du produit chimique dans la phase aqueuse, lorsque la solution est en équilibre avec le composé pur dans sa phase habituelle (gazeuse, liquide ou solide) dans les conditions standards de température et de pression [10]. La détermination expérimentale de la solubilité du composé n'est pas facile à réaliser, ni même possible, lorsqu'on travaille avec de grandes série de composés chimiques.

À cette fin des relations quantitatives structure-solubilité (QSSR) sont utilisées dans ce travail, avec des techniques de régression linéaire multiple (MLR) et de réseau neuronal artificiel (RNA) pour la modélisation de la solubilité de 68 phénols sur la base de descripteurs moléculaires calculés à partir des structures 3D optimisées.

1.1. Origine des données

Les valeurs de log S ont été prélevées du « handbook of physical-chemical properties and environmental fate for organic chemicals » [11].
Les données expérimentales sont disponibles pour 68 phénols et leurs dérivés. Les valeurs de log S varient de 0,73 à 5,04.

Le tableau 1 reproduit les noms des composés avec leurs valeurs de solubilité expérimentales et calculés par les méthodes MLR et ANN.

Tableau 1 : Données expérimentales et calculées (logS) pour 68 phénols

Nom	log (S)			*Nom*	log(S)		
	Expérimenta	*MLR*	*ANN*		*Expérimentale*	*MLR*	*ANN*
Naphthalen-1-ol	2.6415	2.859	2.717	2,3,4,5,6-Pentachlorophénol	1.1461	1.595	1.336
2,3,4,5-Tetrachlorophénol	2.2201	2.007	2.240	4-Ethylphénol	3.9020	3.831	3.764
2,3,5,6-Tetrachlorophénol	2.0000	2.337	2.208	Phénol	4.9463	5.071	4.749
2,3,5-Trimethylphénol	2.9031	3.023	2.894	2,3,4,5-tetrachloro-6-methoxyphénol	1.4150	1.390	1.372
2,3,6-Trichlorophénol	2.6532	2.665	2.708	Naphthalen-2-ol	2.8692	2.863	2.743
2,4,6-Trimethylphénol	3.0792	3.094	2.990	2-Nitrophénol	3.0334	3.450	3.188
naphthalène;2,4,6-trinitrophénol	4.1383	3.323	3.494	3-Ethyl-5-methylphénol	3.3644	3.316	3.565
2,4-Dinitrophénol	2.5250	3.335	3.056	2-Phénylphénol	2.8451	2.474	2.915
2,6-Dichlorophénol	3.4191	3.157	3.116	3,4,5-trichloro-2-methoxyphénol	2.4914	1.982	2.080
2,6-Dimethylphénol	3.7945	3.904	4.044	3,4-Dichlorophénol	3.9664	3.564	3.713
2-Methoxyphénol	4.3945	4.185	4.233	3,5-Dichlorophénol	3.8689	3.976	4.098
5-Tert-butyl-2-methylphénol	2.6128	2.523	2.646	3,5-Dimethylphénol	3.7404	3.649	3.875
3-Nitrophénol	4.0626	3.801	3.805	3,5-Di-tert-butylphénol	1.1461	1.170	1.008
3-Tert-Butylphénol	3.3160	3.057	2.951	3-Methoxyphénol	4.8312	4.348	4.534
4,5-Dichloro-2-methoxyphénol	2.8500	2.746	2.576	2,3,5,6-Tetrachlorophénol	2.2625	2.073	2.249
2-Methyl-4,6-dinitrophénol	2.3464	2.188	2.375	2,3,5-Trichlorophénol	2.6990	2.800	2.786
4-Butylphénol	2.7903	2.927	2.967	2,3-Dichlorophénol	3.9146	2.969	3.199
4-Chloro-2-methoxyphénol	3.7300	3.894	4.092	2,3-Dimethylphénol	3.7782	3.719	3.954
4-Chlorophénol	4.4314	4.771	4.457	2,4,5-Trichlorophénol	2.9768	3.179	3.268
4-Hexylphénol	2.5922	2.044	2.130	2,4,6-Trichlorophénol	2.6375	3.031	2.937
4-Propan-2-ylphénol	3.5136	3.351	3.376	2,4-Dichlorophénol	3.6532	3.533	3.604
4-Methoxyphénol	4.2900	4.464	4.499	2,4-Dimethylphénol	3.9442	3.684	3.924
4-Nitrophénol	4.1303	3.948	4.191	2,5-Dimethylphénol	3.5019	3.776	3.985
4-Nonylphénol	0.7348	0.715	0.781	2-Chlorophénol	4.3918	4.110	4.272
4-Octylphénol	1.1004	1.163	1.243	2-Propan-2-ylphénol	3.6457	3.628	3.606
4-Phénylphénol	0.9912	2.151	1.509	2,3,4-Trichlorophénol	2.6990	2.458	2.580
4-Propylphénol	3.2375	3.357	3.371	2,3,4-Trichloro-6-methoxyphénol	1.7324	1.436	1.815
4-Butan-2-ylphénol	2.9823	3.009	2.951	4,5-Dichloro-2-methoxyphénol	2.7597	2.612	2.454
4-Tert-butylphénol	2.7634	2.974	2.956	5-Chloro-2-methoxyphénol	3.5977	3.553	3.796
Benzene-1,2-diol	4.6532	4.572	4.889	3-Methylphénol	4.3424	4.320	4.458
Benzene-1,4-diol	4.8451	5.090	4.504	Benzene-1,3-diol	5.0414	4.647	4.961
2-Methylphénol	4.4150	4.492	4.648	3,4,5-Trimethylphénol	3.1875	2.909	2.882
2-Ethylphénol	4.1474	4.069	4.260	3,4-Dimethylphénol	3.7076	3.563	3.632
4-methylphénol	4.3010	4.267	4.341	3-Chlorophénol	4.3424	4.267	4.345

1.2. Modèle MLR

Pour obtenir le meilleur modèle, quatre techniques de fractionnement différentes ont été appliquées (tableau 2). (48 composées pour l'ensemble de calibrage et 20 composés pour l'ensemble de validation)

Tableau 2. Paramètres statistiques des modèles obtenus basés sur différentes divisions

Divisions	$Q^2_{LOO}\%$	$R^2\%$	$Q^2_{LMO}\%$	$Q^2F3\%$
Aléatoire	88,66	90,82	88,04	88,81
Selon la réponse	90,32	91,97	89,63	87,16
Selon la similarité structurelle	90,10	90,01	89,01	79,08
CADEX	89,33	91,00	89,00	92,36

Les valeurs statistiques du tableau 2 montrent clairement que les modèles ont des performances similaires et une excellente capacité prédictive, vérifiées par divers critères dans les différentes divisions. Sur la base des résultats ci-dessus, nous avons choisi la sélection par l'algorithme Kennard and Stone.

L'équation (1) reproduit le modèle sélectionné :

$$\log S = 8{,}59 - 0{,}24\ \textbf{Polarizabilité-}\ 2{,}20\ \textbf{DISPe -}\ 0{,}29\ \textbf{nCb-} \qquad (1)$$

n_{tr}=48 R^2= 91,00 % Q^2_{LOO}= 89,33% Q^2_{LMO} = 89,00 %

Q^2_{F1}=87,24 %, Q^2_{F2}= 85,76 % Q^2_{F3} = 92,36 % CCC_{ext} =92,80 %

$RMSE_{tr}$= 0,32 $RMSE_{cv}$=0,35 $RMSE_{pr}$ = 0,30 S = 0,34

F = 148,23

Les paramètres statistiques montrent que le modèle d'équation (1) établit une forte corrélation entre les 3 variables sélectionnées et la propriété étudiée, caractérisée par une grande valeur du F de Fisher = 148,23. Ce qui indique la bonne capacité prédictive du modèle, et une faible valeur de l'erreur standard s = 0,34. En effet, la valeur du coefficient de détermination signifie que 91 % de la variabilité de log s, peut être expliquée par ces trois descripteurs, alors que la valeur élevée de Q^2_{LOO}, renseigne sur la robustesse du modèle. Tous les paramètres statistiques du modèle sélectionné

sont satisfaisants et prouvent que le modèle obtenu est stable, robuste et prédictif. Dans l'équation 1, trois types de descripteurs moléculaires apparaissent :

(1) La polarisabilité est définie comme le moment dipolaire d'une molécule induit par un champ électrique d'intensité unitaire [12] ;

(2) DISPe, valeur pondérée par les électronégativités atomiques de Sanderson, qui est le déplacement (DISPe) entre le centre géométrique et le centre du champ de la propriété considéré, calculé par rapport aux axes moléculaires principaux [13] ;

(3) nCb^{-}, le nombre du carbone dans le benzène substitué C (sp2) [13].

Certains paramètres statistiques importants (comme indiqué dans le tableau 3) ont été utilisés pour évaluer les descripteurs impliqués. La valeur t d'un descripteur mesure la signification statistique de son coefficient de régression. Les valeurs absolues élevées de t indiquent que les coefficients de régression des descripteurs impliqués dans le modèle MLR sont significativement plus grands que l'écart type. La probabilité T d'un descripteur permet de décrire la signification statistique lorsqu'il est combiné avec d'autres descripteurs dans un modèle QSPR collectif global (c'est-à-dire les interactions des descripteurs). Les descripteurs avec des valeurs de la probabilité t inférieures à 0,05 (confiance de 95%) sont généralement considérés comme statistiquement significatifs dans un modèle particulier, ce qui signifie que leurs influences sur la variable dépendante ne sont pas dues au hasard [14]

Tableau 3. Caractéristiques des descripteurs dans le modèle MLR optimal.

Descripteurs	***Type de descripteur***	***Coefficient***	***coeff SE***	***T***	***t-prob***	***VIF***
Constant		8.58640	0.2647	32.43	0.000	
Polarizabilité	Descripteur électronique	-0.24015	0.0138	-17.35	0.000	1.079
DISPe	Descripteur géometrique	-2.2032	0.3572	-6.17	0.000	1.056
nCb-	Compt de groupe fonctionel	-0.29028	0.0503	-5.77	0.000	1.135

La probabilité t la plus petite suggère le descripteur le plus significatif. Les valeurs des probabilités t des trois descripteurs sont nulles, ce qui indique qu'elles sont hautement significatives. Les valeurs du VIF de ces descripteurs (moins de cinq) suggèrent que ces descripteurs ne sont pas corrélés les uns aux autres. Ainsi, le modèle peut être considéré comme une équation de régression optimale.

1.3 Domaine d'application du modèle développé

L'analyse du domaine d'application du modèle est une autre étape de validation importante et nécessaire pour le criblage de nouveaux composés. Dans ce travail, le DA a été vérifié par la méthode du levier effectuée par le diagramme de Williams (Fig. 1) qui présente les résidus de prédiction standardisés en fonction des valeurs des leviers (h_i).

Tous les résidus sont situés dans la plage de ± 3 s (lignes horizontales), à l'exception d'un composé de calibrage : le 4-phénylphénol, ce composé est un point aberrant (son résidu standardisé est supérieur à 3).

Cependant, les valeurs des leviers (hi) des composés des ensembles de calibrage et de validation sont inférieures à la valeur critique de référence (h * = 0,25). Ce qui établit la fiabilité du modèle MLR développé.

Ainsi, il peut être utilisé pour prédire la solubilité de nouveaux phénols.

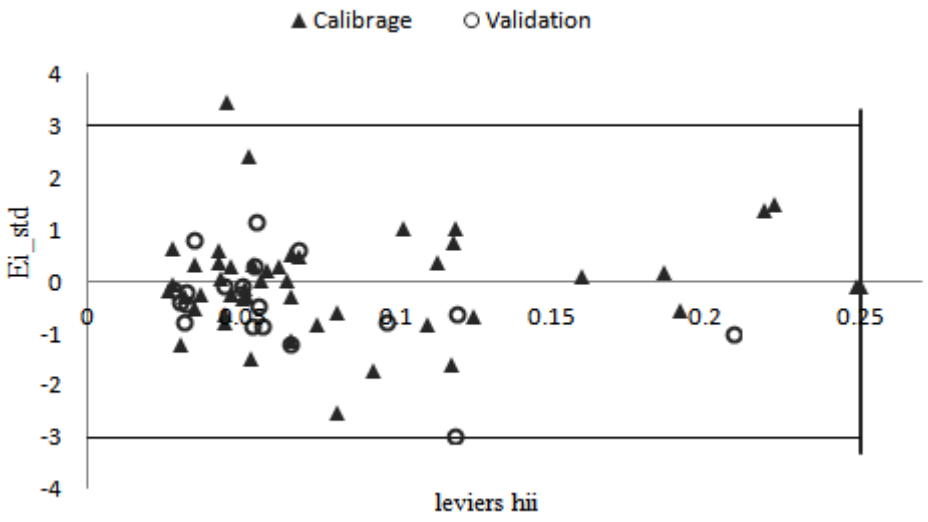


Figure 1 : Diagramme de Williams du modèle développé

Les valeurs prédites de la solubilité aqueuse ont été comparées à leurs valeurs expérimentales dans la figure 2 ; le bon accord entre ces valeurs (R^2 = 91,0%, R^2val = 88,4%) et la forte corrélation entre le log S observé et prédit pour l'ensemble de calibrage et de validation confirme la bonne qualité du modèle.

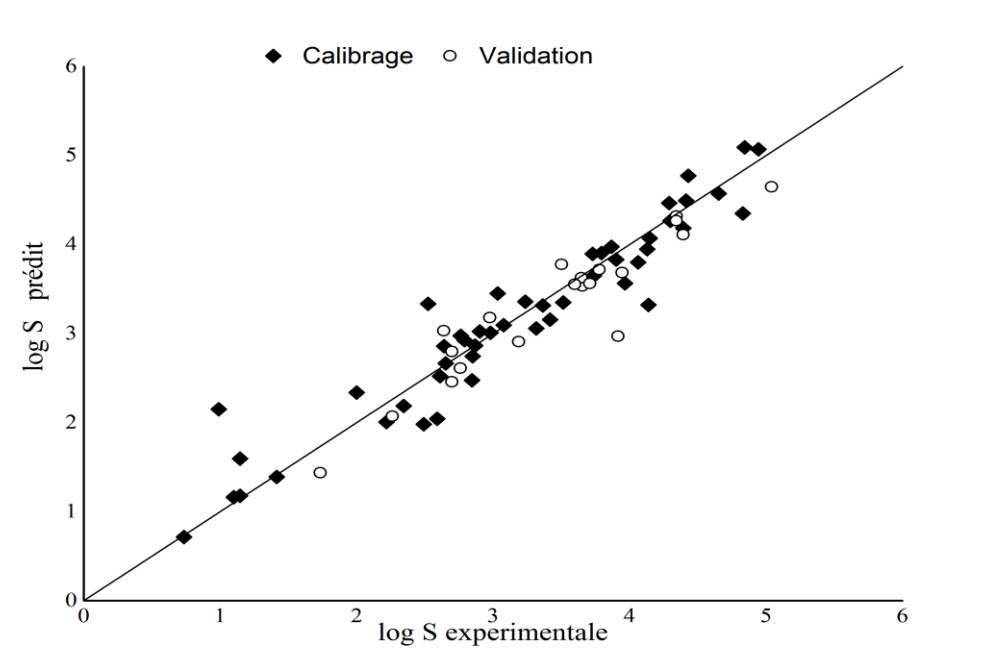


Figure2. Droite d'ajustement du modèle

1.4 Modèle RNA

La méthode des réseaux de neurones artificiels (RNA) est une technique de modélisation non linéaire importante, très utilisée pour les études QSPR ; afin de comparer la capacité prédictive des modèles MLR avec RNA ; l'ensemble de données a été modélisé par RNA en utilisant les descripteurs sélectionnés par le modèle MLR en tant que variables d'entrée pour un modèle RNA à trois couches avec un algorithme d'apprentissage par rétropropagation [15].

