الجمهورية الجزائرية الديمقراطية الشعبية

REPUBLIQUE ALGERIENNE DEMOCRATIQUE ET POPULAIRE

وزارةالتعليمالعالي و البحثالعلمي

BADJI MOKHTAR- ANNABA UNIVERSITY
UNIVERSITE BADJI MOKHTAR- ANNABA

جامعة باجي مختار-عنابة

Année 2021

FACULTE DES SCIENCES

DEPARTEMENT DE CHIMIE

THESE

Pour l'obtention du diplôme de doctorat en sciences

Option : Chimie analytique et environnement

Etude QSRR et QSTR de dérivés benzéniques

Presenté par: Mme DJELLOUL Karima

Mme FERTIKH Nadia	Président	Professeur	Université d'Annaba
Mme. ALIMOKHNACHE Salima	Encadreur	Professeur	Université d'Annaba
Mr. KHORIEF NACEREDDINE Abdelmalek	Examinateur	Professeur	ENSET de Skikda
Mr. CHAWKI Mourad	Examinateur	Maître de conférences	Université d'Ouargla
Mr. DEMS Mohamed Abdessalem	Examinateur	Maître de recherche	CRBT Constantine

Je dédie ce modeste travail à la mémoire de mon père

Remerciement

Je remercie DIEU le tout puissant qui nous a donné la force, la volonté et le courage pour accomplir ce modeste travail.

J'adresse un grand remerciement de manière très particulière et à exprimer ma profonde reconnaissance au Professeur ***ALIMOKHNACHE Salima*** *ma directrice de thèse, pour m'avoir accordé une grande confiance, la bienveillance et la sympathie avec laquelle elle m'a accueilli, Je la remercie pour ses qualités professionnelles incontestables ainsi que son soutien. J'ai pu apprécier l'étendue de ces connaissances, sa disponibilité et ses hautes qualités humaines.*

Je remercie vivement ***FERTIKH Nadia,*** *Professeur à l'université d'Annaba pour sa participation active à mon jury de thèse, pour avoir accepté de juger ce travail et d'en présider le jury de soutenance.*

Je tiens aussi à remercier vivement ***Pr. KHORIEF Nacereddine,*** *professeur à ENSET de Skikda, pour l'honneur qu'il m'a fait pour sa participation à mon jury de thèse, pour avoir accepté d'examiner et d'évaluer ce travail.*

Mes vifs remerciements à ***Mr CHAWKI Mourad,*** *maître de conférences à l'université d'Ouagla, pour l'honneur qu'il m'a fait pour sa participation à mon jury de thèse, pour avoir accepté d'examiner et de juger ce travail.*

Mes sincères remerciements à ***Mr DEMS Mohamed Abdessalem,*** *maître de recherche au CRBt Constantine, our l'honneur qu'il m'a fait pour avoir accepté de participer au membre de jury, d'examiner et de juger ce travail.*

Ma sincère gratitude est exprimée à tous ceux ou celles qui ont contribué à l'aboutissement de ce travail et spécialement Messieurs ***DJILANI Salah Eddine et DADA Noureddine.***

Résumé:

La présence des hydrocarbures dans l'environnement est devenue un souci important. Par conséquent, ces molécules sont sous une grande surveillance. Pour cette raison, il est impératif d'avoir l'identification structurale fiable et les mesures quantitatives précises de ces substances.L'expérimentation est une manière directe d'obtenir des données sur l'activité de ces composés organiques. Une telle expérience peut être déficiente en termes de la nécessité de grands organismes expérimentaux, il peut ne pas être possible à l'expérimentation de fournir les valeurs des activités de tous les composés organiques. Les QSAR (Relations Quantitatives Structure-Activité) et QSPR (Relations Quantitatives Structure-Propriété) peuvent remplacer l'expérience dans le cas où l'activité et/ou la propriété du composé vient à manquer.Une relation quantitative structure-rétention (QSRR) a été produite pour la prédiction des indices de rétention de 38 dérivés benzéniques. Les données exigées pour la validation externe ont été obtenues en séparant « a priori » les données disponibles, en utilisant l'algorithme de Kennard et Stone, deux sous-ensembles disjoints de calibrage (28 éléments) et de validation (10 éléments). Un modèle à 2 descripteurs (indices de connectivité) a été choisi. Leurs contributions dans la construction du modèle QSRR ont été vérifiées et sont significatives. La fiabilité du modèle QSPR a été validée par diverses techniques d'évaluation: validation par omission d'une observation, test de randomisation et validation externe.Une deuxième relation quantitative structure-activité (QSAR) a été développée pour la prédiction de la toxicité de 141 dérivés de benzène vis à vis de *pimephales promelas*, également appelée *fathead minnow* ou la tête de boule. Les descripteurs moléculaires ont été choisis dans un ensemble prolongé de 1664 descripteurs (constitutionnel, topologique, géométrique et quantiques) par 3 méthodes différentes de choix des variables (méthode de remplacement, régression pas-à-pas et par algorithmes génétiques). La méthode RNA est robuste, avec de bonne capacité prédictive interne et externe, et une bonne qualité de l'ajustement (R^2 = 0.8776, SDEC = 0.2477, $SDEP_{ext}$ = 0.2718).

Mots-clés : Dérivés benzéniques, Indices de Rétention, Toxicité, QSAR, QSPR

Abstract:

The presence of hydrocarbons in the environment has become a major concern. Therefore, these molecules are under high surveillance. For this reason, it is imperative to have both reliable structural identification and accurate quantitative measurements of these substances. Experience is a direct way to obtain data on the activity of these organic compounds. Such an experiment may be deficient in terms of the requirement of large experimental organizations, it may not be possible for the experiment to provide the values of the activities of all organic compounds. QSAR (Quantitative Structure-Activity Relationships) and QSPR (Quantitative Structure-Property Relationships) can replace the experiment in the case of compound's activity and/or property lacking. A quantitative structure-retention relationship (QSRR) was performed for the prediction of the retention indices of 38 benzene derivatives. The data required for external validation were obtained by separating "a priori" the available data, using the Kennard and Stone algorithm, two disjoint sets of calibration (28 elements) and test (10 elements). A model of 2 descriptors (connectivity indices) was selected. Their significant contributions in building QSPR were verified. The reliability of the QSPR has been validated using various evaluation techniques: leave-one / more-out, randomization testing, and external validation. A second quantitative structure-activity relationship (QSAR) has been developed for the prediction of the toxicity of 141 benzene derivatives to *pimephales promelas*, also called *fathead minnow* or ball head. The molecular descriptors were selected in an extended set of 1664 descriptors (constitutional, topological, geometric and quantum) by 3 different variable selection methods (replacement method, genetic algorithm and stepwise). The database is reduced to heterogeneous benzene derivatives. The RNA method was robust, with good internal and external predictive abilities, and good quality of fit (R^2 = 0.8776, SDEC = 0.2477, $SDEP_{ext}$ = 0.2718).

Keywords : Benzene derivatives, retention indices, toxicity, QSAR, QSPR

ملخص

أصبح وجود الهيدروكربونات في البيئة مصدر قلق كبير. لذلك ، تخضع هذه الجزيئات لفحص دقيق. لهذا السبب ، من الضروري أن يكون لديك تحديد هيكلي موثوق وقياسات كمية دقيقة لهذه المواد.

التجريب طريقة مباشرة للحصول على بيانات عن نشاط هذه المركبات العضوية. قد تكون هذه التجربة ناقصة من حيث الحاجة إلى كائنات تجريبية كبيرة ، وقد لا يكون من الممكن للتجربة توفير قيم أنشطة جميع المركبات العضوية. يمكن لـ QSAR (العلاقات الكمية بين البنية والنشاط) و QSPR (العلاقات الكمية بين البنية و الخاصية) أن تحل محل الخبرة في حالة عدم وجود نشاط و / أو خاصية للمركب. تم

إنتاج علاقة كمية الاحتفاظ بالبنية (QSRR) للتنبؤ بمؤشرات الاحتفاظ ل 38 من مشتقات البنزين. تم الحصول على البيانات المطلوبة للتحقق الخارجي من خلال فصل البيانات المتاحة "بداهة" ، باستخدام خوارزمية كينارد وستون ، ومجموعتين فرعيتين منفصلتين من المعايرة (28 عنصرًا) والتحقق من الصحة (10 عناصر). تم اختيار نموذج مع 2 واصفات (مؤشرات التوصيل). تم التحقق من مساهماتهم في بناء نموذج QSRR وهي كبيرة. تم التحقق من موثوقية نموذج QSPR من خلال تقنيات تقييم مختلفة: التحقق من الصحة عن طريق حذف الملاحظة ، واختبار العشوائية والتحقق الخارجي. تم تطوير علاقة كمية ثانية بين

البنية والنشاط (QSAR) للتنبؤ بسمية 141 مشتقًا من البنزين إلى pimephales promelas. تم اختيار الواصفات الجزيئية من مجموعة موسعة من 1664 واصفاً (بنيوي، طوبولوجي ، هندسي وكمي) من خلال 3 طرق مختلفة لاختيار المتغيرات (طريقة الاستبدال ، الانحدار التدريجي والخوارزميات الجينية).

طريقة ANN قوية ، مع قدرة تنبؤية داخلية وخارجية جيدة ، ونوعية جيدة من الملائمة.

كلمات مفتاحية: مشتقات البنزين ، مؤشرات الاحتفاظ ، السمية ، QSAR، QSPR

SYMBOLES ET ABREVIATIONS

AM1 : Austin Model 1.

d : Statistique de Durbin-Watson.

DL_{50} Dose létale 50

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 ($\hat{\mathbf{y}}_\mathbf{i}$).

F : Statistique de Fisher.

FIV: Facteur d'inflation de la variance.

GA: Algorithme génétique (Genetic Algorithm).

GA-VSS Genetic Algorithm for Variable Subsets Selection

H Opérateur hamiltonien

$\boldsymbol{h_i}$ Eléments diagonaux de la matrice chapeau

$\boldsymbol{h_{ii}}$ Eléments diagonaux de la matrice influence moléculaire

$\boldsymbol{h^*}$ Valeur critique des leviers

K Pente de la droite de régression passant par l'origine pour les valeurs calculées par rapport aux valeurs observées

k' Pente de la droite de régression passant par l'origine pour les valeurs observées par rapport aux calculées

Kov coefficient de partage **octanol/eau**

IR Indice de rétention

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

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

n: Dimension de la population (échantillon).

n-p : Nombre de degrés de liberté.