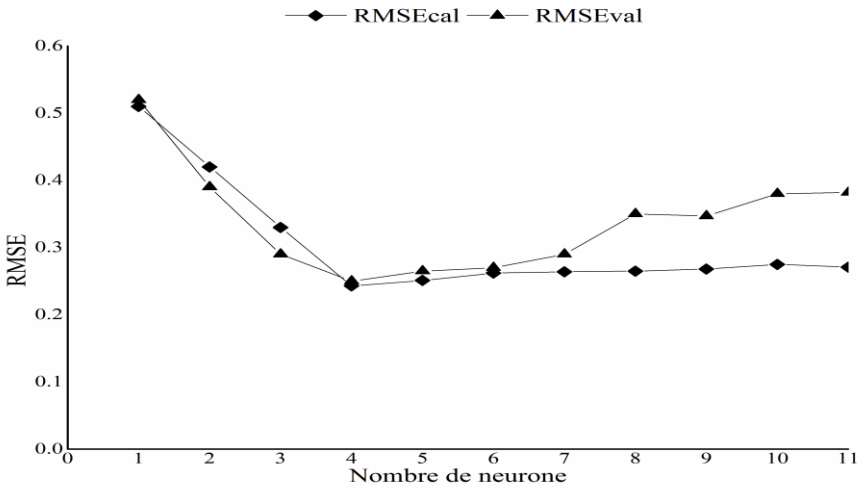
Pour l'optimisation du modèle, le nombre de neurones dans la couche cachée est 4 et le nombre d'itérations 20.

Tableau 4. Structure optimale adoptée pour le réseau de neurones.

Entrées	**3 descripteurs**
Sortie	1 (log (S))
Couche cachée	Une couche cachée
Nombre de neurones dans la couche cachée	4 neurones
Algorithme d'apprentissage	Algorithme Levenberg-Marquardt de rétropropagation

Plusieurs calculs ont été effectués afin d'obtenir les meilleurs résultats. Enfin, le modèle RNA développé qui a une bonne performance pour l'ensemble de calibrage et l'ensemble de validation a été sélectionné comme modèle de prédiction de la solubilité des phénols. L'ensemble de validation a été utilisé pour tester la capacité prédictive du modèle RNA.

La figure 3 présente les graphes des valeurs RMSE en fonction du nombre de neurones dans la couche cachée du modèle. Cette figure montre que les plus petites valeurs RMSE pour l'ensemble de calibrage, de validation et de test sont proches et plus petites lors de l'utilisation de quatre neurones dans la couche cachée.



.

Figure 3. Les valeurs RMSE en fonction du nombre de neurones dans la couche cachée.

Le nombre de neurones dans la couche cachée est un paramètre important qui influe sur les performances du modèle RNA. La règle habituelle est que les poids et le biais doivent être plus petits que les échantillons pour que le modèle obtenu par le réseau soit stable.

Pour notre travail, le nombre de neurones cachés ne devrait pas être supérieur à 10 avec 48 échantillons dans l'ensemble de calibrage.

Les meilleurs résultats sont obtenus en utilisant quatre neurones cachés après une optimisation de l'architecture du réseau en fonction du nombre de neurones cachés.

Ainsi, une architecture (3-4-1) a été obtenue, avec un R^2 de 94,992%, RMSEval = 0,250, RMSEtest = 0,224, RMSEtr = 0,243 et s = 0,245 pour l'ensemble de calibrage.

La qualité de l'ajustement a été vérifiée par la représentation des valeurs prédites de la solubilité en fonction des valeurs expérimentales. La figure 4 montre une faible dispersion des points autour de la première bissectrice, ce qui indique le bon accord entre ces valeurs.

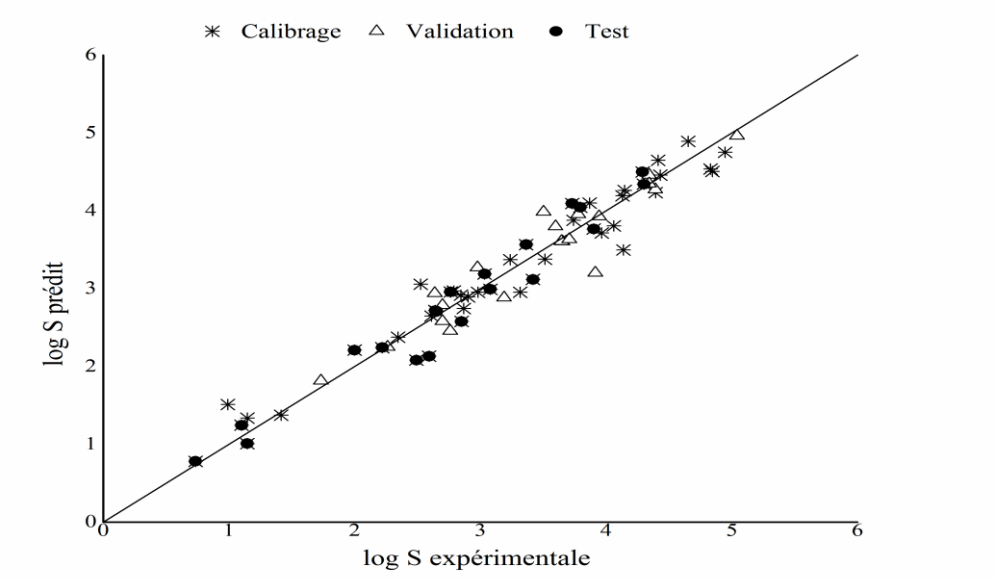


Figure 4. Les valeurs prédites du log s en fonction des valeurs expérimentales

1.5 Comparaison entre les modèles MLR et RNA

Une comparaison entre les paramètres statistiques obtenus par la méthode MLR et la méthode RNA a été effectuée (tableau 5 et figure 5).

Tableau 5. Les paramètres statistiques obtenus par la méthode MLR et ANN

Paramétres	**MLR**	**ANN**
Nombre de descripteurs	3	3
$RMSE_{tr}$	0.325	0.243
$RMSE_{val}$	0.3003	0.250
S	0.340	0.245
R^2, %	91.0	94.99
Q^2_{ext}, %	92.3	94.70

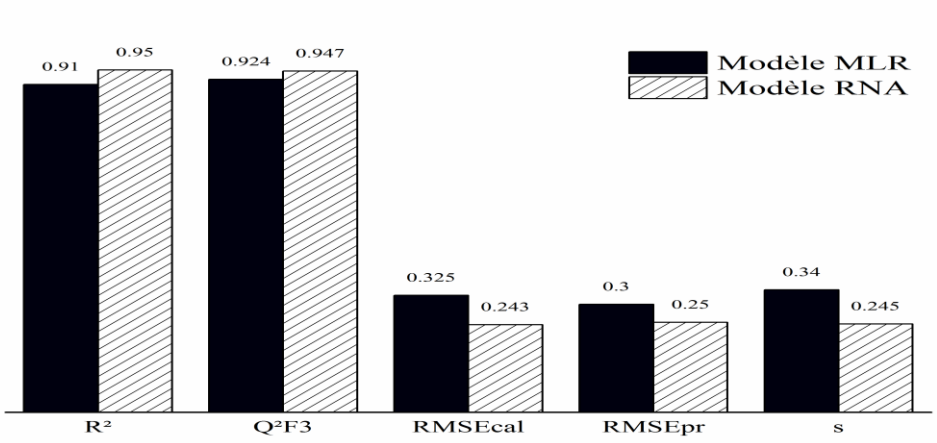


Figure 5. Les paramètres statistiques obtenus par la méthode MLR et RNA

La comparaison des modèles MLR et RNA donne l'avantage aux paramètres statistiques du modèle RNA, ainsi que les RMSE de l'ensemble de calibrage et l'ensemble des données du model RNA sont inférieurs à celle du modèle MLR, ce qui confirme la relation non linéaire entre l'information structurelle et les valeurs de la solubilité des composés.

1.6 Analyse et interprétation des contributions des descripteurs

L'interprétation des modèles non linéaires n'est pas une tâche aisée en raison de la complexité de leur procédure de modélisation, bien que les modèles non linéaires puissent donner de meilleurs résultats prédictifs. Ainsi, l'interprétation des descripteurs a été réalisée sur la base du modèle MLR. Selon une procédure décrite précédemment [16,17], les contributions relatives des trois descripteurs au modèle MLR ont été déterminées et reproduites à la figure 6.

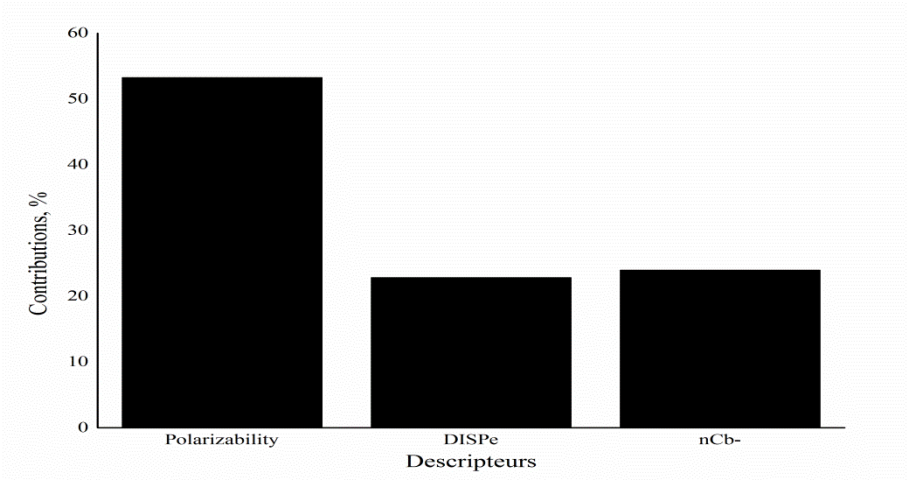


Figure 6. Les contributions relatives des trois descripteurs au modèle MLR

La contribution des descripteurs impliqués dans le modèle de RLM diminue dans l'ordre suivant : polarisabilité (53,23%) nCb- (23,95%)> DISPe (22,82%).

.

2. Modélisation pour la prévision quantitative du point d'ébullition des phénols

Le point d'ébullition ou la température d'ébullition ($T_{éb}$) est une propriété physique très importante, qui présente un intérêt pratique dans les domaines de la chimie, de la protection de l'environnement et de l'industrie pharmaceutique [18]. Elle se définit comme la température à laquelle la pression de vapeur d'un liquide saturé pur est de 760 mmHg [19]. C'est également un indicateur de l'état physique (par exemple, liquide ou gazeux) d'un composé organique. De plus, les points d'ébullition peuvent être utilisés pour prédire ou estimer d'autres propriétés physiques [20], telles que les températures critiques [21], les points d'éclair [22] et les enthalpies de vaporisation [23]. La température d'ébullition d'une molécule dépend de deux groupes principaux de facteurs. Le premier inclut les forces intermoléculaires, telles que les interactions dipôle-dipôle et de Coulomb. Le deuxième groupe explique la taille et la structure de la molécule dans son ensemble, c'est-à-dire comment l'énergie fournie par le chauffage est répartie en modes rotationnel et vibrationnel [24]. Cependant, les données de $T_{éb}$ ne sont souvent pas disponibles et doivent donc être estimées de manière théorique. Les méthodes d'estimation de $T_{éb}$ ont été largement explorées [24-26] en utilisant la topologie de la molécule et / ou des paramètres de chimie quantique calculés pour une structure optimisée de la molécule. L'absence de données complètes, fiables et comparables sur les points d'ébullition a conduit à la mise au point de différentes méthodes d'estimation des points d'ébullition. Avec l'avènement du calcul rapide et peu coûteux, il y a eu une croissance remarquable dans le domaine des relations quantitatives structure-propriété (QSPR), qui corrèlent les propriétés des polluants avec les descripteurs moléculaires pertinents [27].

2.1. Origine des données

Un ensemble de 56 phénols ont été pris en compte. Les données expérimentales de $T_{éb}$ (Tableau 8) ont été extraites de [11]. Les valeurs de $T_{éb}$ rapportées se distribuent entre 174,9 et 305 ° C. Ces composés ont été divisés en deux sous-ensembles en utilisant l'algorithme de Kennard et Stone (CADEX) [6], un ensemble de calibrage de 39 composés et un ensemble de 17 composés pour validation externe.

2.2 Développement du modèle

Un modèle à 4 variables a été retenu comme modèle optimal capable d'expliquer la variation de propriété parmi l'ensemble de calibrage.

L'équation de régression et les paramètres statistiques du modèle sont les suivants

$$\text{Téb} = 314 + 1663\,\boldsymbol{PW5} + 70.9\,\boldsymbol{Hy} - 2091\,\boldsymbol{X5A} + 247\,\boldsymbol{R6m}$$

n_{tr} = 39; S = 11,111; RMSE $_{tr}$ = 10,375; RMSE $_{ext}$ = 11,493; F = 60,455

CCC tr =93,43 ; CCC ext =89,25

Tableau 6. Résultats de l'évaluation du modèle développé.

R^2	Q^2_{LOO}	Q^2 $_{L30\%O}$	Q^2_{F3}	$(R^2-R^2_0)/R^2$	$(R^2-R'^2_0)/R^2$	K	K'
87,67%	84,11%	83,12%	84,87%	0,0059	0,0244	0,9907	1,0068

Le tableau 7 résume les caractéristiques des descripteurs sélectionnés

Tableau. 7 Caractéristiques des descripteurs sélectionnés du modèle.