MAE : Erreur absolue moyenne en anglais Mean absolute error

PRESS : Somme des carrés des erreurs de prédiction.

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

QSAR : Quantitative Structure/ Activity Relationships.

(Relations Quantitatives Structure/ Activité).

QSPR : Quantitative Structure/ Propriety Relationships.

(Relations Quantitatives Structure/ Propriété).

QSRR : Quantitative Structure/ Retention Relationships.

(Relations Quantitatives Structure/ Retention).

QSTR : Quantitative Structure/ Toxicity Relationships.

(Relations Quantitatives Structure/ Toxicité).

Q2 : Coefficient de prédiction interne

Q^2_{EXT} : Coefficient de prédiction externe

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

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

RHF Restricted Hartree-Fock

RLM (MLR): Régression linéaire multiple.

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

RNA: Réseaux de neurones artificiels.

R^2 : Coefficient de détermination.

R^2_{adj} *:* Coefficient de détermination ajusté

r_i : Résidu studentisé interne.

SCE : Somme des carrés des écarts.

SCT : Somme des carrés totale.

T : t de Student.

y_i: Valeur observée.

$\hat{y}_i$: Valeur estimée.

$\hat{y}_{(i)}$: Valeur prédite.

Liste des tableaux :

Liste des figures :

SOMMAIRE

Introduction générale

La pollution de l'environnement, sous ses différentes formes, a fait beaucoup de bruit et a fait couler beaucoup d'ancre ces dernières décennies. Et cela trouve sa justification dans le danger qui menace notre planète à tous les niveaux : sol, eau, atmosphère et produits alimentaires ; ce qui a motivé les chercheurs scientifiques à multiplier les efforts déployés pour résoudre cette problématique en cherchant des moyens de rétention et d'élimination de ces polluants.

Il existe différentes sources de pollution telle que par exemple : les composés organiques volatils qui comprennent les hydrocarbures de la série aliphatique et mono aromatique.

Les Composés Organiques Volatils, ou COV, font partie des principaux polluants atmosphériques. Ils sont souvent évoqués dans le cadre de la surveillance de la pollution atmosphérique, de même que les oxydes d'azote, le dioxyde de soufre ou encore l'ozone. Cependant, la définition de ces composés reste floue, voire ambiguë pour beaucoup. Le benzène fait partie de cette famille de ces composés les plus connus, il s'agit du seul composé organique réglementé à l'heure actuelle. Mais il existe une multitude de substances répondant à la définition de ces derniers. Ces derniers constituent un groupe de substances hétérogènes, pour la plupart encore mal connues et possédant des propriétés variées. Les études concernant les composés oraganiques volatils sont très variables d'un composé à l'autre : certains sont bien étudiés, d'autres très peu.

Or ces composés sont susceptibles d'avoir des effets sur la santé humaine: des effets aigus liés à une exposition à une forte dose sur une courte période, mais aussi des effets chroniques liés à des expositions à de faibles doses sur le long terme, tels que des effets cancérigènes ou toxiques pour la reproduction et le développement de l'homme. Leur présence dans l'atmosphère issue également des réactions chimiques, qui peuvent aboutir à la formation ou l'accumulation dans l'environnement d'autres composés nocifs, tels que l'ozone. Il est donc essentiel d'étudier ces substances et d'en évaluer les risques sanitaires pour la population.

L'évaluation de la pollution par les hydrocarbures moyennant les analyses quantitatives et qualitatives s'avère très onéreuses bien qu'elle soit indispensable en fournissant des données physicochimiques quantifiées. Néanmoins, ces mesures permettent de connaitre la pollution, de mesurer les concentrations des polluants présents et aussi en mesurer les effets**.**

L'expérience est un moyen direct pour obtenir des données de l'activité/propriété des composés organiques. Une telle expérience peut être déficiente en termes d'exigence de grands organismes expérimentaux, coûte beaucoup d'argent et prenant beaucoup de temps, en plus de la différence entre les valeurs mesurées par différents chercheurs selon les conditions expérimentales. Par conséquent, il serait impossible que l'expérience fournisse les valeurs des activités de tous les composés organiques.

Il est donc crucial d'utiliser des méthodes théoriques pour compenser les inconvénients de l'expérience et pour prédire les données (activités ou propriétés) exactes des composés.

Le développement significatif de l'informatique ainsi que des études théoriques de la chimie quantique permettent aux chercheurs d'obtenir des paramètres physicochimiques et quantiques plus précis des composés en un temps plus court.

C'est l'objectif principal des études des relations quantitatives structure-activité QSAR, et des relations quantitatives structure propriété QSPR. Ces études se basent essentiellement sur la recherche de similitudes entre molécules dans de grandes bases de données de molécules existantes dont les activités ou les propriétés sont connues. La découverte d'une telle relation permet de prédire les activités et les propriétés des nouveaux composés. Les relations entre les structures des molécules et leurs activités ou propriétés sont généralement établies à l'aide de méthodes de modélisation moléculaire et des méthodes statistiques. Les techniques usuelles reposent sur la caractérisation des molécules par un ensemble de descripteurs, nombres réels mesurés ou calculés à partir des structures moléculaires. Il est alors possible d'établir une relation entre ces descripteurs et la grandeur modélisée.

L'utilisation de ces méthodes alternatives à l'expérimentation, parmi lesquelles les relations quantitatives structure propriété/activité (QSPR/QSAR) sont devenues d'un grand intérêt et sont même recommandées dans les nouvelles règlementations [1,2] afin d'obtenir les données nécessaires à l'enregistrement des substances.

Ce travail a pour objectif de développer et d'évaluer le potentiel de tels modèles QSAR (***Q****uantitative* ***S****tructure* ***A****ctivité* ***R****elationship*), QSPR (***Q****uantitative* ***S****tructure* ***P****roperty* ***R****elationship)* pour l'explication et la prédiction d'une série d'hydrocarbures mono-aromatiques toxiques.

Le manuscrit de cette thèse est articulé sur deux parties importantes, chaque partie comporte des chapitres:

La première partie est consacrée à une synthèse bibliographique sur les bases de chimie théoriques utilisées pour le calcul des structures (descripteurs moléculaires).

Dans le 2ème chapitre de cette partie, les principes et les méthodes de développement des modèles QSAR/QSPR, seront introduits. Les étapes de développement des modèles, depuis la préparation de la base des données jusqu'à la validation des modèles en passant par la mise en place des modèles obtenus seront présentés.

Dans la seconde partie de cette thèse, nous présenterons et nous discuterons les résultats obtenus pour les études QSAR/QSPR.

Elle est constituée de 3 chapitres :

- Le premier chapitre présentera une synthèse des connaissances actuelles sur les composés étudiés (les dérivés benzéniques).
- Le deuxieme chapitre sera consacré à l'étude QSPR des indices de rétention de 38 dérivés benzéniques séparés par chromatographie en phase gazeuse prélevés dans la littérature [3].
- Le chapitre 3 portera sur la modélisation de la toxicité de 141 dérivés benzéniques prélevés de la littérature [4] vis-à-vis de *PIMEPHALES PROMELAS* également appelé FATHEAD MINNOW ou tête de boule.

  Enfin, nous terminerons par une conclusion générale.

Références bibliographiques :

[1] Règlement (CE) n° 1907/2006 du Parlement Européen et du Conseil du 18 décembre 2006 concernant l'enregistrement, l'évaluation et l'autorisation des substances chimiques, ainsi que les restrictions applicables à ces substances (REACH), instituant une agence européenne des produits chimiques, modifiant la directive 1999/45/CE et abrogeant le règlement (CEE) n°793/93 du Conseil et le règlement (CE) n° 1488/94 de la Commission ainsi que la directive 76/769/CEE du Conseil et les directives 91/155/CEE, 93/67/CEE, 93/105/CE et 2000/21/CE de la Commission.

[2] N. Margossian, Le règlement REACH - La règlementation européenne sur les produits chimiques, Dunod / L'Usine Nouvelle, Paris, 2008.

[3] JALALI-HERAVI, M. et GARKANI-NEJAD, Z. Prediction of gas chromatographic retention indices of some benzene derivatives. *Journal of Chromatography A*, 1993, vol. 648, no 2, p. 389-393.

[4] Lemond B.Kier and Lowell H.Hall, “Molecular Structure Description”, USA. 1999.

Partie I
Synthèse bibliographique

CHAPITRE 1
Les bases de chimie théorique

I Les bases de chimie théorique

I.1 Introduction :

La recherche et la synthèse de nouveaux composés chimiques sont aujourd'hui souvent associées à une étude par modélisation moléculaire.

La modélisation moléculaire au sens général du terme, fait référence à une représentation plus au moins simplifiée d'un processus. Comme son nom l'indique, la modélisation moléculaire s'inscrit dans un contexte chimique / ou biologique afin de pouvoir simuler ce genre de système et prédire certaines propriétés d'intérêt. Pour ce faire, la modélisation moléculaire se base sur des formalismes mathématiques, plus au moins proches de la réalité physique.

La modélisation moléculaire a pour but de prévoir la structure et la réactivité des molécules ou des systèmes de molécules. Les méthodes de la modélisation moléculaire peuvent être rangées en trois catégories [1]:

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

Dans cette partie, les différentes méthodes théoriques utilisées dans cette thèse pour l'étude des molécules, du niveau électronique au niveau moléculaire, seront présentées. Les méthodes de la chimie quantique [2,3], et les méthodes semi-empiriques seront rappelées, puis, les bases de la mécanique moléculaire [4,5], et la dynamique moléculaires seront décrites.

I.2 Méthodes quantiques :

La mécanique quantique est le prolongement de la théorie des quanta, issue des travaux de Planck, de leur interprétation par Einstein et de leur application à la théorie atomique par Bohr et Sommerfeld. Elle explique la quantification de certaines grandeurs (énergie, moment cinétique) et fait émerger le principe d'exclusion de Pauli. La nouvelle conception des particules qui découle de la dualité onde-corpuscule, explicitée dans les travaux de De Broglie (1923) conduit à la mécanique ondulatoire.

Les méthodes de la mécanique quantique, qui font appel à la distribution des électrons répartis en orbitales autour de la molécule, impliquent des temps de calcul souvent élevés qui limitent leur usage à des petites molécules ou nécessitent le recours à de nombreuses approximations. Elles sont particulièrement adaptées au calcul des charges et des potentiels électrostatiques, à l'approche des mécanismes réactionnels ou à la polarisabilité. L'objectif de la mécanique quantique est principalement de déterminer l'énergie et la distribution électronique [6].

La chimie quantique définie la structure moléculaire comme un noyau autour du quel gravitent des électrons, qui sont décrit par leur probabilité de présence en un point et représentés par des orbitales [7]. Les équations de la chimie quantique sont basées sur la résolution de l'équation de SCHRÖDINGER [8].

I.3 Equation de Schrödinger

Les méthodes de chimie quantique reposent toutes sur le même postulat de départ : tout système peut être décrit par une fonction d'onde. Celle-ci est une fonction des coordonnées des noyaux mais aussi des électrons. Cette fonction est solution de l'équation de Schrödinger.
L'équation de Schrödinger non relativiste et indépendante du temps se présente sous la forme suivante :

$$\hat{H}\Psi = E\Psi \qquad (1.1)$$

où

Ψ est la fonction d'onde décrivant le système de noyaux et d'électrons,

$\hat{H}$ est l'opérateur Hamiltonien relatif à ce même système

E est l'énergie correspondante, valeur propre de l'équation.

L'opérateur Hamiltonien est la somme des différentes contributions à l'énergie du système :

$$E_{Totale} = E_{cinétique}(e) + E_{cinétique}(N) + E_{attraction}(N-e) + E_{répulsion}(e-e) + E_{répulsion}(N-N) \qquad (1.2)$$

L'Hamiltonien rend compte des différentes contributions à l'énergie totale du système à partir d'opérateurs pour les énergies cinétiques des électrons et des noyaux ainsi que les interactions noyau-électron, électron-électron et noyau-noyau.

$$\hat{H} = -\frac{1}{2}\sum_i \frac{1}{2}\nabla_i^2 - \frac{1}{2}\sum_k \nabla_k^2 - \sum_i \sum_k \frac{Z_k}{r_{ik}} + \sum_i \sum_{i<j} \frac{1}{r_{ij}} + \sum_k \sum_{k<l} \frac{Z_k Z_l}{r_{kl}} \qquad (1.3)$$

où

∇^2 est l'opérateur d'énergie cinétique,

Z_k est le numéro atomique de l'atome *k*,

r_{ik} est la distance entre un électron *i* et un noyau *k,*

rij est la distance entre deux électrons *i* et *j*

r_{kl} est la distance entre deux noyaux *k* et *l*.

Il n'est pas possible de résoudre cette équation pour des systèmes d'intérêt chimique (au-delà de H_2), de manière exacte. Il faut alors introduire différentes approximations.

Donc cette équation est actuellement impossible à résoudre sans approximation pour des systèmes « réels ».

I.4 Approximation de Born-Oppenheimer :

L'opérateur Hamiltonien peut être simplifié en utilisant l'approximation de Born-Oppenheimer [9] : les noyaux sont considérés comme fixes par rapport aux électrons. En effet, étant donné que les noyaux ont des masses au moins un millier de fois plus grandes que celle des électrons, on peut considérer que leur mouvement est négligeable. L'énergie cinétique des noyaux devient nulle et l'énergie de répulsion entre noyaux reste constante. L'énergie relative aux noyaux devient donc un paramètre et; l'énergie du système devient la somme de l'énergie électronique et du terme constant de la répulsion nucléaire.

$$E_{Totale} = \mathrm{E}_{el} + \sum_{k} \sum_{k \langle l} \frac{Z_k Z_l}{r_{kl}} \qquad (1.4)$$

L'Hamiltonien du système peut donc être réduit à l'Hamiltonien électronique, tout en se rappelant qu'il faudra ajouter l'énergie relative aux noyaux :

$$\hat{H}_{élec} = -\sum_{i} \frac{1}{2} \nabla_i^2 - \sum_{i} \sum_{k} \frac{Z_k}{r_{ik}} + \sum_{i \langle j} \frac{1}{r_{ij}} \qquad (1.5)$$

Soit de façon plus simple :

$$E_{el} = T_e + V_{Ne} + V_{ee} \qquad (1.6)$$

Avec T_e l'énergie cinétique des électrons, V_{Ne} celle d'attraction noyau-électron et V_{ee} celle de répulsion électro-électron.

L'Hamiltonien électronique peut être exprimé comme la somme d'un terme mono-électronique et d'un terme bi-électronique

$$\hat{H}_{el} = \sum_{i=1}^{N}(-\frac{1}{2}\nabla_i^2 - \sum_{A=1}^{M}\frac{Z_A}{r_{iA}}) + \sum_{i=1}^{N}\sum_{j=1}^{N}\frac{1}{r_{ij}} = \sum_{i=1}^{N}\hat{h}(\vec{r}_i) + \sum_{i=1}^{N}\sum_{j=1}^{N}\frac{1}{r_{ij}} \quad (1.7)$$

.

Cette équation permet le calcul de façon exacte de la fonction d'onde ψ_{el} d'un atome à un électron uniquement. Dans le cas d'un système poly-électronique, on ne peut pas obtenir de solution analytique exacte à l'équation.

I.5 Approximation orbitalaire :

Selon l'approximation orbitalaire, la fonction d'onde électronique peut être décomposée comme le produit de plusieurs fonctions mono-électroniques φ_i (appelées spinorbitales) dans lequel les électrons sont indiscernables.

$$\Psi_{el} = \varphi_1(\vec{r_1})\ \varphi_2(\vec{r_2}).....\varphi_N(\vec{r_N}) \quad (1.8)$$

Pour respecter le principe de Pauli, la fonction doit être antisymétrique, et pour cela le produit est écrit sous la forme de déterminant de Slater (voir l'équation (1.9)). Dans ce déterminant, chaque ligne représente les différentes façons de placer un électron dans les spinorbitales : l'échange de deux électrons correspond à l'échange de deux lignes, ce qui conduit au changement de signe de la fonction d'onde.

$$\Psi^{SD} = \frac{1}{\sqrt{N!}}\begin{vmatrix} \varphi_1(\vec{r_1}) & \varphi_2(\vec{r_1}) & ... & \varphi_N(\vec{r_1}) \\ \varphi_1(\vec{r_2}) & \varphi_2(\vec{r_2}) & ... & \varphi_N(\vec{r_2}) \\ ... & ... & ... & ... \\ \varphi_1(\vec{r_N}) & \varphi_2(\vec{r_N}) & ... & \varphi_N(\vec{r_N}) \end{vmatrix} = \frac{1}{\sqrt{N!}}\left\|\varphi_1(\vec{r_1})\varphi_2(\vec{r_2})...\varphi_N(\vec{r_N})\right\| \quad (1.9)$$

Déterminant de slater

Où r_i est le vecteur de position de l'électron i.

L'énergie associée au déterminant de Slater est la suivante :

$$E_{el} = \langle \psi | H_{el} | \psi \rangle = \left\langle \psi \left| \sum_{i=1}^{N} \hat{h}(\overrightarrow{r_i}) \right| \psi \right\rangle + \left\langle \psi \left| \sum_{i=1}^{N} \sum_{j=1}^{N} \frac{1}{r_{ij}} \right| \psi \right\rangle \quad (1.10)$$

$$E_{el} = \sum_{i=1}^{N} h(\overrightarrow{r_i}) + \sum_{i=1}^{N} \sum_{j=1}^{N} (J_{ij} - K_{ij}) \quad (1.11)$$

Avec Ji l'intégrale bi-électronique coulombienne et KJ l'intégrale bi-électronique d'échange. Une méthode de résolution variationnelle est utilisée, avec la contrainte que les spinorbitales sont orthogonales, pour obtenir l'énergie minimale.

I.6 Méthodes ab-*initio* (Hartree- Fock- Roothann):

Une approche quantique est la méthode Hartree-Fock (HF). Il s'agit de l'approche de Roothaan, une approche variationnelle. Il s'agit donc de trouver les spin-orbitales minimisant l'énergie électronique tout en conservant leur orthogonalité. Une approche itérative dite de champ auto-cohérent (*Self Consistent Fields,* SCF) est donc utilisée.

I.6.1 Equation de Hartree-Fock :

Pour un système moléculaire, au sein duquel les électrons sont appariés (on parle alors de système à couches fermées), l'énergie HF se décompose en une somme de termes mono et bi-électroniques.

$$E^{HF} = 2\sum_{i} h_{ii} + \sum_{i} \sum_{j \succ i} (2J_{ij} - K_{ij}) \quad (1.12)$$

avec

$$h_{ii} = \int \psi_i^*(l)\, \hat{h}_l \psi_i(l) \overrightarrow{dr} \quad (1.13)$$

$$J_{ij} = \int \psi_i^*(1) \psi_j^*(2)\, r_{12}^{-1}\, \psi_i(1) \psi_j(2)\, \overrightarrow{dr_1} \overrightarrow{dr_2} \quad (1.14)$$

Où

$$\overrightarrow{h_1} = -\frac{1}{2} \nabla_1^2 - \sum_{A} \frac{Z_A}{r_{1A}} \quad (1.15)$$

Dans cette expression, *Jij* et *Kij* sont respectivement des intégrales de Coulomb et d'échange, qui caractérisent les répulsions entre électrons. Les intégrales d'échange résultent de la nature antisymétrique de la fonction d'onde multiélectronique.

L'interaction de chaque électron avec les noyaux et les autres électrons environnants est introduite via l'approximation du champ moyen qui considère que chaque électron subit un potentiel V^{eff} formé par les noyaux et le champ moyen des autres électrons.

A chaque électron est alors associée une équation mono-électronique similaire à l'équation de Schrödinger. Ce sont les équations Hartree-Fock.

$$\hat{F}\ \psi_i = \varepsilon_i\ \psi_i \qquad (1.16)$$

où ε_i sont les énergies des spin-orbitales.

Ces dernières sont les valeurs propres de l'opérateur Hamiltonien monoélectronique, $\hat{F}$ appelé opérateur de Fock, dont l'expression pour l'électron 1 est la suivante :

$$\hat{F}\ (1) = \hat{h}_1(l) + V^{eff}(1) = \hat{h}_1(1) + \sum_j (2\hat{J}_j(1) - \hat{K}_j(1)) \qquad (1.17)$$

dans laquelle les opérateurs d'échange *j J*^ et de Coulomb *j K* ^ ont les expressions suivantes :

$$\hat{J}_j\ (1)\ \psi_i(1) = \int \psi_j^*(2)\ r_{12}^{-1} \psi_j\ (2)\ \overrightarrow{dr}_2\ \psi_i(1) \qquad (1.18)$$

$$\hat{K}_j(1)\ \psi_i(1) = \int \psi_j^*(2)\ r_{12}^{-1}\ \psi_i(2)\ \overrightarrow{dr}_2 \psi_j(1) \qquad (1.19)$$

I.6.2 Equations de Roothaan-Hall

La résolution des équations Hartree-Fock, sous forme matricielle, a été proposée par Roothaan et Hall au début des années 1950 [10].