Prédicteur	Coef	SE Coef	T	P	VIF
Constants	313,58	22,68	13,83	0,000	
PW5	1662,9	173,3	9,59	0,000	1,358
Hy	70,871	6,910	10,26	0,000	1,364
X5A	-2091,3	235,5	-8,88	0,000	1,222
R6m	247,16	57,83	4,27	0,000	1,471

D'après ces deux tableaux, nous pouvons conclure que les paramètres statistiques du modèle sont satisfaisants et que le modèle MLR est stable, robuste, prédictif et satisfont aux conditions d'acceptation « paramètres » de Tropsha.

Les descripteurs du modèle sélectionné et leurs valeurs respectives sont reproduits dans les tableaux 8 l'ensemble de calibrage et 9 l'ensemble de prédiction ;

Tableau 8. Les descripteurs moléculaires et les valeurs de $T_{éb}$ pour les phénols dans l'ensemble de calibrage

N°	Nom	$T_{éb}$ Exp.	PW5	Hy	X5A	R6m	$T_{éb}$ Pred.
1	2-Méthyllphénol	191.04	0.057	-0.16	0.102	0.001	184.1044
2	4-Méthyllphénol	201.98	0.08	-0.16	0.109	0.016	211.4188
3	2,6-Diméthyllphénol	201.07	0.054	-0.21	0.089	0.005	203.3931
4	3,4-Diméthyllphénol	227	0.071	-0.21	0.095	0.015	221.5858
5	2,4,6-Triméthyllphénol	220	0.065	-0.26	0.082	0.018	236.4183
6	2-Propylphénol	220	0.082	-0.26	0.104	0.048	226.0939
7	4-Propylphénol	232.6	0.085	-0.26	0.1	0.051	240.1891
8	2-Isopropylpheonol	213.5	0.062	-0.26	0.096	0.022	203.1405
9	4-Isopropylphénol	230	0.072	-0.26	0.091	0.073	242.8309
10	4-Butylphénol	248	0.088	-0.29	0.1	0.068	246.7572
11	2-sec-Butylphénol	228	0.073	-0.29	0.096	0.068	230.1793
12	2-tert-Butylphénol	223	0.059	-0.29	0.091	0.029	207.7162
13	3-tert-Butylphénol	240	0.076	-0.29	0.093	0.058	238.9701
14	4-sec-Butylphénol	241	0.074	-0.29	0.09	0.092	250.3218
15	4-tert-Butylphénol	237	0.066	-0.29	0.084	0.095	250.308
16	4-tert-Octylphénol	279	0.076	-0.4	0.085	0.161	274.0002
17	1-Naphthol	288	0.094	-0.29	0.084	0.013	276.6008
18	2-Naphthol	285	0.101	-0.29	0.084	0.036	293.9256
19	2-Phénylphénol	286	0.095	-0.35	0.087	0.099	289.1354
20	4-Phénylphénol	305	0.095	-0.35	0.087	0.093	287.6524
21	2-Allylphénol	220	0.082	-0.26	0.104	0.037	223.3751
22	4-Chlorophénol	220	0.08	-0.04	0.109	0.01	218.3695
23	2,3-Dichlorophénol	206	0.054	-0	0.089	0	217.182
24	3,4,5-Trichlorophénol	275	0.065	0.031	0.082	0.01	254.8519
25	4-Nitrophénol	279	0.072	0.031	0.091	0.103	270.6567
26	1,2-Dihydroxybenzéne	245	0.057	0.846	0.102	0.001	255.2589
27	1,3-Dihydroxybenzéne	276.5	0.06	0.846	0.097	0.001	270.7039
28	Hydroquinone	285	0.08	0.846	0.109	0.007	280.3489
29	2-Méthoxyphénol	205	0.063	-0.11	0.101	0.015	203.2476
30	3-Méthoxyphénol	244	0.074	-0.11	0.102	0.028	222.6609
31	2,6-Dimethoxyphénol	261	0.075	-0.12	0.09	0.051	254.2534
32	3-Méthyllphénol	202.27	0.06	-0.16	0.097	0.002	199.7965
33	3,5-DiMéthylphénol	221.74	0.058	-0.21	0.082	0.007	225.1778
34	2,3,5-TriMéthylphénol	233	0.065	-0.26	0.082	0.016	235.924
35	3,4,5-TriMéthylphénol	248.5	0.065	-0.26	0.082	0.016	235.924
36	o-Ethylphénol	204.5	0.063	-0.21	0.101	0.013	195.2409
37	p-Ethylphénol	217.9	0.078	-0.21	0.098	0.04	233.1311
38	2,5-Dichlorophénol	211	0.071	-0	0.095	0	232.903
39	2,4,6-Trichlorophénol	246	0.065	0.031	0.082	0.01	254.8519

Tableau 9. Descripteurs moléculaires et valeurs de $T_{éb}$ des phénols dans l'ensemble de prédiction.

No	Nom	$T_{éb}$ Exp.	PW5	Hy	X5A	R6m	$T_{éb}$ Pred.
40	m-Ethylphénol	218.4	0.074	-0.213	0.102	0.023	213.9128
41	Phénol	181.87	0.062	-0.088	0.113	0	174.1285
42	2-Chlorophénol	174.9	0.057	-0.039	0.102	0	192.2909
43	3,4-Dichlorophénol	253	0.071	-0.001	0.095	0.01	235.3746
44	4-Chloro-m-cresol	235	0.071	-0.107	0.095	0.011	228.1095
45	2-Nitrophénol	216	0.062	0.031	0.096	0.008	220.0911
46	4-Méthoxyphénol	243	0.078	-0.107	0.098	0.036	239.6548
47	4-Hydroxy-3-méthoxybenzaldehyde	285	0.086	-0.119	0.09	0.088	281.6899
48	2,3-Diméthyllphénol	216.9	0.054	-0.213	0.089	0.003	202.8988
49	2,4-Diméthyllphénol	210.98	0.071	-0.213	0.095	0.016	221.833
50	2,5-Diméthyllphénol	211.1	0.071	-0.213	0.095	0.014	221.3386
51	2,4,5-Triméthyllphénol	232	0.076	-0.257	0.088	0.025	243.8923
52	3-Chlorophénol	214	0.06	-0.039	0.097	0	207.7358
53	2,4-Dichlorophénol	210	0.071	-0.001	0.095	0.01	235.3746
54	2,6-Dichlorophénol	220	0.054	-0.001	0.089	0	217.182
55	3,5-Dichlorophénol	233	0.058	-0.001	0.082	0	238.4723
56	2,4,5-Trichlorophénol	247	0.076	0.031	0.088	0.01	260.5958

Le diagramme de dispersion (la droite d'ajustement) de la température d'ébullition prédite par rapport au point d'ébullition expérimental est présenté à la figure 7, qui montre la courbe du modèle proposé avec des performances très satisfaisantes, illustrant également un bon étalonnage du modèle.

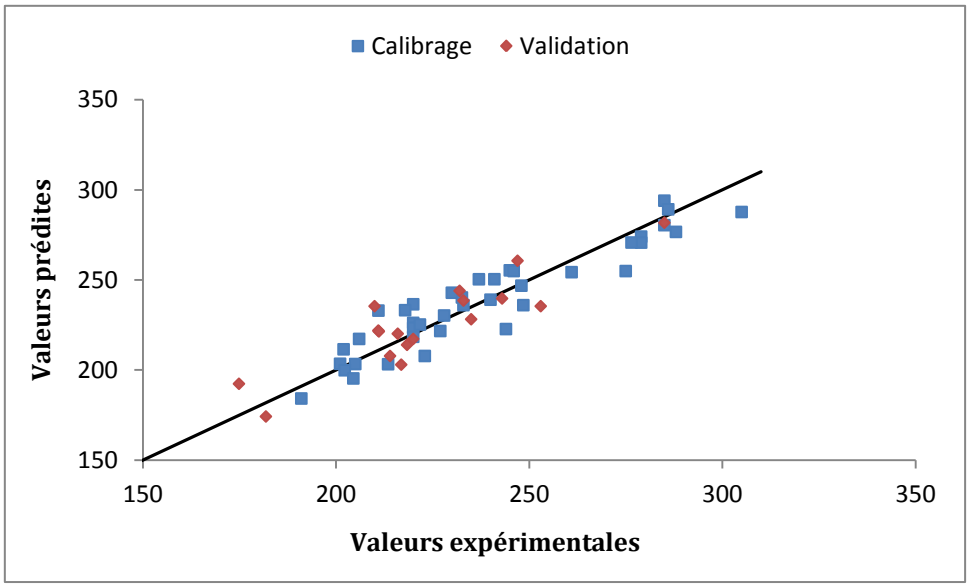


Figure 7. Diagramme de dispersion des valeurs prédites de $T_{éb}$ en fonction des valeurs expérimentales

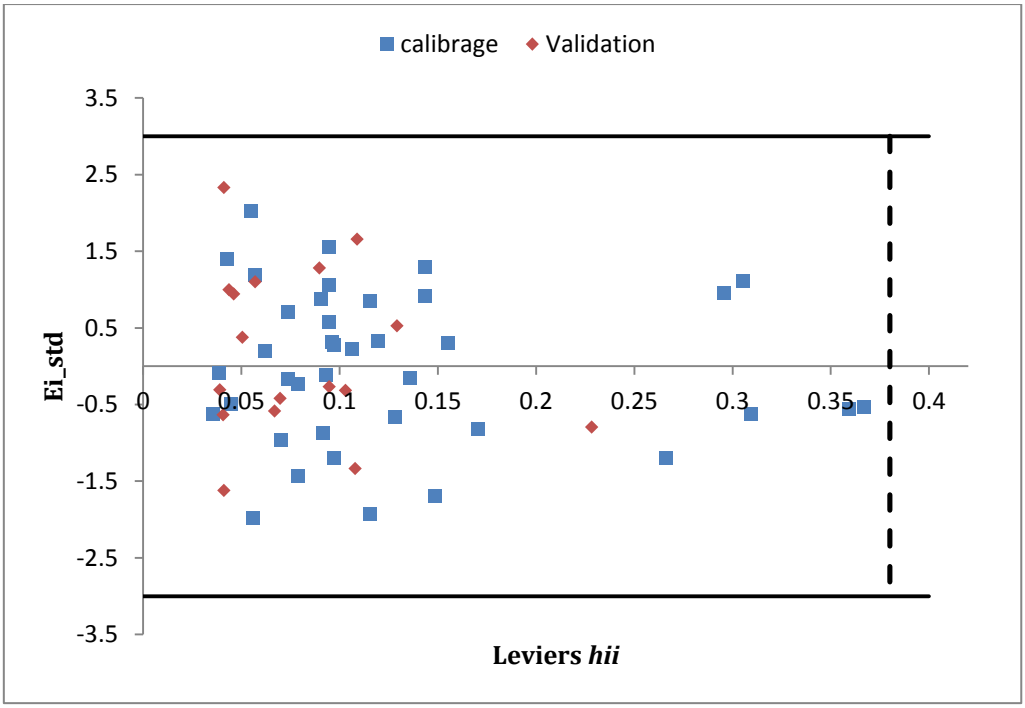


Figure 8. Diagramme de Williams

La figure 8 représente les erreurs standardisées de prédiction en fonction des valeurs des leviers (*hii*). On constate que tous les résidus se situent dans la plage (± 3 SD) (lignes horizontales), Nous notons l'absence de point aberrant et /ou influent pour l'ensemble de calibrage et de prédiction (validation externe), *h** étant égal à 0,385, ce qui signifie que le modèle a une bonne capacité prédictive.

Ainsi, Le modèle proposé pourrait être utilisé pour cribler des bases des données existantes ou des structures chimiques virtuelles afin d'identifier le point d'ébullition d'autres phénols, le domaine d'application servira, alors, comme un outil précieux pour filtrer les structures chimiques « dissemblables ».

Ensuite, le modèle a été validé en appliquant le test de randomisation (figure 9) qui compare les résultats obtenus pour les modèles randomisés au modèle réel de départ. Il est clair que les statistiques R^2_{Yscr} (= 0.1107) et Q^2_{Yscr} (= -0.1905) sont plus petites que celles du modèle QSPR réel. Ce qui permet de s'assurer que le modèle établi a une base réelle et qu'il n'est pas fortuit.

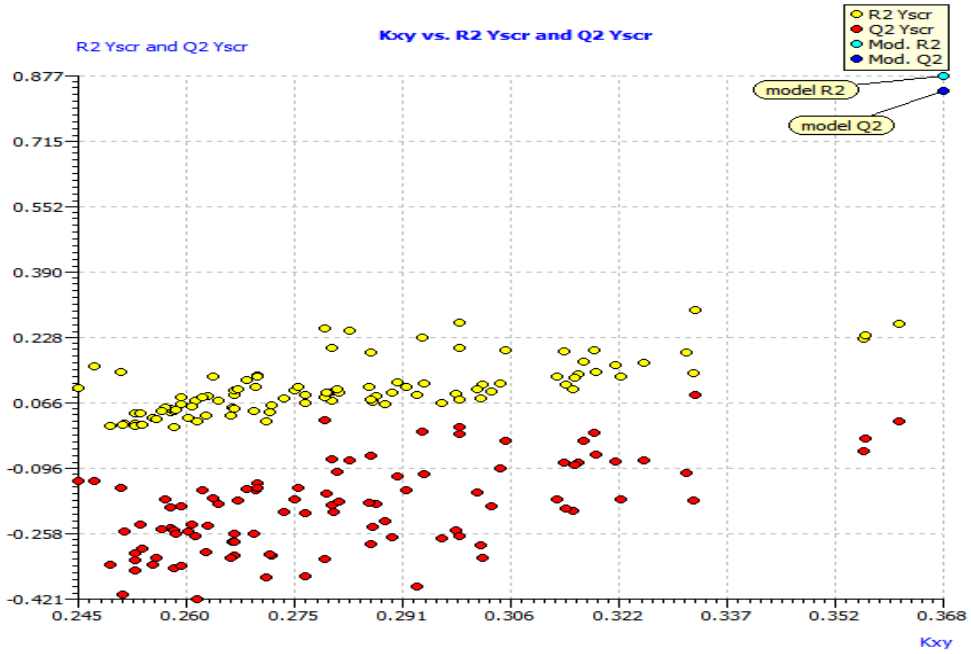


Figure 9. Test de randomisation

2.3 Contribution des descripteurs et interprétation

Les contributions relatives des quatre descripteurs au modèle MLR sont illustrées à la figure 10. L'importance des descripteurs impliqués dans le modèle diminue dans l'ordre suivant : Hy (28,155%)> PW5 (27,476%)> X5A (26,844%)> R6m (17,524%).