En effet, dans le cadre de l'approximation LCAO (combinaison linéaire d'orbitales atomiques), les équations Hartree-Fock peuvent être réécrites sous la forme suivante (pour l'électron 1) :

$$\hat{F}(1)\sum_v c_{1/i}\,\chi_1(1) = \varepsilon_i \sum_v c_{1/i}\,\chi_v(1) \qquad (1.20)$$

Multiplier à gauche chaque terme de cette équation par φ_μ^* permet de transformer l'équation précédente sous une forme matricielle.

$$\sum_v F_{\mu v} C_{vi} = \varepsilon_i \sum_v S_{\mu v} C_{vi} \qquad (1.21)$$

avec

$$F_{\mu v} = \int \chi_{\mu}^{*}(1)\hat{F}(1)\chi_{v}(1)\ \overrightarrow{dr_1} \qquad (1.22)$$

$$S_{\mu v} = \int \chi_{\mu}^{v}(1)\chi_{v}(1)\overrightarrow{dr_1} \qquad (1.23)$$

Les matrices *F* et *S* ainsi définies sont respectivement les matrices de Fock et de recouvrement. Ces équations, appelées équations de Roothaan-Hall, peuvent s'exprimer sous la forme simplifiée suivante.

$$FC = SC_{\varepsilon} \qquad (1.24)$$

Par orthogonalisation des OM, le problème peut être ramené à la résolution de l'équation *FC=CE*. Or, l'opérateur de Fock dépend des spinorbitales, et donc des solutions de l'équation. Aussi, la résolution du problème passe nécessairement par un processus itératif dit du champ auto-cohérent (SCF, pour *self consistent field*).

A partir d'une géométrie donnée (et en utilisant une base donnée), un jeu d'orbitales moléculaires initial est établi. Ce jeu initial peut être obtenu, par exemple, par calcul semi empirique de type Hückel étendu. La valeur du potentiel HF est alors déterminée, donnant ainsi accès à l'opérateur de Fock. La résolution des équations aux valeurs propres mène ensuite aux énergies *Ei* ainsi qu'à un nouveau jeu d'orbitales moléculaires. A partir de ce nouveau jeu d'orbitales, un nouveau cycle peut alors débuter.

La procédure itérative prend fin lorsque que la variation des énergies devient inférieure à une certaine limite, le critère de convergence de la procédure SCF.

I.6.3 Limites de la méthode de Hartree-Fock :

Dans la plupart des cas, la méthode HF donne des résultats satisfaisants. Malgré tout, certaines limitations ont mené au développement de nouvelles méthodes. Le principal problème posé par cette approche découle du fait que la corrélation existant entre les mouvements des électrons n'est pas prise en compte.

L'énergie de corrélation E_{corr} est ainsi définie comme la différence entre l'énergie exacte E_{exacte} et l'énergie calculée en HF pour une base complète E_{HF}.

$$E_{corr} = E_{exacte} - E_{HR} < 0 \qquad (1.25)$$

D'autres méthodes ont été développées sur des approximations différentes et permettent ainsi la prise en compte de cette corrélation électronique. C'est le cas des

méthodes post-HF (méthodes multiconfigurationnelles, perturbatives ou encore d'agrégats couplés (*coupled clusters*)) [2], qui nécessitent des temps de calculs plus importants, ou la théorie de la Fonctionnelle de la densité (DFT), basée sur une approche différente.

I.7 Théorie de la fonctionnelle de la densité (DFT)

La théorie de la fonctionnelle de la densité (DFT) s'est beaucoup développée ces dernières années. Dans cette approche l'énergie de l'état fondamental d'un système est une fonctionnelle d'une densité électronique tridimensionnelle. L'application du principe variationel donne les équations appelées équations de Kohn-Sham qui sont similaires aux équations de Hartree-Fock. En principe, il suffit de remplacer la contribution d'échange de l'opérateur de Fock par un potentiel d'échange et de corrélation qui correspond à la dérivation de la fonctionnelle d'énergie d'échange et de corrélation par rapport à la densité. Le point crucial en DFT est que l'énergie d'échange et de corrélation n'est pas connue de façon exacte. Néanmoins les formules approchées pour cette énergie donnent des résultats qui sont comparables ou meilleurs que ceux donnés par MP2 à un moindre coût de ressource informatique.

Les premières approximations de la DFT sont similaires à celles appliquées aux méthodes HF. L'équation de Schrödinger est non-dépendante du temps et non-relativiste.
A partir de l'approximation de Born-Oppenheimer le formalisme et les approximations divergent [11].
HyperChem (version 6.02) [12] généralement exécute les calculs d'ab initio (SCF). Il peut aussi calculer l'énergie de la corrélation (peut être ajouté à l'énergie totale par la méthode SCF) par la procédure d'option Hartree-Fock, appelé MP2 qui font le calcul de Møller-Plesset au second ordre de perturbation. La procédure MP2 est disponible pour les calculs de ‘’single point’’. L'énergie de corrélation MP2, peut être ajoutée à l'énergie totale (SCF) à cette configuration du ‘’single point’’.

I.8 Méthodes semi-empiriques:

I.8.1 Introduction :

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

Pour des problèmes d'intérêt chimique impliquant souvent plusieurs dizaines d'atomes, des méthodes simplifiées semi-empiriques, qui négligent ou paramétrisent un grand nombre

d'intégrales, sont couramment utilisées et permettent une reproduction satisfaisante des résultats expérimentaux.

Une méthode semi empirique est une méthode dans laquelle une partie des calculs nécessaires aux calculs Hartree-Fock est remplacé par des paramètres ajustés sur des valeurs expérimentales (l'Hamiltonien est toujours paramétré par comparaison avec des composés référence). En générale toutes ces méthodes sont très précises pour des familles de produits donnés voisines de celles utilisées pour la paramétrisation.

Les calculs semi-empiriques traitent seulement les électrons de valence et utilisent un Hamiltonien plus simple ayant des facteurs de correction basés sur des données expérimentales. L'équation de SchrÖdinger d'un système moléculaire peut être résolue sans approximation *(ab initi*o) ou en introduisant des approximations *(semi-empirique).*

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

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

I.8.2 Définition du semi-empirisme :

Une méthode est semi-empirique si elle admet le cadre de Hartree-Fock-Roothaan, en y incorporant un certain nombre de simplification. On arrive ainsi à réduire considérablement le nombre d'intégrales. En particulier on élimine les intégrales biélectroniques à 3 et 4 centres, qui sont très faibles.

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

Les méthodes semi empiriques ne considèrent que les électrons de la couche de valence ; les électrons des couches internes sont inclus dans le cœur nucléaire.

Toutes les méthodes semi- empiriques modernes sont basées sur l'approche MNDO (Modified Neglect of Differential Overlap) [13] dans laquelle des paramètres sont assignés aux différents types d'atomes puis ajustés de telle sorte à reproduire certaines propriétés comme les chaleurs de formation, les variables géométriques, les moments dipolaires et les énergies de première ionisation.

I.8.3 Quelques méthodes semi-empiriques :

I.8.3.1 Approximation du recouvrement différentiel ; méthode MNDO :

Une des approches des schémas semi-empiriques repose sur une approximation mathématique explicite qui consiste à négliger certains termes de recouvrements différentiels. Introduite en 1953 par Pariser et Parr [14,15], et utilisée la même année par Pople [16], elle permet d'étudier les systèmes conjugués sans tenir compte du squelette σ.

Ces approximations ont été modifiées en 1965 par Pople, Santry et Segal [17], pour être appliquées à tous les électrons de valence de molécules quelconques organiques ou minérales.

L'approximation du recouvrement différentiel, consiste à poser :

$$S_{\mu\nu} = \int \varphi_\mu^A(\overrightarrow{r_1})\ \varphi_\mu^B(\overrightarrow{r_2})\ \overrightarrow{d\tau_1} << 1 \qquad (1.26)$$

La plus simple des approximations est **l'approximation RDN: Recouvrement différentiel Nul (NDO : Neglect of Differential Overlap).**

$$\varphi_\mu\ \varphi_\nu = \delta_{\mu\nu}\ \varphi_\mu\ \varphi_\mu \qquad (1.27)$$

Une autre approximation consiste à négliger **le Recouvrement Différenteiel Diatomique : RDDN (NDDO : Neglect of Diatomic Differentiel Overlap).**

$$\varphi_\mu^A\ \varphi_\nu^B = \varphi_\mu^A\ \varphi_\nu^A\ \delta_{AB} \qquad (1.28)$$

$\delta_{\mu\nu}\,(\delta_{AB})$ représente le symbole de Kronecker.

La méthode MNDO, pour Modified Neglect of Diatomic Overlap, est 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

Cette méthode a été proposée et développée par M.J.S. Dewar et ses collaborateurs [18,20]

I.8.3.2 La méthode AM1 :

L'une des méthodes semi-empiriques les plus utilisées est le modèle AM1 (*Austin Model 1*) [21]. Cette approche emploie un schéma de type NDDO dans lequel les recouvrements des intégrales bi-électroniques mono-centrées sont paramétrés sur des données spectroscopiques pour des atomes isolés, les autres considérants des interactions entre

multipôles. Si cette méthode est en particulier largement utilisée pour les composés organiques, elle présente quelques limitations reconnues dans l'estimation des énergies d'activation, stabilité de certains composés ou enthalpies de liaison [22].

La méthode AM1 permet d'obtenir des chaleurs de formation, des barrières d'activation et des structures géométriques d'équilibre en meilleur accord avec les résultats expérimentaux [23-25] que ne le faisait la méthode MNDO. Ceci est dû essentiellement à l'amélioration apportée à l'énergie de répulsion entre les noyaux ainsi qu'à la différentiation des exposants des orbitales s et ceux des orbitales p (en MNDO les deux exposants sont pris égaux).

L'objectif de cette méthode est de réduire le nombre d'intégrales bi-électroniques à calculer. Elle est fondée sur les approximations suivantes :

- La base d'orbitales utilisée est constituée par les orbitales de Slater des couches de valence.

- Les intégrales de recouvrement sont négligées dans la résolution des équations SCF.

- Toutes les intégrales bi-électroniques, à trois ou quatre centres, sont supposées nulles. En outre, certaines intégrales bi-électroniques, à un ou deux centres, sont également négligées.

- Les termes non diagonaux de la matrice « Hamiltonien de coeur » sont estimés au moyen de relations empiriques qui reposent toutes sur l'hypothèse de proportionnalité de ces intégrales à l'intégrale de recouvrement des orbitales atomiques concernées.

- La plupart des intégrales mono ou bi-électroniques à un centre sont souvent estimées à partir de données spectrales des atomes ou des ions des éléments considérés.

I.8.3.3 Méthode semi-empirique (PM3) :

PM3 (Parametric Method 3) : proposée par Stewart en 1989. Elle utilise une procédure de paramétrisation automatique au cour des calculs.

PM3 est une méthode semi–empirique (SCF) « Self-Consistent Field » pour les calculs chimiques ; il est une paramétrisation de la méthode AM1.

PM3 et AM1 généralement sont les méthodes les plus rigoureuses dans le logiciel HyperChem (7.0), PM3 a été paramétré pour beaucoup des éléments principaux des groupes

et quelques métaux de transition. PM3 est différent d'AM1 seulement dans les valeurs des paramètres.

I.9 Mécanique moléculaire:

I.9.1 Introduction :

Parmi les méthodes dont dispose le chimiste théoricien, les méthodes quantiques et les méthodes dites de champ de force ou **"MÉCANIQUE MOLÉCULAIRE"** qui conduisent à la connaissance de la structure **3D** des molécules.

I.9.2 Définition :

L'expression « Mécanique Moléculaire » désigne actuellement une méthode de calcul qui permet, à priori, d'obtenir des résultats de géométries et d'énergies moléculaires en se basant sur la mécanique classique. La MM est apparu en 1930 [26], mais s'est développée à partir des années soixante, quand les ordinateurs furent plus accessibles et plus performants.

La MM est basée sur l'approximation de Born- Oppenheimer selon laquelle les électrons sont beaucoup plus rapides que les noyaux.

La mécanique moléculaire est une méthode non quantique, mais elle a un intérêt pour les grands systèmes ; comme dans le cas des systèmes biologiques qu'on ne peut aborder avec les méthodes quantiques. Dans ces méthodes, on associe une fonction énergie potentielle à chaque degré de liberté de la molécule : élongation des liaisons, variation des angles de valence, des dièdres (rotation autour d'une liaison). Ces fonctions sont empiriques. L'optimisation de tous les paramètres par minimisation de l'énergie fournit la géométrie d'équilibre des divers conformères et leurs énergies relatives. Pour les molécules possédant un grand nombre de conformères, il existe des procédures automatiques de recherche des minimums locaux d'énergie (recuit simulé) [27].

Les fonctions de potentiel et les paramètres exploités pour l'évaluation des interactions sont désignés par "champ de force".

La mécanique moléculaire a pour but de calculer l'énergie potentielle d'une molécule (ou d'un système de molécules) en fonction des coordonnées des atomes :

$$E_p = \mathrm{f}(\vec{r_1}, \vec{r_2}, ...\vec{r_n}) \qquad (1.29)$$

Où

$\vec{r_i}$ représente le vecteur de la position de l'atome *i*.

La mécanique moléculaire utilise les approximations suivantes :

- chaque atome constitue une particule ;

- l'atome est considéré comme une sphère rigide ayant un rayon et une charge déterminée;
- les énergies sont calculées par des formules dérivées de la mécanique classique [1].

I.9.3 Champs de force :

Compte tenu de la taille des systèmes étudiés en biologie, l'utilisation de la mécanique quantique n'est pas possible. Les macromolécules sont représentées comme un ensemble d'atomes ponctuels dont les interactions sont décrites par un potentiel semi-empirique ou champ de force.

On appelle champ de force le modèle mathématique représentant l'énergie potentielle d'une molécule en mécanique moléculaire.

Le champ de force exprime réellement à la moyenne les interactions électroniques entre les atomes [28].

Le champ de force permet d'accéder à l'hypersurface énergétique d'une molécule en établissant un lien entre les déformations structurales du système et son énergie potentielle. Il désigne à la fois l'équation mathématique (fonction d'énergie potentielle) et les paramètres qui la composent [29]. La fonction d'énergie potentielle définit une énergie empirique, l'énergie totale étant décomposée en une somme de termes additifs représentant chacun des interactions inter atomiques. Elle est exprimée comme une somme de contributions de plusieurs types d'interaction [30,34]. Elle peut se décomposer en terme d'interaction intramoléculaire et un terme d'interaction intermoléculaire.

Les interactions intramoléculaires ne dépendent que des coordonnées internes des molécules c'est-a-dire des liaisons, des angles de valence, et de torsions. En fait pour, affiner l'expression du terme potentielle est rendre plus fidèle la description du système, des termes de couplages entre différents atomes ont été introduit. Le potentielle intramoléculaire peut s'écrire de façon générale.

$$V_{intra} = \sum_{liaison} V_{élongation} + \sum_{angles} V_{courbure} + \sum_{angles\ dièdres} V_{torsion} + \sum V_{croisé} \qquad (1.30)$$

Les interactions intermoléculaires prennent en compte les interactions qui n'interagissent pas par des termes de liaison, d'angle de courbure et d'angle de torsion. Le potentiel non liant s'exprime en deux termes : un terme de Van der Walls et un terme d'énergie électrostatique. On a donc ;

$$V_{\text{int}ermoléculaire} = \sum_{atomes\ \text{non liés}} V_{Van\ \text{der Walls}} + \sum_{atomes\ \text{non liés}} V_{éléctrostatique} \qquad (1.31)$$

I.10 Différents champs de force en mécanique moléculaire :

Différents champs de force utilisent le même type de termes énergétiques mais de manières différentes. Les champs de forces en MM peuvent être groupés en trois classes principales [35] :

- Champs de force contenant uniquement les termes harmoniques.
- Champs de force utilisant les termes d'ordre supérieur (cubique, quadratique,…).
- Champs de force suggérés par *Allinger et col*. [36] ne considérant pas que les termes de la mécanique moléculaire classique mais aussi les effets chimiques comme l'électronégativité.

- ✓ **MM2/MM3/MM4 :**
  MM2, MM3, et MM4 : introduit par Allinger *et al*.[37-40], largement utilisé pour le traitement de petites molécules.
- ✓ **AMBER:** (Assisted Method Building and Energy Refinement) introduit par Cornell *et al*.[41], très largement utilisé dans le traitement des protéines et des acides nucléiques.
- ✓ **CHARMM:** (Chemistry at Harvad Macrmolecular Mechanics) développé par Mackerall, Karplus *et al*. [41] qui est largement utilisé pour la simulation de petites molécules et jusqu‟aux complexes solvatés de grandes macromolécules biologiques
- ✓ **MMFF:** (Merck Molecular Force Field) développé par Halgren [42,43], il est similaire à MM3 dans la forme, mais en diffère par son application focalisée sur les processus de condensation de phases (en dynamique moléculaire). Il reproduit l'exactitude MM3 pour les petites molécules, et est applicable aux protéines et autres systèmes d'importance biologique.

I.10.1 Limite de la mécanique moléculaire:

La mécanique moléculaire présente cependant des limites : par définition, les électrons ne sont pas pris en compte, ce qui rend cette méthode non adaptée aux problèmes dans lesquels les effets électroniques sont prédominants. De plus, les champs de force sont optimisés pour une famille de molécules et ne peuvent pas être généralisés à toutes les molécules.

I.11 LA DYNAMIQUE MOLECULAIRE

I.11.1 Introduction :

Les premiers pas de la dynamique moléculaire n'ont pu se faire que grâce à l'arrivée des premiers ordinateurs (1957) [44]; mais les premières réelles simulations ont été faites par Rahman [45], grâce à ses travaux sur la simulation de l'argon liquide en 1964 avec un temps de simulation de 10^{-11} s, puis de l'eau liquide [46] en 1971.

I.11.2 Principe de la dynamique moléculaire :

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

$$\overrightarrow{F_i} = \mathrm{m}_i a_i = \mathrm{m}_i . \frac{d^2 \overrightarrow{r_i}(t)}{dt^2} \qquad (1.34)$$

$\overrightarrow{F_i}$: Vecteur force agissant sur l'atome i.

mi : masse de l'atome i.

$\overrightarrow{a_i}$: Vecteur accélération de l'atome i.

$\overrightarrow{r_i}$: La position de l'atome i.

Grâce aux vitesses et aux positions de chaque atome dans le temps, il est possible d'évaluer les données macroscopiques, comme l'énergie cinétique et la température. L'énergie cinétique se calcule selon la formule :

$$E_C = \sum_{i=1}^{N} \frac{|p_i|^2}{2m_i} \qquad (1.35)$$

Où pi est la quantité de mouvement de l'atome i.

La température s'obtient à partir de l'énergie cinétique par :

$$E_C = \frac{2K_b T}{2} (3\mathrm{N}\text{-}\mathrm{N}_C) \qquad (1.36)$$

k_b : constante de Boltzmann

N_c : nombre de contrainte

3N-Nc : nombre total de degré de liberté.

La force $\overrightarrow{F_i}$ qui s'exerce sur un atome i se trouvant en position ri(t) est déterminée par dérivation de la fonction potentielle :

$$\overrightarrow{F_i} = \frac{d\overrightarrow{E}(r_i .....r_n)}{dr_i(t)} \quad (1.37)$$

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

ri : coordonnées cartésiennes de l'atome i.

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

$$\overrightarrow{a_i} = \frac{d\overrightarrow{V_i}}{dt} \quad (1.38)$$

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

$$\overrightarrow{V_i} = \frac{d\overrightarrow{r_i}}{dt} \quad (1.39)$$

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

$$r_i(t+\Delta t) = r_i(t) + v_i \Delta t + \frac{1}{2} a_i \Delta t^2 \quad (1.40)$$

I.12 Conclusion :

Cette partie a présenté les bases de la chimie théorique, des méthodes quantiques de l'équation de Schrödinger à la Théorie de la fonctionnelle de la densité DFT, ainsi que les méthodes semi-empiriques, la mécanique moléculaire et la dynamique moléculaire. Ces méthodes nous servirons pour l'optimisation de la géométrie des molécules avant d'utiliser leur structure pour le développement des modèles QSAR/QSPR.

Références bibliographiques :

[1] J. Debord.*Introduction à la modélisation moléculaire*.**2004**, pp.37-41.