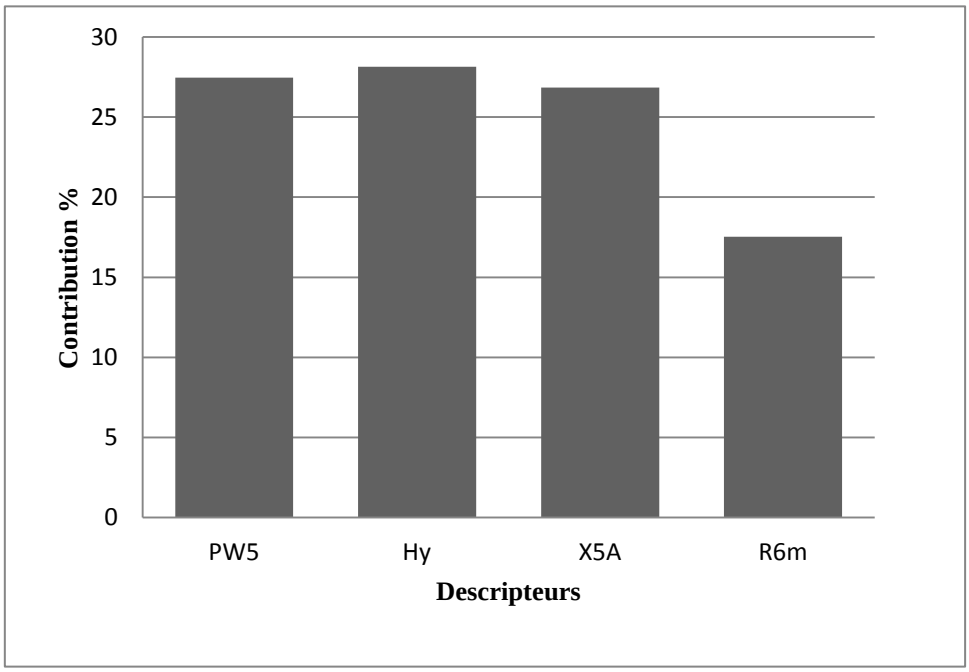


Figure 10. Contributions relatives des descripteurs au modèle MLR.

Le descripteur le plus significatif est le facteur hydrophilie (Hy), qui donne une corrélation avec $T_{éb}$ de 0,196 et explique 28,155% des contributions. Le facteur hydrophile est un simple indice empirique lié aux hydrophobicités des composés basé sur les descripteurs de comptage [28]. Il s'agit d'une mesure du nombre de groupes fonctionnels hydrophiles (-OH, -SH, -NH) [28]. Le descripteur de propriété moléculaire numérise ici les propriétés hydrophiles des molécules provoquées par le groupe «OH» [25].

Le deuxième descripteur significatif est PW5, ce qui donne une corrélation avec Bp de 0,566 et explique 27,47% des contributions. PW5 est un indice topologique qui considère la forme des molécules comme des propriétés moléculaires dans les différents composés [29]. La forme des molécules avec un type de ramification spécifique a également été sélectionnée comme descripteur significatif dans le modèle QSAR développé par Mitra et al., [30]. La variation des caractéristiques structurelles de ramification a également été prise en compte par Ray et al. [31] pour développer le modèle QSAR spécifique. PW5 fait référence aux proportions de chemin de longueur 5 par rapport à l'indice de forme moléculaire Randic [32]. Caractérise l'indice de forme d'un graphe moléculaire en considérant à la fois les chemins et les allées de longueur différente dans un graphe, puis en faisant en sorte que les proportions du nombre de trajets et du nombre de marches soient de la même longueur.

Le troisième descripteur X5A, appartient aux descripteurs d'indices de connectivité ; X5A est l'indice de connectivité moyen d'ordre cinq montre principalement les caractéristiques topologiques. Les indices topologiques sont des quantificateurs numériques de la topologie moléculaire et le graphe moléculaire dépourvu des atomes H. Ils impliquent une ou plusieurs caractéristiques structurelles de la molécule, telles que la taille, la forme, la symétrie et la ramification, et peuvent également codifier des informations chimiques sur le type d'atome et la multiplicité des liaisons [33].

Le dernier descripteur R6m appartient aux descripteurs GETAWAY, R6m est l'autocorrélation à effet de levier de distance topologique 6 pondérée par la masse atomique. R6m est un descripteur géométrique codant des informations sur la position efficace des substituants et des fragments dans l'espace moléculaire [34,35].

Conclusion

Dans cette étude, des modèles QSPR linéaires et non linéaires ont été développés pour prédire la solubilité (log S) d'un ensemble de données de 68 phénols utilisant MLR et RNA. La procédure de validation détaillée a démontré la robustesse des modèles proposés ainsi que la distinction du modèle non linéaire qui a donné des meilleurs résultats capables de prédire la solubilité dans l'eau des phénols avec plus de précision que l'approche MLR.

En outre la régression linéaire multiple été a utilisée pour construire des modèles QSPR au point d'ébullition d'un ensemble de 56 phénols. Les résultats de cette étude indiquent que l'utilisation des descripteurs PW5, Hy, X5A et R6m fournit une bonne estimation des points d'ébullition.

Références

[1]. HyperChem 6.03 Package. Hypercube, Inc., Gainesville, Florida, USA, 1999; software available at: http://www.hyper.com

[2]. Dewar M J S, Zoebisch E G, Ealy E F, Stewart J J P, (1985)"AMI: A New General Purpose Quantum Mechanical Model", J Am, Chem, Soc, Vol, 107, pp, 3902-3909.

[3]. Todeschini R, Consonni V, Pavan, M (2006) DRAGON Software for the Calculation of Molecular Descriptors, Release 5.3 for windows, Milano

[4]. Benfenati E, Chrétien J. R, Gini G, Piclin N. Pintore M., Roncaglioni A. Validation of the models. In Quantitative Structure-Activity Relationships (QSAR) for Pesticide Regulatory Purposes, Elsevier. 2007 185.

[5]. Jackson J. E. A User's Guide to Principal Component, Wiley, New York, United States,1991.

[6]. Kennard R. Stone L. A. Technometrics. 1969 11. 137

[7]. Daszykowski M, WalczakB, MassartDL, Representative subset selection. Analytica Chimica Acta. 2002. 468, pp, 91.

[8]. Gramatica P, Chirico N, Papa E, Cassani S, Kovarich S. QSARINS, Software for the Development and validation of QSAR MLR Models, available on request in (http://www.qsar.it)

[9]. Micha Lowicz J, Ożadowicz Wirgiliusz Duda R., Water. Air. Soil. Poll. 2005.16.205.

[10]. Palmer S. D, O'Boyle N. M., Glen R. C. J, Mitchell B. O. J. Chem. Inf. Model. 2007. 47.150.

[11]. Mackay D, Shiu W. Y, Ma K. C, Lee S. C. Handbook of physical-chemical properties and environmental fate for organic chemicals Second edition, CRC Press Inc, Boca Raton, USA, 2006.

[12]. Famini G. R., Penski C. A., Wilson L. Y. J. Phys. Org. Chem. 1992 .5.395.

[13]. Talete Srl. Dragon for windows (Software for Molecular Descriptor Calculation) Version 5.5 Milano, Italy, 2007; software available at

[14]. Xu H, Liu W, Li H, Zou W, Xu, Macromol. Theory. Simul. 2008.17.470.

[15]. Rétrop

[16]. Zheng, F., Bayram, E., Sumithran, S.P., Ayers, J.T., Zhan, C.G., Schmitt, J.D., Dwoskin, L.P. and Crooks, P.A. QSAR modeling of mono- and bis-quaternary ammonium salts that act as antagonists at neuronal nicotinic acetylcholine receptors mediating dopamine release, Bioorganic & Medicinal Chemistry, 2006.14. 3017.

[17]. Guha, R. and Jurs, P.C. Interpreting computational neural network QSAR models: a measure of descriptor importance, J Chemical Information and Modeling, 2005. 45. 800.

[18]. Todeschini R., Gramatica P., Provenazi R. & Marengo E.,. Weighted holistic invariant molecular descriptors. Part2. Theory of development and applications on modeling physicochemical properties of polyaromatic hydrocarbons, Chemometrics and Intelligent Laboratory Systems. 1995. 27. 221.

[19]. Gharagheizi F., Mirkhani S. A., Ilani-Kashkouli P., Mohammadi A. H., Ramjugernath D. & Richon D.,. Determination of the normal boiling point of chemical compounds using a quantitative structure–property relationship strategy:Application to a very large dataset, Fluid Phase Equilibria. 2013. 354. 250.

[20]. Cao D. Liang Y., Xu Q.,Yun, Y. & Li, H.,. Toward better QSAR/QSPR modeling: simultaneous outlier detection and variable selection using distribution of model features. Journal of Computer-Aided Molecular Design. 2011. 25. 67.

[21]. Katritzky A. R., Mu L. & Lobanov V. S., Correlation of Boiling Points with Molecular Structure. 1. A Training Set of 298 Diverse Organics and a Test Set of 9 Simple Inorganics, Journal of Physical Chemistry. 1996. 100 10400.

[22]. Yi-min D, Zhi-ping Z, Zhong C,Yue-fei Z, Ju-lan Z, Xun L.. Prediction of boiling points of organic compounds by QSPR tools, Journal of Molecular Graphics & Modelling. 2013. 44.113.

[23]. White C.M., 1986. Prediction of the boiling point, heat of vaporization, and vapor pressure at various temperatures for polycyclic aromatic hydrocarbons, Journal of Chemical And Engineering Data. Vol 31 (2), pp. 198–203

[24]. Smeeks, F.C. and Jurs, P.C. Prediction of boiling points of alcohols from molecular structure. Anal. Chim. Acta. 233, 111-119 (1990).

[25]. Katritzky, A.R.; Mu, L.; Lobanov, V. S.; Karelson, M. Correlation of Boiling Points with Molecular Structure. 1. A Training Set of 298 Diverse Organics and a Test Set of 9 Simple Inorganics. J. Phys. Chem. 100, 10400-10407 (1996).

[26]. Chunhui Lu; Weimin Guo; Yang Wang; Chunsheng Yin. Novel distance-based atom-type topological indices DAI for QSPR/QSAR studies of alcohols. J. Mol. Model. 12, 749–756 (2006).

[27]. Cronin MTD, Livingstone DJ, Predicting Chemical Toxicity and Fate, CRC PressLLC, Baca Raton, FL, 2004

[28]. Todeschini, R.; Consonni, V. Molecular Descriptors for Chemoinformatics; Wiley-VCH: Weinheim, Germany, 2009

[29]. Randic M., Razinger M., 1995, Molecular Topographic Indices, Journal of Chemical Information and Modelling 35 (1), 140-147.

[30]. Mitra I., Saha A., Roy K., , Exploring quantitative structure–activity relationship studies of antioxidant phenolic compounds obtained from traditional Chinese medicinal plants, Molecular Simulation 2010.36 . 1067.

[31]. Ray S., Sengupta C., Roy K. QSAR modeling of antiradical and antioxidant activities of flavonoids using electrotopological state (E-state) atom parameters, Central European Journal of Chemistry.2007. 5 . 1094.

[32]. Randic M., 2001, Novel Shape Descriptors for molecular Graphs, Journal of Chemical Information and Modelling. 2007.41. 607.

[33]. Abbasi M, Sadeghi-Aliabadi H, Amanlou M. Prediction of new Hsp90 inhibitors based on 3,4-isoxazolediamide scaffold using QSAR study, molecular docking and molecular dynamic simulation. DARU Journal of Pharmaceutical Sciences. 2017.25.17

[34]. Consonni V, Todeschini R, Pavan M. J. Chem. Inf. Comput. Sci. 2002. 42. 682.

[35]. Consonni V, Todeschini R, Pavan M, Gramatica P. J. Chem. Inf. Comput. Sci. 2002, 42, 693.

CHAPITRE 6

Modèle QSRR pour la prédiction des indices de rétention des pyrazines

Introduction

1. Données expérimentales

2. Développement du modèle

3. Validation du modèle

Conclusion

Introduction

L'identification des pyrazines se fait généralement par chromatographie gazeuse (CG) en comparant leurs pics à ceux obtenus pour les standards des composés suspectés.

La Relation Quantitative Structure/Activité (QSAR) initiée par Hansch et Fujita [1], a trouvé de nombreuses applications en chimie, en particulier dans la prédiction de la rétention chromatographique [2-5].

Mihra et Enomoto (1985) [6], ont décrit une relation structure/rétention pour un ensemble de pyrazines substituées pour lesquelles les incréments d'indices relatifs à différents substituants sur le cycle ont été déterminés pour une petite série de substituants présents.

La méthode fut ensuite étendue pour intégrer d'autres substituants, et ajouter un terme qui tient compte de la position sur le cycle d'un substituant par rapport aux autres [7]. Dans une approche analogue, Masuda et Mihara [8] décrivent l'utilisation d'indices de connectivité modifiés pour calculer à l'avance les indices de rétention d'une série de pyrazines substituées. Les méthodes conduisent à de bons résultats, pour autant que les incréments d'indices déterminés expérimentalement soient disponibles pour les composés inconnus impliqués, ce qui constitue leur défaut principal.

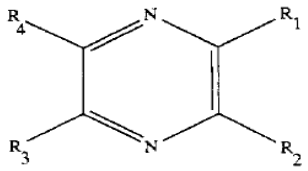
Stanton et Jurs (1989) [9], ont utilisé la méthodologie QSRR pour développer des modèles reliant les caractéristiques structurales de 107 pyrazines diversement substituées, à leurs indices de rétention obtenus sur deux colonnes de polarités très différentes (OV-101 et Carbowax-20M). Les équations ont été calculées à l'aide de la régression multilinéaire, le choix des variables explicatives (topologiques, électroniques et propriétés physiques) étant réalisé par élimination progressive **[10],** parmi les 85 descripteurs moléculaires individuels obtenus pour chaque molécule entière. Les indices de rétention (Ir) obtenus sur chaque colonne ont été traités séparément, en puisant dans les mêmes ensembles de descripteurs. Les modèles calculés avec 6 variables explicatives fournissent des erreurs standards élevées (S = 23 unités d'indice (u.i)- sur OV-101 et S= 36.33(u.i). Sur Carbowax -20M) qui ne présagent pas de bonnes capacités prédictives pour ces modèles, et qui laissent supposer des relations non linéaires entre descripteurs et propriété (Ir) étudiée [11].

1. Données expérimentales :

L'approche hybride algorithme génétique/ régression multilinéaire (AG/RML) a été utilisée pour modéliser les indices de rétention de 113 pyrazines rapportés dans [9], les descripteurs moléculaires étant uniquement calculés à partir de la structure chimique des composés.

Les pyrazines ont été séparées par chromatographie gazeuse à température programmée tour à tour sur les colonnes capillaire OV-101 et Carbowax-20M.

Les composés impliqués dans cette étude, présentent la structure générale suivante :



R_1 : H, alkyl, alkoxy, alkylthio, aryloxy, arylthio, acetyl, chloro.

R_2: H, alkyl, chloro.

R_3: H, alkyl.

R_4: H, alkyl.

2. Développement du modèle

L'ensemble des données obtenues sur chaque colonne a été éclaté, séparément, en un ensemble de calibrage (90 composés) et un ensemble de validation (23composés) à l'aide de l'algorithme DUPLEX. En opérant sur les pyrazines d'essai, des sous-ensembles de descripteurs ont été sélectionnés par algorithme génétique, en utilisant le logiciel MobyDigs [12] et en maximisant le coefficient de prédiction Q^2.

L'optimisation par algorithme génétique est appliquée pour les dimensions de 1 à 9. Nous avons fixé la taille adéquate c'est-à-dire le nombre optimal du modèle, en étudiant l'influence du nombre de descripteurs sur les valeurs de Q^2, R^2 pour chacune des deux colonnes. L'objectif étant de construire, sur un nombre minimal de descripteurs significatifs, des modèles conduisant à de faibles erreurs, et de meilleures qualités de prédiction et de détermination ; il est important de réduire le pool de descripteurs disponibles, de façon à ne retenir que ceux encodant l'information la plus riche.

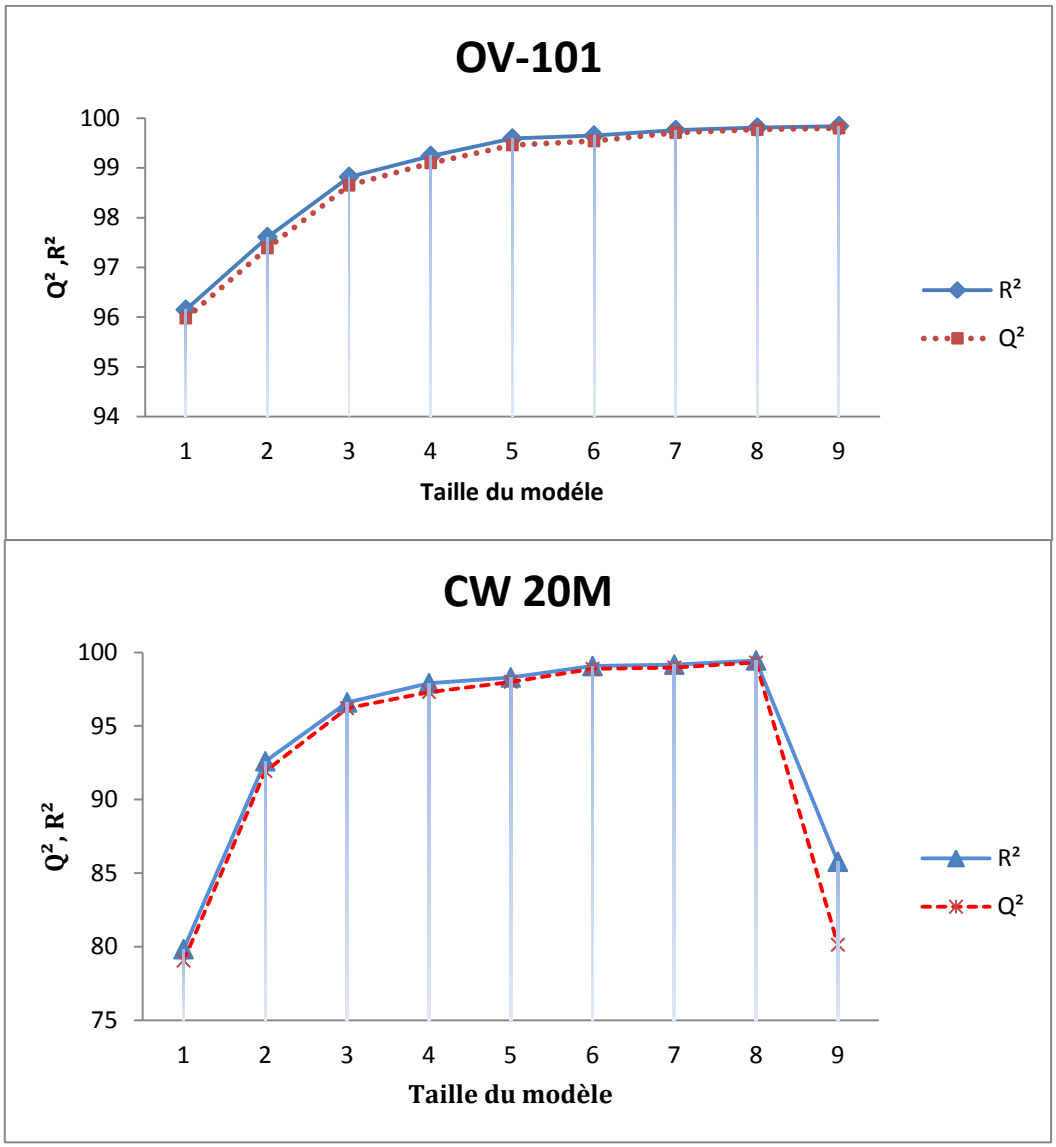


Figure 1 Variation de R² et Q² en fonction de la taille du modèle pour les deux colonnes

Manifestement les valeurs de Q^2 et R² augmente graduellement avec le nombre de descripteurs. Lorsque l'ajout d'un descripteur (augmentation de la dimension du modèle d'une unité) n'améliore pas de manière significative les statistiques d'un modèle, on a atteint la taille optimale du sous-ensemble, ce qui nous a conduit aux choix des tailles 6 et 5 respectivement pour les colonnes carbowax 20M et OV-101.

Après le choix de la taille du modèle, on a procédé à la recherche des descripteurs qui expliquent au mieux la variation de la propriété à modéliser ou variable à expliquer.

2.1. Equations et analyse de régression

Nous avons choisi pour chaque colonne un modèle avec les descripteurs les plus significatifs, dont les équations sont :

- Le modèle obtenu pour la CW a pour équation :

$$IR_{cw} = -50,68(\pm 29,75) + 40,09(\pm 1,20)\,Polarisabilité + 344,79(\pm 16,40)\,nCIR + 1778,80(\pm 298,60)\,X4Av + 132,48(\pm 13,01)\,Mor06v - 47,95(\pm 7,10)\,nCt + 5364(\pm 271,20)\,Qmean$$

- Le modèle obtenu pour la colonne OV-101 a pour équation :

$$IR_{ov-101} = 644(\pm 23,38) + 6,48(\pm 0.09)\,Mass + 935(\pm 47.31)\,MATS1v + 61,4(\pm 140,89)\,Mor20m - 360(\pm 24,38)\,H0v - 38,9(\pm 2,30)\,nCp$$

Tableau 1 : Paramètres statistiques des modèles sélectionnés

	Ov-101	**Cw-20M**			**Ov-101**	**Cw-20M**
EQMP	18,220	26,726	**90 observations**	R^2	99,6 %	99,07%
EQMC	16,644	24,409		Q^2	99,51%	98,88%
$EQMP_{ext}$	18,980	31,331	**23 observations**	Q^2_{ext}	99,47 %	99,46%

Pour les deux colonnes (CW et OV-101) les indices de rétention pour les 90 composés utilisés pour le calibrage sont bien corrélés avec les descripteurs d'où la grande valeur des coefficients de détermination R^2.

Nos modèles ont de très bonnes capacités prédictives confirmées par les valeurs de Q^2 qui sont supérieures à 90%. Les très faibles différences entre Q^2 et R^2 renseignent sur la robustesse des modèles qui sont très hautement significatifs (valeurs élevées de la statistique F de Fisher). De plus, la similitude de $EQMC$ et $EQMP$ signifie que les capacités de prédiction internes des modèles ne sont pas trop dissemblables de leurs pouvoirs d'ajustement.

Par ailleurs les matrices de corrélation (tableaux **2,3)** suggèrent que ces descripteurs sont faiblement corrélés les uns avec les autres. Ainsi, chaque modèle peut être considéré comme une équation de régression optimale.
Les descripteurs de chaque modèle, leurs classes et leurs définitions sont réunis dans les tableaux 4 et 5, alors que les tableaux 6 et 7 reproduisent les résultats obtenus sur chaque colonne.

Tableau 2 : Matrice de corrélation （CW-20M)

	CW	Polariz	nCIR	X4Av	Mor06v	nCt
Polariz	0,851					
nCIR	0,723	0,465				
X4Av	0,346	0,455	-0,202			
Mor06v	0,382	0,308	-0,152	0,556		
NCt	0,374	0,612	0,138	0,234	0,191	
Qmean	0,129	-0,343	0,222	-0,254	0,096	-0,316

Tableau 3 : Matrice de corrélation (OV-101)

	OV- 101	Mass	MATS1v	Mor20m	H0v
Mass	0,981				
MATS1v	0,702	0,647			
Mor20m	0,543	0,494	0,191		
H0v	0,542	0,580	0,534	0,361	
nCp	0,496	0,592	0,482	-0,025	0,166

Tableau 4 : Descripteurs sélectionnés pour la colonne CW-20M

N°	Descripteur	Classe	Définition
1	Polarizabilité	Descripteur électronique	La polarisabilité mesure le moment dipolaire de la molécule induit par un champ électrique
2	nCIR	Descripteur constititionnnel	Nombre de circuits
3	X4Av	Indice de connectivité	Indice de connectivité de valence moyenne d'ordre 4
4	Mor06v	Descripteur MoRSE 3D	Signal Morse 3D n°6 pondéré par le volume de van der Waals
5	nCt	Comptage des Groupements fonctionnels	Nombre total de carbones tertiaires (sp3)
6	Qmean	Descripteur de charge	La moyenne des charges absolues

Tableau 5: Descripteurs sélectionnés pour la colonne OV-101

N°	Descripteur	Classe	Définition
1	Masse	Descripteur constititionnnel	Masse moléculaire
2	MATS1v	autocorrelation 2D	Autocorrelation de Moran d'une structure topologique de distance 1 pondérée par le volume de van der waals
3	Mor20m	MoRSE3D	Signal Morse 3D numéro 20 pondéré par la masse atomique
4	H0v	GETAWAY	Autocorrelation de distance topologique 0 non pondérérée
5	nCp	Comptage des Groupements fonctionnels	Nombre total de carbones primaire (sp)

Tableau 6 : Valeurs calculées et observées des indices de rétention des pyrazines sur CW

	Composés	$Ir_{cw\ obs}$	$Ir_{cw\ calc}$	h_i	$e_{i,}$	$e_{i\ std}$
1	Pyrazine	1179	1199,7849	0,072	0,8489	0,9147
2	2,5-diméthylpyrazine	1290	1276,3194	0,055	-0,5535	-0,5855
3	2,6-diméthylpyrazine	1300	1304,7900	0,056	0,194	0,2055
4	Triméthylpyrazine	1365	1347,7415	0,079	-0,7076	-0,7683
5	Tetraméthylpyrazine	1439	1402,5743	0,106	-1,5157	-1,6953
6	Ethylpyrazine	1300	1301,7608	0,042	0,0708	0,0739
7	2-éthyl-5-méthylpyrazine	1357	1333,6998	0,038	-0,9344	-0,9708
8	2-éthyl-6-méthylpyrazine	1353	1339,7631	0,036	-0,5305	-0,5505
9	2,5-diméthyl-3-éthylpyrazine	1400	1399,659	0,044	-0,0137	-0,0143
10	2,6-diméthyl-3-éthylpyrazine	1415	1412,282	0,047	-0,1095	-0,1149
11	2,3-diméthyl-5-éthylpyrazine	1421	1405,7556	0,058	-0,618	-0,656
12	2,3-diéthylpyrazine	1417	1426,2363	0,048	0,3725	0,3914
13	2,3-diéthyl-5-méthylpyrazine	1459	1474,8054	0,044	0,6361	0,6655
14	Propylpyrazine	1374	1378,3687	0,051	0,1764	0,1859
15	2-méthyl-3-propylpyrazine	1438	1430,0898	0,037	-0,3172	-0,3296
16	2,3-diméthyl-5-propylpyrazine	1500	1482,7208	0,044	-0,6952	-0,7271
17	2,5-diméthyl-3-propylpyrazine	1474	1476,0091	0,041	0,0807	0,0842
18	2,6-diméthyl-3-propylpyrazine	1493	1484,6011	0,040	-0,3373	-0,3514
19	Isopropylpyrazine	1316	1311,3066	0,079	-0,1924	-0,2088
20	2,3-diméthyl-5-isopropylpyrazine	1431	1434,4568	0,064	0,1406	0,1502
21	Butylpyrazine	1474	1453,8431	0,047	-0,8122	-0,852
22	2-butyl-3-méthylpyrazine	1459	1506,821	0,043	1,9232	2,0094
23	3-butyl-2,5-diméthylpyrazine	1487	1556,4581	0,053	2,8074	2,9631
24	3-butyl-2,6-diméthylpyrazine	1514	1567,0937	0,050	2,1426	2,2542
25	5-butyl-2,3-diméthylpyrazine	1600	1577,4102	0,053	-0,9131	-0,9638
26	Isobutylpyrazine	1406	1401,1484	0,080	-0,199	-0,2162
27	2,3-diméthyl-5-isobutylpyrazine	1525	1526,0322	0,037	0,0414	0,043
28	2-isobutyl-3,5,6-triméthylpyrazine	1556	1596,7035	0,054	1,6466	1,7408
29	sec-butylpyrazine	1394	1392,0689	0,056	-0,0782	-0,0829
30	5-sec-butyl-2,3-diméthylpyrazine	1500	1517,1384	0,046	0,6902	0,7231
31	2,3-diméthyl-5-pentypyrazine	1700	1680,4554	0,064	-0,795	-0,8497
32	Isopentylpyrazine	1530	1497,287	0,054	-1,3231	-1,3984
33	2,3-diméthyl-5-isopentylpyrazine	1655	1625,6883	0,034	-1,1732	-1,2142
34	(2-méthylbutyl)pyrazine	1527	1481,6734	0,061	-1,8404	-1,9602
35	2,3-diméthyl-5-(2-méthylbutyl)pyrazine	1636	1623,777	0,036	-0,4897	-0,5077
36	2-(2-méthylbutyl)-3,5,6-triméthylpyrazine	1661	1696,7569	0,06	1,4512	1,5444
37	(2-méthylpentyl)pyrazine	1606	1598,0499	0,121	-0,3336	-0,3794
38	(2-éthylpropyl) pyrazine	1449	1469,5526	0,068	0,8377	0,8989
39	(1-méthylbutyl)pyrazine	1471	1502,7117	0,075	1,2972	1,4023
40	2,3-diméthyl-5-(2-méthylpentyl)pyrazine	1710	1741,5822	0,037	1,2659	1,314
41	Hexylpyrazine	1668	1653,9747	0,071	-0,5724	-0,616
42	Octylpyrazine	1845	1869,5784	0,115	1,0278	1,1609
43	2-méthyl-3-octylpyrazine	1956	1921,5325	0,127	-1,4516	-1,6633
44	2-méthyl-6-(2-méthylbutyl)-3-octylpyrazine	2254	2257,7774	0,276	0,1746	0,2411
45	2-methoxy-3-méthylpyrazine	1339	1357,2273	0,039	0,7316	0,7614