[2]C. J. Cramer.*Essentials of computational chemistry - Theories and models*. Wiley,Chicherster, U.K., **2004**.

[3]A. Szabo, N. S.Ostlund.*Modern quantum chemistry : introduction to advanced electronic structure theory*. Dover Publications: Mineola, N.Y., **1996**.

[4]A. R. Leach. *Molecular modelling: principles and applications*. Addison Wesley: Harlow, **1996**.

[5]D. Frenkel, B. Smit. *Understanding molecular simulation : from algorithms to applications*.Academic Press: San Diego, **2002**.

[6]H. Dugas.*Principes de base en modélisation moléculaire, Aspects théoriques et pratiques*.Chapitre 3, Introduction aux méthodes de minimisationd'énergie. 4ème, Ed., Librairie de l'Université de Montréal, **1996.**

[7] D. B. Boyd, K. B. Lipkowitz. *J. Chem. Educ.*, **1982**, 59, 269-274.

[8] E. Schrudinger. *Ann.Phys*., 79, pp.361, 489, 734, Leipzig,**1926**.

[9] M. B. R. Oppenheimer. *Ann. Phys.***1927**, 389, 457-484.

[10] C.C.J. Roothaan, *Rev. Mod. Phys.*, **1951**, 23, 69-89.

[11] N. Vulliermet, *Thèse de Doctorat*, *Université de Genève (Suisse),* **2000**.

[12] M. J. S. Dewar, W. J. Thiel.*J. Am. Chem, Soc*.,**1977**,99, 4899- 4907.

[13] Hyperchem 6.02, Hypercube,**2000**.(http://www.hyper.com).

[14] R. Pariser, R. G. Parr.*J.Chem. Phy*.**1953**, 21, 466-471.

[15] R. Pariser, R. G. Parr.*J.Chem. Phy*.**1953**, 21, 767-776.

[16] J.A. Pople.*Trans. Faraday. Soc*. **1953**, 49, 1375-1385.

[17] J.A. Pople, D.P. Santry, G.A. Segal.*J.Chem. Phys*, **1965**, 43,S129-S135.

[18] M.J.S. Dewar, M. L. McKee.*J.Am.Chem.Soc*. **1977**, 99,5231-5241.

[19] M.J. S. Dewar, H.S. Rzepa. *J.Am.Chem.Soc*. **1978**, 100, 58-67.

[20] L.P.Davis, R.M. Guidry, J.R. Williams, M. J.S. Dewar, H.S. Rzepa.*J. Comp. Chem*, **1981, 2**. 433-445.

[21] M.J.S. Dewar, E.G. Zoebisch, E.F. Healy, J.J. P. Stewart.*J. Am. Chem. Soc.*, **1985**, *107*, 3902-3909.

[22] D.C. Young, *Computational chemistry: A Practical guide for applying techniques to real-world problems*. John Wiley & Sons Inc., New York, **2001**.

[23] R. Volets, J. P. François, J. M. L. Martin, J. Mullens, J. Yperman, L.C. Van Poucke.*J. Comp. Chem*., **1989**, 10 , 449-467.

[24] R. Volets, J. P. François, J. M. L. Martin, J. Mullens, J. Y. perman, L. C. Van Poucke.*J. Comp. Chem*., **1990**, 11, 269-290.

[25] W. J. Welsh.*J. Comput. Chem*., **1990**, 11, 644-653.

[26] D. H. Andrews. *Phys. Rev.*, **1930**, 36, 544-554.

[27] P.Chaquin.*Manuel de chimie théorique :Application à la structure et à la réactivité en chimie moléculaire*.Ellipses,**2000,** pp.190.

.[28] F. Jensen.*Introduction to computational chemistry*. John Wiley & Sons, Chichester, **1999**

[29] G. Monard.*Introduction a la modelisation moleculaire*.Formation continue CNRS-Nancy, **2003.**

[30] J. P .Browen, N. L. Allinger.In: K. B. Boyd (Eds.), *reviews in computational chemistry*.VCH, New York, **1991**, pp.2,81.

[31] J. R. Maple.In: P.V .R. Schleyer (Ed.), *Encyclopedia of computational chemistry*. Wiley, Chichester, **1998**, 2, 1015.

[32] J. Goodman. *Chemical applications of molecular modelling*. Royal society of chemistry, Cambridge, UK.,**1998**.

[33] U. Brkert, N. L. Allinger. *Molecular mechanics*. ACS Monograph 177, ACS, Washington, D. C, **1982**.

[34] P. Comba, T. W. Hambley.*Molecular modelling of inorganic compounds*. VCH, NewYork, **1995**.

[35] U. Dinur, A. Hagler.*Reviews in computational chemistry*. (K. B. Lipkowitz, D. B. Boyd, Eds). VCH, Weinheim,**1991**, 2, 99.

[36] N. L. Allinger, K. Chen, J. A. Katzenellenbogen, S. R.Willson, G. M. Anstead.*J. Comp. Chem*.,**1996**, 17, 747-755.

[37] M. J. S.Dewar, W. J.Thiel.*J. Theoretica Chimica Acta.*, **1977**, 46, 89- 104.

[38] U. N. L.Burkert, D. C.Allinger. *Molecular mechanics*. American chemical society, Washington,**1982**.

[39] N. L.Allinger, Y. H.Yuh, J. H.Lii. *J. Am. Chem. Soc.***1989**, 111, 8551-8565.

[40] N. L.Allinger, K.Chen, J. H.Lii.*J.Comp Chem*.**1996**, 17, 642- 668.

[41] Jr.A. D.Mckerell, D.Bashford, M.Bellott, Jr.R. L.Dunbrack, J. D.Evanseck, M. J.Field, S. Fischer, J. Gao, H. Guo, S. Ha.*J. chem.Phys.*, **1998**, 102, 3586- 3616.

[42] T. A.Halgren. *J Comp. Chem.*, **1996**, 17, 490- 519.

[43] T. A.Halgren, R. B.Nackbar. *J. Comp. Chem.*, **1996**, 17, 587- 615.

[44] B. J. Alder, T. E. Wainwright.*J. Chem. Phys*.,**1957**, 27, 1208-1209.

[45] A. Rahman.*Phys. Rev.*, **1964**, 136, A 405-A 411.

[46] A. Rahman, F. H. Stillinger.*J. Chem. Phys.,* **1971**, 55, 3336-3359.

Chapitre 2
Etude QSAR/QSPR : Principe et méthodologie

II LES ETUDES QSAR/QSPR : Relations quantitatives structures activités/propriétés

II.1 Introduction :

L'utilisation des outils informatiques chez les chimistes est devenue obligatoire afin de bien manipuler les informations moléculaires qui ont été, au cours des dernières années, stockées numériquement sur les ordinateurs dans des bases de données en très grandes quantités. De plus, la multiplication des données exploitables par les chimistes a donné lieu à une obligation de la numérisation, afin d'être capable de stocker, visualiser et traiter ces mêmes données aisément.

Les avancées technologiques de la dernière décennie ont rendu possibles de nombreuses découvertes et applications inaccessibles auparavant. Par exemple, le nombre de composés disponibles dans les études de criblage a augmenté de manière exponentielle. En parallèle, les développements techniques dans le domaine de l'informatique et des technologies de communication ont permis la création de bases de données de composés comportant des millions d'entrées.

Le développement de nouvelles techniques de modélisation a permis la mise en place de nombreuses méthodes RQSP (en anglais QSPR : *Quantitative Structure Property Relationships*) et RQSA (en anglais QSAR : *Quantitative Structure-Activity Relationships*) ; elles reposent pour la plupart sur « la recherche d'une relation entre un ensemble de nombres réels, appelés descripteurs moléculaires, et la propriété ou l'activité que l'on souhaite prédire». Ces méthodes permettent de justifier les données expérimentales disponibles et de prédire les propriétés/activités pour des nouveaux composés ou des composés pour lesquels les données expérimentales ne sont pas disponibles.

Les relations quantitatives structure-propriété (QSPR pour *Quantitative Structure Property Relationships*) sont de plus en plus utilisées, du fait de la croissance des moyens de calculs et de logiciels de calcul permettant le calcul systématique de très nombreux descripteurs moléculaires.

Un modèle QSAR/QSPR relie, d'une manière qualitative ou quantitative, la structure des molécules à une activité ou propriété donnée. La stratégie de développement de tels modèles, en respectant les cinq règles mises en place par l'OCDE (*Organisation de Coopération et de Développement Economique*) pour la validation des modèles RQSA/RQSP (voir plus loin : les principes OECD de validité des modèles RQSA/RQSP), suit les étapes suivantes :

- Constituer la base de données structure – activité (ou propriété) à partir de mesures quantitatives, fiables et normalisées de l'activité (ou propriété) cible, pour chaque composé, et sélectionner des descripteurs moléculaires en relation avec l'activité (ou la propriété) cible afin de traduire de manière numérique la structure des molécules ;
- Diviser ce jeu de données en un jeu d'apprentissage et un jeu de test ;
- Construire des modèles à partir de jeu d'apprentissage à l'aide des méthodes statistiques ;
- Caractériser ces modèles par leurs indices statistiques et par une validation interne ;
- Valider les modèles avec le jeu de test et calculer leur indice de corrélation externe ;
- Répéter l'opération de division pour obtenir d'autres jeux d'apprentissage et de test, et répéter les mêmes étapes (facultative) ;
- Définir le domaine d'applicabilité des modèles proposés afin d'éviter des extrapolations hasardeuses ;
- Explorer et exploiter les modèles validés pour comprendre les mécanismes possibles et faire des prévisions d'activité/propriété pour de nouvelles molécules, si cela est possible.

Dans ce chapitre, une étude bibliographique sur les différentes méthodologies, QSAR/QSPR a été présentée, les différentes étapes de développement, de validation et d'application de ces méthodes sont aussi mises en œuvre.

II.2 Définition :

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

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

II.3 Historique :

Il y a plus d'un siècle et demi, en 1863, *Cros* [1] a observé que le point d'ébullition et le point de fusion des alcanes augmente avec le nombre d'atomes de carbone et la masse moléculaire.

Il a observé également une diminution de la solubilité dans l'eau des alcools avec l'augmentation du nombre d'atomes de carbone et la masse moléculaire, cela est considéré depuis comme la première formulation générale en QSPR.

Cinq ans après, en 1868, *Crum-Brown* et *Fraser* [2] postulèrent que « l'activité biologique d'une molécule est une fonction de sa constitution chimique ».