Tableau 6: suite et fin

	Composés	$Ir_{cw\ obs}$	$Ir_{cw\ calc}$	h_i	e_i	$e_{i\ std}$
46	2-méthoxy-5-méthylpyrazine	1358	1336,6102	0,048	-0,8625	-0,906
47	3-éthyl-2-méthoxypyrazine	1400	1389,3905	0,031	-10,9543	-0,4379
48	3-isopropyl-2-méthoxypyrazine	1400	1423,5833	0,05	24,8321	1,0025
49	5-isopropyl-3-méthyl-2-méthoxypyrazine	1467	1469,7399	0,049	2,8797	0,1162
50	5-isobutyl-3-méthyl-2-méthoxypyrazine	1556	1559,2717	0,031	3,378	0,135
51	2-éthoxy-3-méthylpyrazine	1385	1402,1302	0,044	17,9157	0,7208
52	2-éthoxy-5-méthylpyrazine	1418	1396,7858	0,045	-22,2027	-0,8937
53	2-éthoxy-3-éthylpyrazine	1439	1442,9049	0,079	4,2384	0,1737
54	2-éthoxy-5-isopropyl-3-méthylpyrazine	1500	1508,4726	0,077	9,1766	0,3757
55	2-éthoxy-5-isobutyl-3-méthylpyrazine	1584	1605,367	0,075	23,091	0,9444
56	5-sec-butyl-2-éthoxy-3-méthylpyrazine	1566	1589,1105	0,095	25,5426	1,0565
57	2-éthoxy-3-méthyl-5-(2-méthylbutyl)pyrazine	1693	1689,2041	0,096	-4,1981	-0,1737
58	(méthylthio)pyrazine	1600	1606,9031	0,107	7,7289	0,3218
59	3-éthyl-2-(méthylthio)pyrazine	1695	1666,6582	0,098	-31,4145	-1,3012
60	3-isopropyl-2-(méthylthio)pyrazine	1692	1686,6334	0,095	-5,9308	-0,2453
61	3-isopropyl-3-méthyl-2-(méthylthio)pyrazine	1737	1710,6272	0,044	-27,5731	-1,1092
62	5-sec-butyl-3méthyl-2-(méthylthio)pyrazine	1800	1781,8136	0,041	-18,9609	-0,7617
63	5-isobutyl-3-méthyl-2-(méthylthio)pyrazine	1816	1792,5622	0,048	-24,6239	-0,993
64	3-méthyl-5-(2-méthylpentyl)-2-(méthylthio)pyrazine	2008	1981,8015	0,093	-28,8707	-1,1924
65	(éthylthio)pyrazine	1635	1668,7319	0,162	40,269	1,731
66	2-éthylthio-5-isopropyl-3-méthylpyrazine	1769	1781,1267	0,049	12,7531	0,5145
67	5-sec-butyl-2-éthylthio-3-méthylpyrazine	1832	1859,1346	0,054	28,6736	1,1597
68	2-éthylthio-5-isobutyl-3-méthylpyrazine	1843	1867,9719	0,058	26,5219	1,0754
69	Phénoxypyrazine	2104	2082,2611	0,195	-27,0175	-1,185
70	2-méthyl-3-phénoxypyrazine	2103	2135,5025	0,227	42,0449	1,8814
71	5-isopropyl-3-méthyl-2-phénoxypyrazine	2114	2114,5483	0,173	0,6627	0,0287
72	5-sec-butyl-3-méthyl-2-phénoxypyrazine	2173	2185,2613	0,182	14,9946	0,6524
73	3-méthyl-5-(2-méthylbutyl)-2-phénoxypyrazine	2301	2291,3927	0,172	-11,6003	-0,5015
74	5-isopropyl-3-méthyl-2-(phénylthio)pyrazine	2375	2360,9499	0,167	-16,8699	-0,7273
75	5-sec-butyl-3-méthyl-2-(phénylthio)pyrazine	2430	2430,0845	0,166	0,1013	0,0044
76	Acetylpyrazine	1571	1591,8725	0,212	26,4765	1,1732
77	2-acétyl-3-méthylpyrazine	1567	1609,8967	0,114	48,4103	2,0233
78	2-acétyl-5-méthylpyrazine	1625	1592,9327	0,131	-36,9064	-1,5577
79	2-acétyl-6-méthylpyrazine	1618	1587,7577	0,121	-34,4085	-1,444
80	2-acétyl-3-éthylpyrazine	1617	1551,1287	0,166	-78,9556	-3,4009
81	Chloropyrazine	1351	1416,9042	0,077	71,4003	2,9239
82	2,3-dichloropyrazine	1581	1578,3757	0,119	-2,9802	-0,1249
83	2-chloro-3-méthylpyrazine	1399	1412,8009	0,06	14,6847	0,596
84	2-chloro-3-éthylpyrazine	1467	1458,0984	0,036	-9,2329	-0,37
85	2-chloro-3-isobuthylpyrazine	1575	1544,8202	0,04	-31,4338	-1,2621
86	2-chloro-5-isopropyl-3-méthylpyrazine	1505	1500,1676	0,068	-5,1828	-0,2112
87	5-sec-butyl-2-chloro-3-méthylpyrazine	1577	1575,1005	0,049	-1,9965	-0,0805
88	2-chloro-5-isobutyl-3-méthylpyrazine	1600	1588,0334	0,036	-12,4155	-0,4975
89	2-chloro-3-méthyl-5-(2-méthylbutyl)pyrazine	1710	1676,5075	0,043	-35,0147	-1,4086
90	2-chloro-3-méthyl-5-(2-méthylpentyl)pyrazine	1789	1792,0039	0,044	3,1413	0,1264

Tableau 7 : Valeurs calculées et observées des indices de rétention des pyrazines sur OV-101

	Composés	$Ir_{ov\ obs}$	$Ir_{ov\ cal}$	h_i	e_i	$e_{i\ std}$
1	Pyrazine	710	751,289	0,1	41,2894	2,7892
2	2,3-diméthylpyrazine	897	895,859	0,05	-1,1409	-0,0714
3	2,5-diméthylpyrazine	889	913,644	0,07	24,6441	1,5821
4	2,6-diméthylpyrazine	889	898,738	0,05	9,7385	0,6111
5	Triméthylpyrazine	981	979,227	0,08	-1,7728	-0,1159
6	Ethylpyrazine	894	886,094	0,04	-7,9059	-0,4902
7	2-ethyl-5-méthylpyrazine	980	981,892	0,03	1,8926	0,1151
8	2-ethyl-6-méthylpyrazine	977	966,767	0,03	-10,233	-0,6222
9	2,5-diméthyl-3-ethylpyrazine	1059	1043,04	0,04	-15,958	-0,9809
10	2,6-diméthyl-3-ethylpyrazine	1064	1045,21	0,04	-18,781	-1,1564
11	2,3-diméthyl-5-ethylpyrazine	1066	1048,47	0,04	-17,524	-1,0815
12	2,3-diethylpyrazine	1065	1049,80	0,03	-15,193	-0,919
13	2,3-diethyl-5-méthylpyrazine	1137	1132,90	0,03	-4,0992	-0,2475
14	Propylpyrazine	986	996,059	0,04	10,0595	0,6196
15	2-méthyl-3-propylpyrazine	1072	1062,73	0,02	-9,2703	-0,5566
16	2,3-diméthyl-5-propylpyrazine	1154	1145,56	0,03	-8,4393	-0,5111
17	2,5-diméthyl-3-propylpyrazine	1142	1138,99	0,03	-3,0052	-0,1814
18	2,6-diméthyl-3-propylpyrazine	1151	1142,32	0,03	-8,6786	-0,5245
19	2,3-diméthyl-5-isopropylpyrazine	1112	1099,29	0,07	-12,708	-0,8154
20	Butylpyrazine	1088	1089,27	0,05	1,2741	0,0797
21	2-butyl-3-méthylpyrazine	1121	1150,43	0,02	29,4385	1,7698
22	3-butyl-2,5-diméthylpyrazine	1184	1223,77	0,02	39,7781	2,3838
23	3-butyl-2,6-diméthylpyrazine	1196	1226,56	0,02	30,5605	1,8297
24	5-butyl-2,3-diméthylpyrazine	1254	1228,94	0,02	-25,06	-1,5035
25	Isobutylpyrazine	1043	1049,17	0,03	6,1765	0,3738
26	2,3-diméthyl-5-isobutylpyrazine	1200	1187,58	0,05	-12,412	-0,7728
27	2-isobutyl-3, 5,6-triméthylpyrazine	1263	1249,69	0,09	-13,301	-0,8867
28	sec-butylpyrazine	1040	1038,07	0,04	-1,9295	-0,1181
29	5-sec-butyl-2,3-diméthylpyrazine	1194	1182,25	0,05	-11,742	-0,7343
30	Isopentylpyrazine	1157	1149,54	0,03	-7,4548	-0,4537
31	2,3-diméthyl-5-isopentylpyrazine	1317	1286,19	0,04	-30,81	-1,9015
32	(2-méthylbutyl)pyrazine	1151	1144,48	0,03	-6,511	-0,3979
33	2,3-diméthyl-5-(2-méthylbutyl)pyrazine	1306	1279,75	0,04	-26,243	-1,6284
34	(2-méthylpentyl)pyrazine	1240	1246,58	0,05	6,5843	0,4123
35	(2-ethylpropyl)pyrazine	1121	1132,16	0,04	11,1632	0,6859
36	(1-méthylbutyl)pyrazine	1133	1130,1	0,04	-2,9005	-0,1781
37	2,3-diméthyl-5-(2-méthylpentyl)pyrazine	1377	1377,12	0,05	0,1255	0,0079
38	Hexylpyrazine	1293	1287,49	0,1	-5,5024	-0,3744
39	2-méthyl-3-octylpyrazine	1546	1532,91	0,13	-13,083	-0,93
40	2-méthyl-6-(2-méthylbutyl)-3-octylpyrazine	1962	1944,25	0,36	-17,745	-2,0226
41	Methoxypyrazine	877	838,538	0,19	-38,462	-3,0386
42	2-methoxy-3-méthylpyrazine	954	941,417	0,1	-12,583	-0,8597
43	2-methoxy-5-méthylpyrazine	969	957,085	0,11	-11,914	-0,8281
44	3-ethyl-2-methoxypyrazine	1037	1023,74	0,07	-13,255	-0,8609
45	3-isopropyl-2-methoxypyrazine	1078	1086,28	0,05	8,2859	0,5225

Tableau 7 : Suite et fin

	Composés	$Ir_{ov\ obs}$	Ir_{ov}	h_i	e_i	$e_{i\ Std}$
46	5-isopropyl-3-méthyl-2-methoxypyrazine	1170	1183,3	0,05	13,3048	0,8296
47	5-isobutyl-3-méthyl-2-methoxypyrazine	1257	1278,8	0,04	21,877	1,3462
48	3-méthyl-2-methoxy-5-(2-méthylbutyl)-pyrazine	1362	1377,1	0,05	15,1178	0,9424
49	Ethoxypyrazine	959	932,66	0,1	-26,335	-1,7935
50	2-ethoxy-3-méthylpyrazine	1029	1016,6	0,07	-12,325	-0,7935
51	2-ethoxy-5-méthylpyrazine	1047	1039,6	0,08	-7,3352	-0,4838
52	2-ethoxy-3-ethylpyrazine	1101	1095,2	0,05	-5,7344	-0,3591
53	2-ethoxy-5-isopropyl-3-méthylpyrazine	1230	1248,3	0,07	18,3131	1,1835
54	2-ethoxy-5-isobutyl-3-méthylpyrazine	1314	1340,4	0,06	26,4525	1,6838
55	5-sec-butyl-2-ethoxy-3-méthylpyrazine	1306	1333,4	0,06	27,4382	1,749
56	2-ethoxy-3-méthyl-5-(2-méthylbutyl)pyrazine	1415	1436,6	0,07	21,6474	1,3927
57	(méthylthio)pyrazine	1076	1086,2	0,08	10,2959	0,6747
58	3-méthyl-2-(méthylthio)pyrazine	1151	1175,9	0,07	24,9623	1,6022
59	5-méthyl-2-(méthylthio)pyrazine	1163	1171,1	0,06	8,1949	0,5233
60	3-ethyl-2-(méthylthio)pyrazine	1237	1227,9	0,04	-9,0293	-0,559
61	5-sec-butyl-3méthyl-2-(méthylthio)pyrazine	1441	1442,7	0,02	1,766	0,1063
62	5-isobutyl-3-méthyl-2-(méthylthio)pyrazine	1446	1447,7	0,02	1,7716	0,1062
63	3-méthyl-5-(2-méthylbutyl)-2-(méthylthio)pyrazine	1552	1537,0	0,03	-14,957	-0,9141
64	3-méthyl-5-(2-méthylpentyl)-2-(méthylthio)pyrazine	1638	1633,2	0,06	-4,7248	-0,3012
65	2-ethylthio-3-méthylpyrazine	1215	1239,2	0,05	24,2798	1,5183
66	2-ethylthio-5-isopropyl-3-méthylpyrazine	1418	1428,6	0,05	10,6186	0,665
67	5-sec-butyl-2-ethylthio-3-méthylpyrazine	1494	1503,3	0,04	9,3963	0,5769
68	2-ethylthio-5-isobutyl-3-méthylpyrazine	1496	1507,0	0,04	11,0236	0,6764
69	2-ethylthio-3-méthyl-5-(2-méthylbutyl)pyrazine	1602	1596,3	0,05	-5,6347	-0,353
70	Phenoxypyrazine	1415	1424,5	0,15	9,5655	0,7093
71	5-sec-butyl-3-méthyl-2-phenoxypyrazine	1694	1717,8	0,1	23,8178	1,6093
72	5-isobutyl-3-méthyl-2-phenoxypyrazine	1706	1726,8	0,07	20,8893	1,3527
73	3-méthyl-5-(2-méthylbutyl)-2-phenoxypyrazine	1807	1822,7	0,09	15,7268	1,0428
74	3-méthyl-2-(phenylthio)pyrazine	1658	1629,0	0,16	-28,945	-2,1879
75	5-isopropyl-3-méthyl-2-(phenylthio)pyrazine	1806	1785,0	0,21	-20,928	-1,7386
76	5-isobutyl-3-méthyl-2-(phenylthio)pyrazine	1882	1868,2	0,17	-13,779	-1,0595
77	3-méthyl-5-(2-méthylpentyl)-2-(phenylthio)pyrazine	2064	2043,4	0,16	-20,548	-1,5394
78	Acetylpyrazine	993	996,36	0,05	3,368	0,2106
79	2-acetyl-5-méthylpyrazine	1093	1086,7	0,04	-6,2536	-0,3879
80	2-acetyl-6-méthylpyrazine	1088	1069,3	0,03	-18,697	-1,1311
81	2-acetyl-3-ethylpyrazine	1138	1117,2	0,03	-20,757	-1,2699
82	Chloropyrazine	861	883,53	0,12	22,5304	1,5918
83	2-chloro-3-méthylpyrazine	951	957,04	0,11	6,0423	0,4151
84	2-chloro-3-ethylpyrazine	1044	1024,8	0,2	-19,187	-1,5435
85	2-chloro-5-isopropyl-3-méthylpyrazine	1173	1183,8	0,1	10,8678	0,7345
86	5-sec-butyl-2-chloro-3-méthylpyrazine	1256	1271,2	0,11	15,2247	1,0479
87	2-chloro-3-méthyl-5-(2-méthylbutyl)pyrazine	1371	1380,1	0,08	9,1062	0,5998
88	3-isopropyl-2-(méthylthio)pyrazine	1273	1273,2	0,03	0,2525	0,0152
89	5-isopropyl-3-méthyl-2-(méthylthio)pyrazine	1362	1366,2	0,03	4,2698	0,261
90	(ethylthio)pyrazine	1148	1159,6	0,05	11,6709	0,7293

3.Validation du modèle :

Pour vérifier les capacités prédictives de nos modèles on a eu recours à leurs validations. Ces ensembles de validation, sont constitués des composés numérotés de 91 à 113 pour les deux colonnes (tableaux 8et 9),

Tableau 8 *:* Quelques caractéristiques des éléments de l'ensemble de validation externe pour les pyrazines sur la colonne CW 20M

	Composés	**$Ir_{cw\ obs}$**	**$Ir_{cw\ pred}$**	**h_i**	**$ei_{\ std\ pre}$**
91	Méthylpyrazine	1235	1270,1438	0,05	1,4187
92	2,3-diméthylpyrazine	1309	1309,3532	0,071	0,0144
93	Pentylpyrazine	1575	1553,0944	0,065	-0,8911
94	2-méthyl-5-(2-méthylbutyl)-3-octylpyrazine	2200	2228,1769	0,268	1,2961
95	Methoxypyrazine	1306	1326,5524	0,055	0,8319
96	5-sec-butyl-3-méthyl-2-methoxypyrazine	1536	1551,0968	0,039	0,606
97	3-méthyl-2-methoxy-5-(2-méthylbutyl)-pyrazine	1664	1651,1225	0,039	-0,5168
98	3-méthyl-2-methoxy-5-(2-méthylpentyl)pyrazine	1737	1758,0189	0,036	0,8422
99	Ethoxypyrazine	1348	1376,7188	0,073	1,1739
100	2-ethoxy-3-isopropylpyrazine	1431	1472,7176	0,064	1,6966
101	2-ethoxy-3-méthyl-5-(2-méthylpentyl)pyrazine	1771	1797,2446	0,1	1,0885
102	3-méthyl-2-(méthylthio)pyrazine	1616	1645,6009	0,125	1,245
103	3-méthyl-5-(2-méthylbutyl)-2-(méthylthio)pyrazine	1941	1881,6423	0,059	-2,4071
104	2-ethylthio-3-méthylpyrazine	1655	1702,1847	0,134	1,9949
105	2-ethylthio-3-méthyl-5-(2-méthylbutyl)pyrazine	1951	1953,6073	0,076	0,1067
106	2-ethylthio-3-méthyl-5-(2-méthylpentyl)pyrazine	2026	2060,0015	0,114	1,421
107	5-isobutyl-3-méthyl-2-phenoxypyrazine	2209	2226,8984	0,152	0,7645
108	(phenythio)pyrazine	2400	2333,4251	0,334	-3,2096
109	3-méthyl-2-(phenylthio)pyrazine	2399	2381,9811	0,382	-0,852
110	5-isobutyl-3-méthyl-2-(phenylthio)pyrazine	2452	2451,3189	0,177	-0,0295
111	3-méthyl-5-(2-méthylbutyl)-2-(phenylthio)pyrazine	2569	2532,3174	0,183	-1,5962
112	3-méthyl-5-(2-méthylpentyl)-2-(phenylthio)pyrazine	2669	2633,1782	0,216	-1,5917
113	2-acetyl-3,5-diméthylpyrazine	1629	1638,8172	0,082	0,4031

Tableau 9 : Quelques caractéristiques des éléments de l'ensemble de validation externe pour les pyrazines sur la colonne OV-101

	Composés	$Ir_{ov\ obs}$	$Ir_{ov\ pred}$	h_i	$e_{i\ std\ pred}$
91	Isopropylpyrazine	949	945,7342	0,04	-0,1934
92	2-chloro-3-isobutylpyrazine	1187	1193,2564	0,15	0,3947
93	2-chloro-5-isobutyl-3-méthylpyrazine	1264	1282,752	0,08	1,1351
94	Méthylpyrazine	801	818,0761	0,05	1,0173
95	Tetraméthylpyrazine	1067	1051,2435	0,12	-0,9732
96	Pentylpyrazine	1192	1191,881	0,07	-0,0072
97	2,3-diméthyl-5-pentylpyrazine	1352	1323,7413	0,03	-1,6651
98	2-(2-méthylbutyl)-3,5,6-triméthylpyrazine	1363	1342,556	0,08	-1,2335
99	Octylpyrazine	1495	1481,9857	0,2	-0,8434
100	2-méthyl-5-(2-méthylbutyl)-3-octylpyrazine	1923	1899,1095	0,37	-1,7448
101	5-sec-butyl-3-méthyl-2-methoxypyrazine	1250	1270,9673	0,04	1,2412
102	3-méthyl-2-methoxy-5-(2-méthylpentyl)pyrazine	1444	1479,0004	0,07	2,1036
103	2-ethoxy-3-isopropylpyrazine	1143	1151,4397	0,06	0,5046
104	2-ethylthio-3-méthyl-5-(2-méthylpentyl)pyrazine	1686	1689,6558	0,07	0,2206
105	2-méthyl-3-phenoxypyrazine	1465	1456,8463	0,07	-0,4918
106	5-isopropyl-3-méthyl-2-phenoxypyrazine	1620	1629,1271	0,1	0,5571
107	(phenylthio)pyrazine	1606	1582,9617	0,22	-1,5135
108	5-sec-butyl-3-méthyl-2-(phenylthio)pyrazine	1874	1866,3768	0,18	-0,4871
109	3-méthyl-5-(2-méthylbutyl)-2-	1985	1952,2687	0,14	-2,0535
110	2-acetyl-3-méthylpyrazine	1061	1079,2741	0,03	1,0761
111	2-acetyl-3,5-diméthylpyrazine	1153	1160,9824	0,05	0,4746
112	2,3-dichloropyrazine	1032	1007,4456	0,48	-1,9724
113	2-chloro-3-méthyl-5-(2-méthylpentyl)pyrazine	1456	1484,7702	0,08	1,7424

L'application du modèle, calculé sur l'ensemble de calibrage, aux composés de l'ensemble de validation externe permet de vérifier de manière fiable la capacité prédictive du modèle obtenu. La valeur de Q^2_{ext} nous renseigne sur la validité du modèle et sa capacité à prédire des valeurs qui n'ont pas servi à le générer.

3.1. Qualité de l'ajustement

Les figures 2 et 3 représentent la droite d'ajustement pour l'ensemble de calibrage (Ir_{cal} en fonction de Ir_{obs}) et celle pour l'ensemble de validation (Ir_{pred} en fonction de Ir_{obs}).

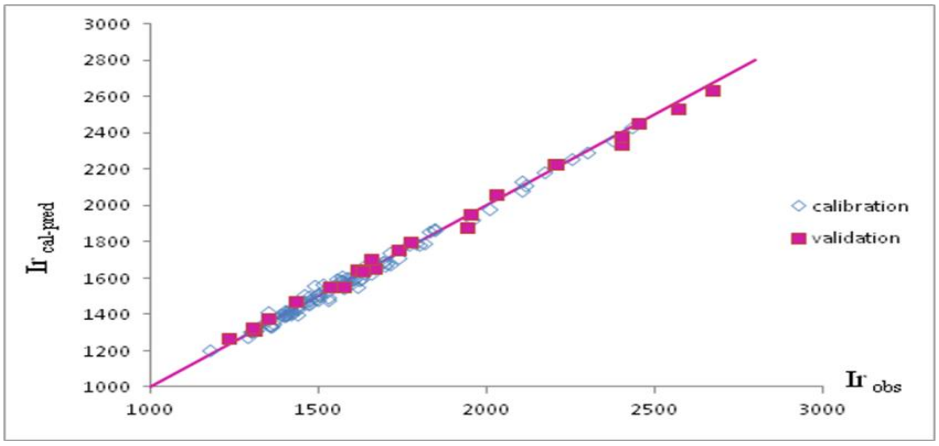


Figure 2 : Droite d'ajustement des Ir pour les pyrazines séparées sur la colonne CW-20M.

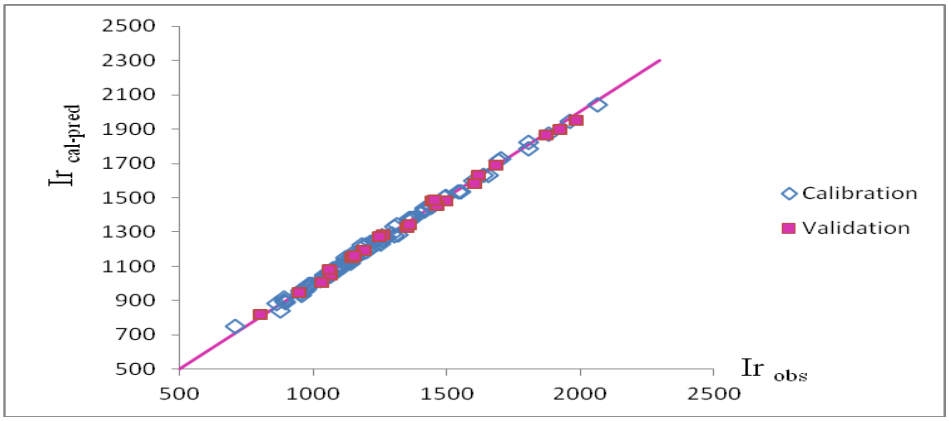


Figure 3 : Droite d'ajustement des Ir pour les pyrazines séparées sur la colonne OV-101.

On remarque une faible dispersion autour de la droite d'ajustement ce qui traduit la faiblesse des erreurs lors du calcul (Calibrage) et de la prédiction (Validation). On déduit donc, des figures précédentes, qu'on a un bon ajustement confirmé par les valeurs du coefficient de régression R^2 supérieures à 90%, pour les deux ensembles sur chacune des deux colonnes, ce qui renforce la performance des modèles établis.

3.2 Test de randomisation

Les modèles QSRR, à cause (souvent) de leur complexité et de la sophistication des outils de chimiométrie employés, peuvent constituer une source de corrélations fortuites, Dans le but d'établir que le modèle n'est pas dû au hasard, on a appliqué le test de randomisation de Y [13]. Ce test consiste à générer un vecteur de la propriété étudiée par permutation aléatoire des composantes du vecteur réel, On calcule alors sur le vecteur obtenu un modèle QSRR, selon la méthode habituelle (figure 7).

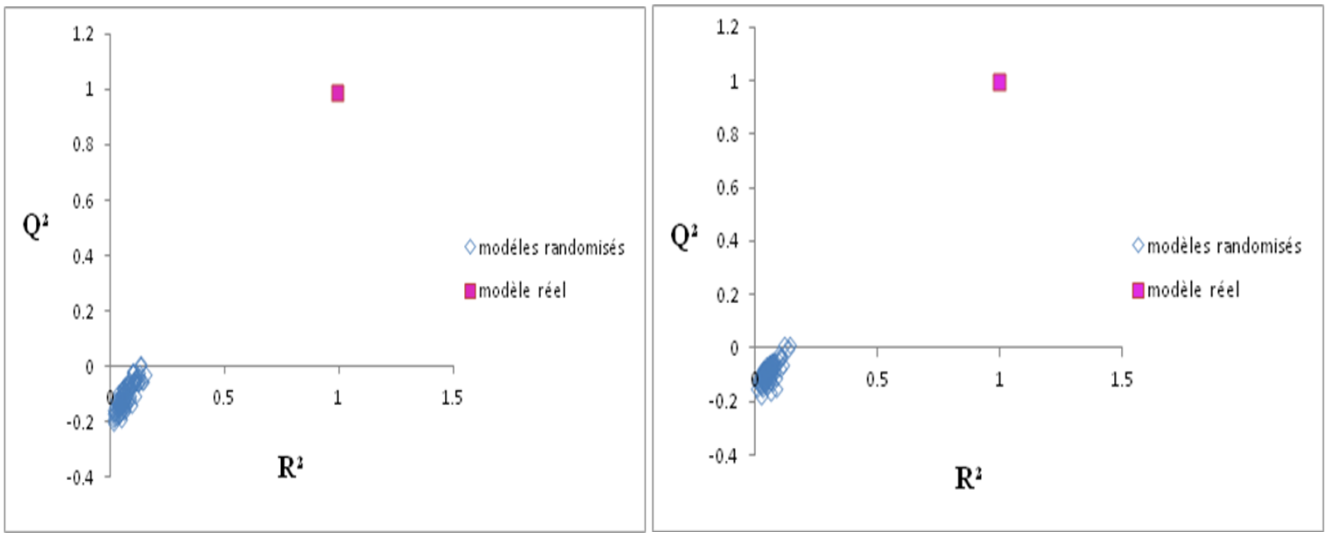


Figure 4 *:* Test de randomisation CW-20M et OV-1O1.

Les cercles en bleu regroupés dans la région des valeurs négatives de Q^2 ont des valeurs petites de R^2 ; ce sont les représentations des modèles randomisés. Seuls les deux carrés remplis en rose ont des valeurs élevées et proches pour ces deux statistiques, ils représentent nos modèles qui, par conséquent, ne sont pas dû au hasard.

3.3 Diagramme de Williams

On a représenté dans les figures 4 et 5 le diagramme de Williams, pour chaque modèle, qui permettent de visualiser les valeurs des résidus de prédiction standardisés en fonction de levier, pour les deux ensembles (calibrage et validation).