Quelques décennies plus tard, en 1893, *Richet* [3] a montré que la cytotoxicité de certains composés organiques était inversement proportionnelle à leur solubilité dans l'eau.

A la fin du 19ème siècle, *Meyer* en 1899 et *Overton* en 1901 [4-6], ont indépendamment observé « une relation linéaire entre l'activité des narcotiques et leur coefficient de partage huile-eau ».

Six ans après, en 1907, *Fühner* et *Neubauer* [7] ont montré pour une série de narcotiques homologues, que l'activité augmentait en fonction de la progression géométrique de la série de composés, ceci montrant l'importance de la contribution d'additivité de groupements fonctionnels pour l'activité biologique.

En 1962, *Hansen* [8] a montré l'existence d'une corrélation entre la toxicité des acides benzoïques substitués et les constantes électroniques « *σ* » des substituants.

L'année 1964 est considérée comme le début des méthodes QSAR modernes. *Hansch* et *Fujita* ont établi les premières corrélations entre les propriétés physico-chimiques (log P, pKa, paramètres stériques et électroniques) et l'activité biologique (activité enzymatique, pharmacologique). Ces méthodes seront appelées par la suite l'analyse de *Hansch* et l'analyse de *Free Wilson* [9-10]). Sept ans plus tard, *Hansch* et *Lien* ont réalisé une étude QSAR sur différentes familles d'antifongiques : benzoquinones, sels d'alkylpyridinium, imidazoles et phénols. Ils ont observé que quels que soient la famille et le champignon utilisé, l'activité antifongique dépend du coefficient de partage Eau-Octanol, expérimental ou calculé [11].

Ces études ont été extrapolées aux techniques séparatives en corrélant les propriétés physico-chimiques des analytes avec les temps de rétention obtenus expérimentalement : c'est l'étude quantitative des relations structure temps de rétention noté RQSR [12].

Maintenant, des méthodes 3D comme l'étude CoMFA (*Comparative Molecular Field Analysis*) et CoMSIA (*Comparative Molecular Similarity Indices Analysis*) [13-14] permettent de traiter les relations structure-activité en trois dimensions, 3D-QSAR/QSPR.

II.4 Principe :

Le principe d'une étude QSAR/QSPR (Figure 1), consiste à trouver une relation mathématique reliant de manière quantitative une activité biologique, ou une propriété,

mesurée pour une série de composés similaires dans les mêmes conditions expérimentales, avec des descripteurs moléculaires à l'aide des méthodes statistiques. L'objectif de ces études est d'analyser les données structurales afin de détecter les facteurs déterminants pour l'activité ou la propriété étudiée. Pour ce faire, différents types de méthodes statistiques peuvent être employées.

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

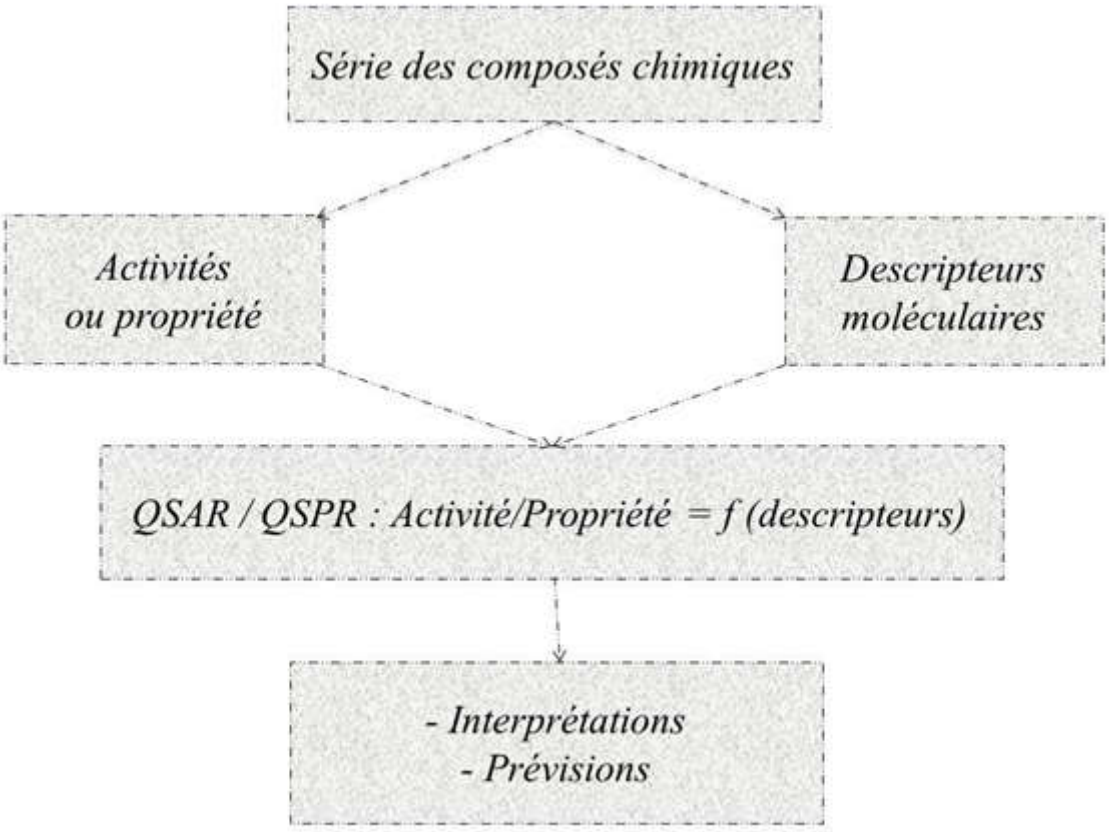


Figure 1: Modèle de l'étude de relation structure activité/propriété **[15]**

Un modèle QSAR/QSPR relie, d'une manière qualitative ou quantitative, la structure des molécules à une activité ou propriété donnée. La stratégie de développement de tels modèles, en respectant les cinq règles mises en place par l'OCDE (*Organisation de Coopération et de Développement Economique*) pour la validation des modèles QSAR/QSPR (les principes OECD de validité des modèles QSAR/QSPR), suit les étapes suivantes :

- Constituer la base de données structure – activité (ou propriété) à partir de mesures quantitatives, fiables et normalisées de l'activité (ou propriété) cible, pour chaque composé, et sélectionner des descripteurs moléculaires en relation avec l'activité (ou la propriété) cible afin de traduire de manière numérique la structure des molécules ;
- Diviser ce jeu de données en un jeu d'apprentissage et un jeu de test ;
- Construire des modèles à partir de jeu d'apprentissage à l'aide des méthodes statistiques ;
- Caractériser ces modèles par leurs indices statistiques et par une validation interne ;
- Valider les modèles avec le jeu de test et calculer leur indice de corrélation externe ;
- Répéter l'opération de division pour obtenir d'autres jeux d'apprentissage et de test, et répéter les mêmes étapes (facultative) ;
- Définir le domaine d'applicabilité des modèles proposés afin d'éviter des extrapolations hasardeuses ;
- Explorer et exploiter les modèles validés pour comprendre les mécanismes possibles et faire des prévisions d'activité/propriété pour de nouvelles molécules, si cela est possible.

II.5 LES DESCRIPTEURS MOLECULAIRES : [15]

II.5.1 Introduction :

De nombreuses recherches ont été menées, au cours des dernières décennies, pour trouver la meilleure façon de représenter l'information contenue dans la structure des molécules, et ces structures elles-mêmes, en un ensemble de nombres réels appelés descripteurs ; une fois que ces nombres sont disponibles, il est possible d'établir une relation entre ceux-ci et une propriété ou activité moléculaire, à l'aide d'outils de modélisation classiques. Ces descripteurs numériques réalisent de ce fait un codage de l'information chimique en un vecteur de réels. Ils peuvent être obtenus de manière empirique ou non-empirique, mais les descripteurs calculés, et non mesurés, sont à privilégier : ils permettent en effet d'effectuer des prédictions sans avoir à synthétiser les molécules, ce qui est un des objectifs de la modélisation.

II.5.2 Définition:

L'une des définitions possible des descripteurs est donnée ci-dessous par Todeschini et Consonni dans le « Handbook of Molecular Descriptor » [16]**:** *The molecular descriptor is the final result of a logic and mathematical procedure which transforms chemical information encoded within a symbolic representation of a molecule into a useful number or the result of some standardized experiment*". Le mot "useful" est à considérer ici avec un double sens : le

descripteur aide à l'interprétation des propriétés moléculaires et/ou il est capable de participer à l'amélioration du pouvoir prédictif. En effet, il est fondamental de prédire mais aussi de comprendre les faits expérimentaux d'un point de vue chimique à partir d'une représentation des molécules par les descripteurs. Ces descripteurs peuvent être obtenus par l'observation des structures, l'utilisation d'algorithmes avec des logiciels tels que Dragon [17], Codessa [18], ISIDA [19] (In SIlico design and Data Analysis) ou être des propriétés physico-chimiques expérimentales.

Les descripteurs numériques réalisent un codage de l'information chimique en un vecteur de réels. Tout simplement, un descripteur moléculaire est une représentation mathématique d'une molécule, qui contient à la fois des informations sur la structure, et donc, implicitement ou explicitement, sur ses propriétés physico-chimiques. Ces informations peuvent être encodées par des valeurs scalaires, des vecteurs ou des chaînes de bits [20,22].

II.5.3 Les types de descripteurs moléculaires :

On dénombre aujourd'hui plus de 10000 descripteurs moléculaires, qui quantifient des caractéristiques physico-chimiques ou structurelles de molécules.

Nous allons présenter les descripteurs moléculaires les plus courants, en commençant par les descripteurs les plus simples, qui nécessitent peu de connaissances sur la structure moléculaire, mais véhiculent peu d'informations. Nous verrons ensuite comment les progrès de la modélisation moléculaire ont permis d'accéder à la structure 3D de la molécule, et de calculer des descripteurs à partir de cette structure.

II.5.3.1 *Les descripteurs 1D*

Les descripteurs 1D : sont accessibles à partir de la formule brute de la molécule (par exemple C_6H_6O pour le phénol), et décrivent des propriétés globales du composé. Il s'agit par exemple de sa composition, c'est-à-dire les atomes qui le constituent, ou de sa masse molaire. On peut remarquer que ces descripteurs ne permettent pas de distinguer les isomères de constitution [20,21].