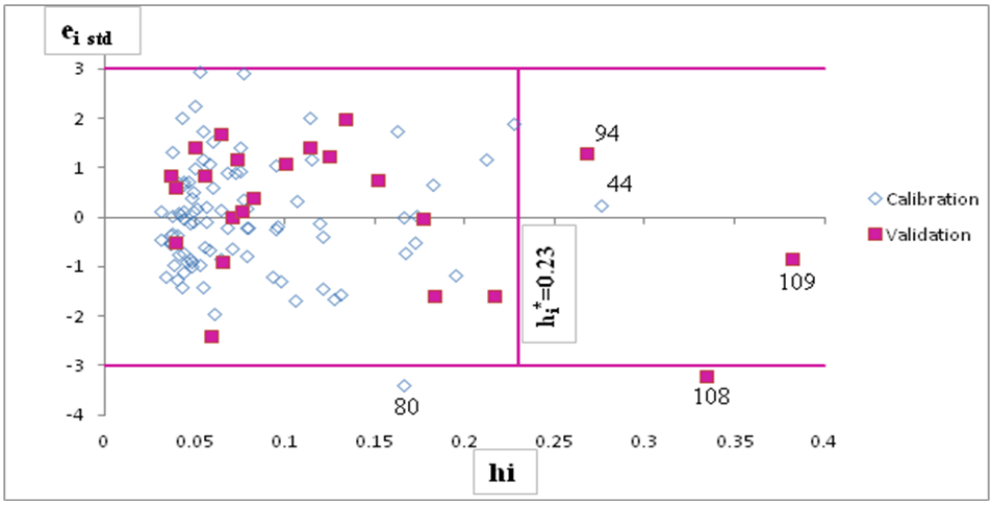


Figure 5 : Diagramme de Williams (colonne CW-20M).

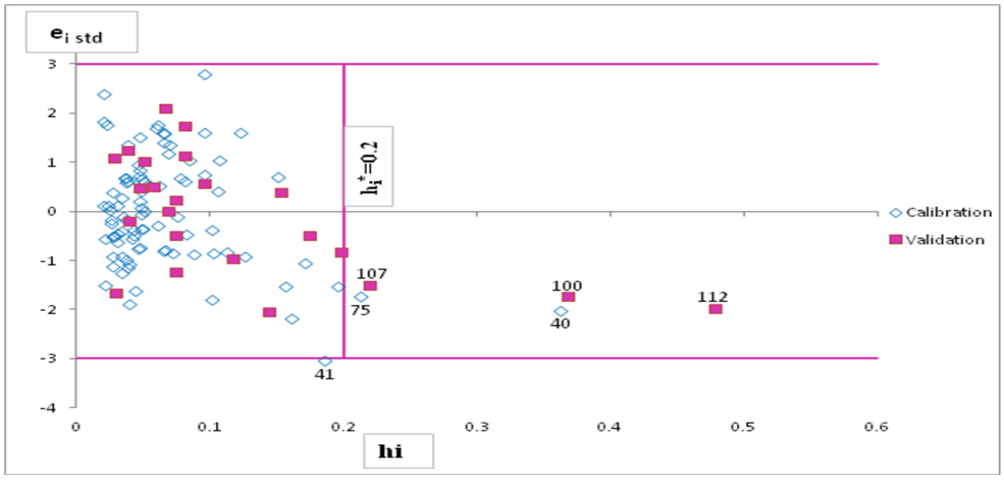


Figure 6 : Diagramme de Williams (colonne OV-101).

Manifestement les diagrammes de Williams pour chacune des colonnes (Figures 4-5) sont caractérisés par des composés aberrants précisés dans le tableau 10.

Tableau 10 : Les composés aberrant pour les deux colonnes

CW		OV-101	
44	2-méthyl-6-(2-méthylbutyl)-3-octylpyrazine	**40**	2-méthyl-6-(2-méthylbutyl)-3-octylpyrazine
80	2-acetyl-3-ethylpyrazine	**41**	Methoxypyrazine
94	Méthylpyrazine*	**75**	5-isopropyl-3-méthyl-2-(phenylthio)pyrazine
108	(phenythio)pyrazine*	**100**	2-méthyl-5-(2-méthylbutyl)-3-octylpyrazine*
109	3-méthyl-2-(phenylthio)pyrazine*	**107**	(phenylthio)pyrazine*
*** pour les composés de validation**		**112**	2,3-dichloropyrazine*

En règle générale, si la valeur des erreurs standardisées est supérieure à ± 3s, l'échantillon peut être considéré comme une valeur aberrante de la réponse « valeur aberrante en Y » (cas des composés 41, 80, 108), ce qui peut être associé à des erreurs dans les valeurs expérimentales. Les valeurs aberrantes en X, celles qui présentent un effet de levier supérieure à la valeur critique, sont des composés structurellement influents (composés : 94, 44, 109, 107, 100, 75, 40, 112), elles peuvent être associés à des composés présentant des caractéristiques particulières mal représentés dans l'ensemble d'apprentissage, ce qui pourrait affecter la sélection des variables pour une meilleure modélisation de ces composés.

Il est pertinent de noter que 86,96% du domaine est couvert par les modèles lorsqu'il a été appliqué pour prévoir l'indice de rétention mesuré sur les deux colonnes des 23 pyrazines de l'ensemble de validation.

La suppression des observations aberrantes modifie légèrement les valeurs de R^2 et Q^2 à 99,09% et 98,91% pour CW20M et 99,56% et 99,48% pour OV 101. Ainsi que les nouveaux diagrammes de Williams illustrés dans la figure 4.

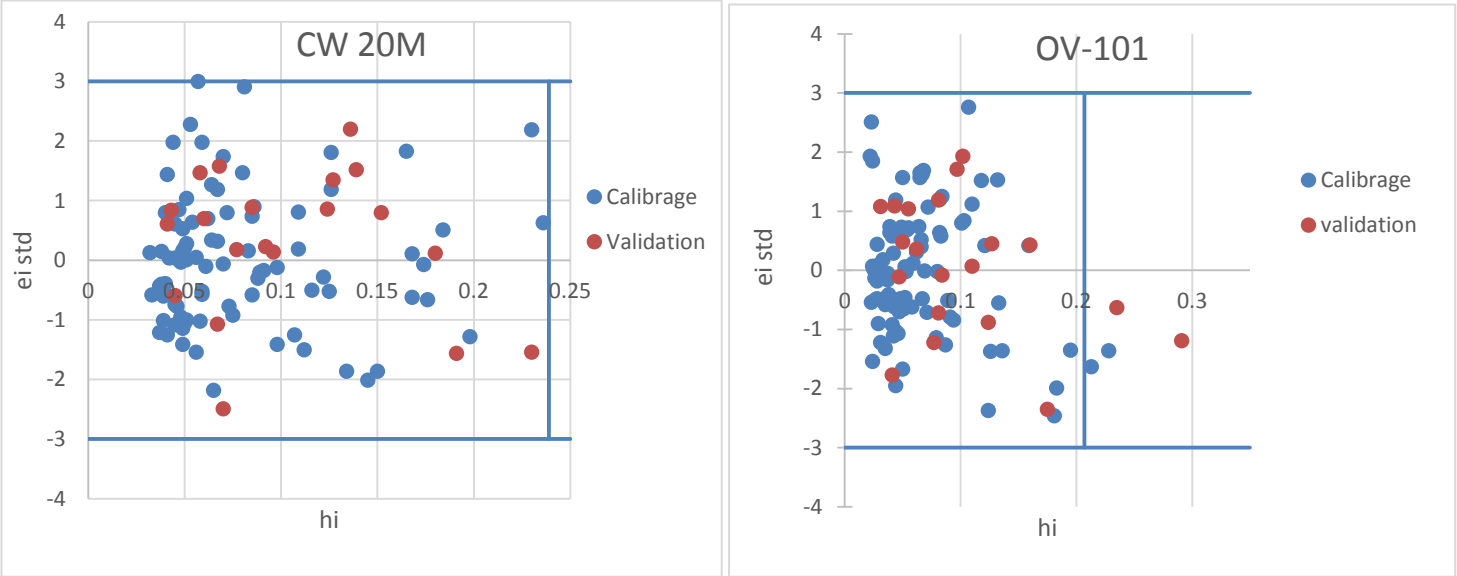


Figure 7 : Diagramme de Williams de CW20M et OV101 après suppression des composés aberrants

Conclusion

Les indices de rétention de 113 pyrazines éluées sur 2 colonnes de polarités très différentes ont été corrélés respectivement avec 6 (colonne Carbowax20M) et 5 (colonne OV-101) descripteurs théoriques calculés uniquement à partir de la structure des molécules, et sélectionnés par algorithme génétique parmi plus de 1600 descripteurs moléculaires obtenus à l'aide du logiciel DRAGON.

Les modèles QSRR présentés sont robustes, avec de bonnes capacités prédictives internes et externes, et une bonne qualité de l'ajustement. Des composés aberrants sont apparus dans le diagramme de williams, leur suppression modifie légèrement les paramètres statistiques et les diagrammes de williams des deux colonnes.

Références

[1]. Hansch, C., Fujita, T. ρ-σ-π analysis. A method for the correlation of biological activity and chemical structure.. American Chemistry Society., 1964. 86. 1616.

[2]. Kaliszan, R, Quantitative relationships between molecular structure and chromatographic retention, CRC Crit, Rev, Anal, Chem. 1986.16.323.

[3]. Kaliszan, R, Quantitative structure-chromatographic retention relationships, J, Wiley,New York, Energy Citations Database. 1987. 303

[4]. Wang, Q,S,, Zhang, L,M,, Zhang, X,D, Xing, G,, Tang, Z, , A system for predicting the retentions of O-alkyl, n-(1-methylthioethylideneamino) phosphoramidates on RPHP, Chromatographia, 1999.49. 444.

[5]. Lee, Seung Ki,, Polyakova, Yulia,, Row, Kyung Ho, Evaluation of predictive retention factors for phenolic compounds with QSPR equations, Liq, Chromatogr, and Rel,Tech. 2004,27. 629.

[6]. Mihara S. Enomoto N.. calculation of retention indices of pyrazines on the basis of molecular structure. Journal of Chromatography. 1985.324. 428.

[7]. Mihara, S., Masuda, H. Correlation between molecular structures and retention indices of pyrazines. Journal of Chromatography. 1987. 402. 309.

[8]. Masuda, H., Mihara, S. Use of modified molecular connectivity indices to predict retention indices of monosubstituted alkyl, alkoxy, alkylthio, phenoxy and (phenylthio) pyrazines. Journal of Chromatography. 1986. 366. 373.

[9]. Stanton, D.T., Jurs, P.C. Computer-assisted prediction of gas chromatographic retention indexes of pyrazines. Analytical Chemistry. 1989.611.328.

[10]. Small, G.W., Jurs, P.C. Interactive computer system for the simulation of carbon-13 nuclear magnetic resonance spectra. Analytical Chemistry. 1983.55. 1121.

[11]. TOUHAMI. I, MOKRANI. K et MESSADI. D. modèles QSRR hybrides algorithme génétique-régression linéaire multiple des indices de rétention de pyrazines en chromatographie gazeuse. Lebanese Science Journal. 2012.13 .75.

[12]. Todeschini , R, Ballabio, D, Consonni, V, Mauri, A, Pavan, V.. MobyDigs 1.1, Copyright TALETE srl. 2009.

[13]. Wold S, Eriksson L, , Statistical validation of QSAR results, In: H, Van de Waterbeemd ed, Chemometrics methods in molecular design, VCH, New York,, 1995, 2,309.

CONCLUSİON GÉNÉRALE

Conclusion générale

La caractérisation expérimentale de l'ensemble des propriétés de composés chimiques est contraignante pour des raisons de temps, de coûts, d'éthique (essais sur animaux) et de faisabilité au niveau recherche et développement. Ainsi, le développement de méthodes prédictives, est recommandé. En particulier l'utilisation des modèles QSPR fournit un compromis nécessaire qui permet l'estimation des paramètres physicochimiques de grandes classes de composés qui sont importants du point de vue théorique ou pour les applications industrielles.

Le point de départ de telles méthodes se construit sur la définition des descripteurs moléculaires. Ces derniers prennent en compte des informations sur la structure et les caractéristiques physico-chimiques des molécules

L'objet de notre travail est de trouver le lien entre les descripteurs et la base de données grâce à des outils d'analyse comme les régressions multilinéaires (MLR), et les réseaux de neurones (RNA).

Deux approches, la régression linéaire multiple, et les réseaux de neurones artificiels, (MLR, RNA) ont été appliqués pour relier la solubilité aqueuse d'une série de phénols à des descripteurs de plusieurs types.

Les 68 données de base ont été éclatées en deux ensembles par plusieurs méthodes de fractionnement, un ensemble de 48 composés utilisées pour la construction de modèle et un autre ensemble de 20 composés pour la prédiction externe. Un modèle de 3 descripteurs à été choisi, la sélection des variables explicatives à été réalisée par algorithme génétique, dans la version QSARINS de GRAMATICA, en maximisant Q^2_{LOO}.

La comparaison de la qualité des modèles MLR et RNA favorise la relation non linéaire entre l'information structurelle et les valeurs de la solubilité des composés.

La mise en évidence de la modélisation de la température d'ébullition est effectuée pour un ensemble de 56 phénols par la méthode de régression linéaire multiple. Les résultats obtenus indiquent que l'utilisation des descripteurs PW5, Hy, X5A et R6m fournit une bonne estimation des points d'ébullition. Ainsi que la qualité de l'ajustement, la robustesse et la prédictivité du modèle GA-MLR étaient significatives pour les validations internes et externes.

Conclusion générale

L'approche hybride algorithme génétique/régression linéaire multiple a été appliquée pour modéliser, séparément, les indices de rétention d'un même ensemble de 113 pyrazines éluées tour à tour sur les colonnes OV-101 et Carbowax-20M, en utilisant des descripteurs moléculaires théoriques calculés à l'aide du logiciel DRAGON.

Deux modèles de tailles différentes pour les deux colonnes ont été obtenus, ils sont caractérisés par de très bonnes statistiques d'ajustement et de validation. Sur la colonne non polaire OV-101 ce sont les interactions de dispersion et la complexité des molécules qui gouvernent la rétention, alors que sur la colonne Carbowax- 20M ce sont les interactions spécifiques et la symétrie des molécules qui s'imposent.

Des validations rigoureuses interne et externe on été utilisées pour juger la stabilité, la justesse et la capacité prédictive des modèles obtenus pour les différentes propriétés.

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

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

Le test de randomisation associé à chaque modèle obtenu permet d'assurer qu'une relation structure/activité réelle a été établie.