II.5.3.2 *Les descripteurs 2D :*

sont calculés à partir de la formule développée de la molécule. Ils peuvent être de plusieurs types.

- ***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…

- ***Les indices topologiques*** peuvent être obtenus à partir de la structure 2D de la molécule, et donnent des informations sur sa taille, sa forme globale et ses ramifications. Les plus fréquemment utilisés sont l'indice de Wiener [23], l'indice de Randić [24], l'indice de connectivité de valence de Kier-Hall [25] et l'indice de Balaban [26]. L'indice de Wiener permet de caractériser le volume moléculaire et la ramification d'une molécule : si l'on appelle distance topologique entre deux atomes le plus petit nombre de liaisons séparant ces deux atomes, l'indice de Wiener est égal à la somme de toutes les distances topologiques entre les différentes paires d'atomes de la molécule. L'indice de Randić est un des descripteurs les plus utilisés ; il peut être interprété comme une mesure de l'aire de la molécule accessible au solvant.

Ces descripteurs 2D reflètent bien les propriétés physiques dans la plupart des cas, mais sont insuffisants pour expliquer de façon satisfaisante certaines propriétés ou activités, telles que les activités biologiques. Des descripteurs, accessibles à partir de la structure 3D des molécules, ont pu être calculés grâce au développement des techniques instrumentales et de nouvelles méthodes théoriques.

II.5.3.3 *Les descripteurs 3D*

Les descripteurs 3*D* d'une molécule sont évalués à partir des positions relatives de ses 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 3*D* de la molécule. Ces descripteurs s'avèrent donc relativement coûteux en temps de calcul, mais apportent davantage d'informations, et sont nécessaires à la modélisation de propriétés ou d'activités qui dépendent de la structure 3*D*. On distingue plusieurs familles importantes de descripteurs 3*D* :

- ***Les descripteurs géométriques***, parmi ceux qui sont 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 de 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 et de la plus basse vacante sont des descripteurs fréquemment sélectionnés. Le moment dipolaire,

le potentiel d'ionisation, et différentes énergies relatives à la molécule sont d'autres paramètres importants.

- ***Les descripteurs spectroscopiques*** : les molécules peuvent être caractérisées par des mesures spectroscopiques, par exemples 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. Les spectres infrarouges peuvent être obtenus soit de manière expérimentale, soit par calcul théorique, après recherche de la géométrie optimale de la molécule. Ces spectres sont alors codés en vecteurs de descripteurs de taille fixe. Le descripteur EVA [27] est ainsi obtenu à partir des fréquences de vibration de chaque molécule. Les descripteurs de type *MoRSE* [28] *(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.

II.6 Préparation des données :

II.6.1 Introduction :

Dans de nombreux domaines d'application, il est courant de travailler avec des ”tableaux” de grande dimension qui rendent l'interprétation des données difficile. Une solution envisageable pour simplifier l'analyse et l'interprétation de ces grands tableaux de données serait d'effectuer une sélection permettant de conserver l'information la plus explicative. Selon les données et l'objectif de l'étude, nous chercherons à réduire la dimension en sélectionnant soit des lignes, soit des colonnes.

Le développement d'un modèle prédictif et validé nécessite le partage de la base de données expérimentale en deux jeux distincts. Le jeu d'entraînement va permettre la construction d'un modèle tandis que le jeu de validation sert à calculer la prédictivité du modèle. Ce partage en deux jeux et leur utilité est illustré par la Figure 2. Peduzzi [29] et Babyak [30] estiment que pour obtenir des estimations correctes, le nombre minimal de molécules par descripteur est compris entre 10 et 15.

En effet, lorsque le nombre de descripteurs du modèle est élevé par rapport au nombre de molécules on se trouve souvent face à un problème dit de sur-paramétrisation (ou *overfitting* [31]) : le nombre de descripteurs choisi ne correspond pas au nombre minimal de descripteurs nécessaires (principe de parcimonie). La sur-paramétrisation lors du développement de modèle entraîne généralement des difficultés en termes de prédictivité puisque le modèle n'est

applicable qu'aux molécules utilisées pour l'entraînement. Il n'y a pas de règle concernant la taille optimale du jeu de validation mais un minimum de 10 molécules semble être accepté [32].

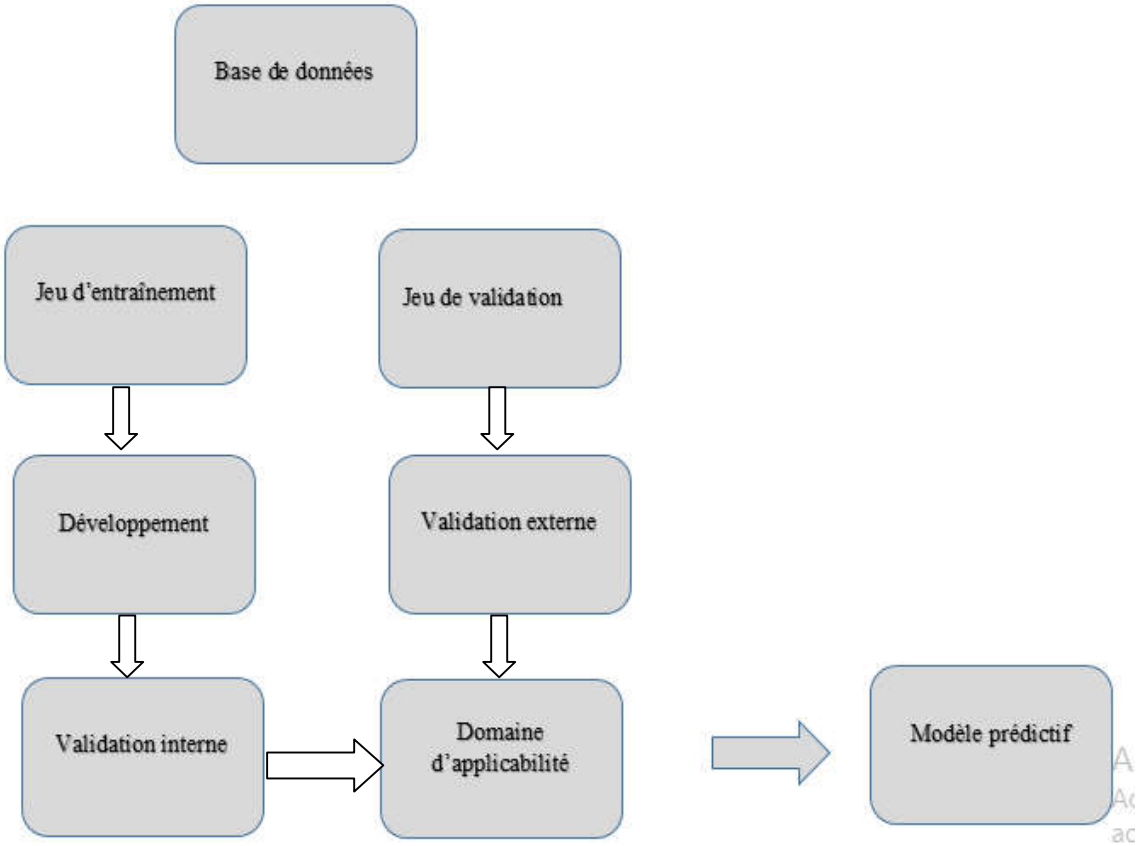


Figure 2: Partage des données expérimentales pour le développement d'un modèle [33]

Dans ce contexte, la sélection aléatoire pour le sous-ensemble d'entraînement (calibration) n'est pas suffisante car elle ne garantit ni sa représentativité de l'ensemble initial ni de lui inclure les points extrêmes dans l'espace des variables qui sont susceptibles de présenter des comportements particuliers qu'il est important de considérer dans l'étape d'apprentissage.

II.6.2 Sélection des points :

Il existe de nombreux algorithmes de sélection de points (les sous ensembles) qui se différencient principalement par leur technique de base : certains reposent sur des calculs de distance alors que d'autres utilisent des clusters de points dans l'espace. Les méthodes basées sur les distances entre points sélectionnent un sous-ensemble de points dans un ensemble initial en considérant les distances entre les points.

L'objectif des méthodes basées sur les clusters est de regrouper les données en clusters et à partir des résultats du "clustering", de choisir les objets représentatifs de l'ensemble initial

pour chaque cluster. Tout d'abord, nous présenterons certaines de ces méthodes, les plus représentatives, avec leur algorithme respectif :

- Kennard et Stone,
- DUPLEX,
- Optisim,
- k-means

II.6.3 Méthodes de sélection de points basées sur les distances

Les méthodes de construction de plans uniformes reposant sur des critères de distance, considèrent généralement des distances euclidiennes.

II.6.3.1 Algorithme de Kennard et Stone (KS)

L'algorithme de Kennard et Stone [34] est une méthode séquentielle qui permet d'extraire un sous-ensemble de N points d'un ensemble de Nc points candidats en D dimensions. A chaque itération, l'algorithme sélectionne le point le plus éloigné des points déjà retenus. L'algorithme peut être décrit ainsi (Algorithme .1) :

Algorithme de Kennard et Stone (Algorithme .1):

Considérer un ensemble de Nc points candidats dans l'espace à D dimensions

Calculer la matrice des distances euclidiennes de l'ensemble des points candidats :

$d_{ij} = \left\| x_i - x_j \right\| = \sqrt{\sum_{r=1}^{D} (x_{ir} - x_{jr})^2}$ = distance euclidienne entre les points iet j

Choisir les points I et J tels que : $d_{IJ} = \max(dij)$

JUSQU'A ce que $N = N_C - N$) points souhaités

- Calculer les distances des ($N_C - N$) points restants par rapport aux N points choisis et retenir la valeur minimale :

$$\Delta_i(N) = \min\{d_{i1}, d_{i2}, d_{i3}, \ldots\ldots d_{iN}\}$$

avec $\Delta_i(N)$, la distance du point candidat i non encore dans la matrice des points sélectionnés au point le plus proche dans les points sélectionnés.

- Pour le $(N+1)^{ème}$ point de la matrice, choisir parmi les ($N_C - N$) points candidats restants celui pour lequel :

$$\Delta(N+1) = \max\{\Delta_i(N)\}$$

pour que le nouveau point soit le plus éloigné des points déjà sélectionés.

FIN

Il existe une variante de l'algorithme de Kennard et Stone [35] qui se différencie par la première étape. En effet, l'algorithme débute par la recherche du point candidat le plus proche du centre du domaine puis de son point le plus éloigné. A partir de ces deux premiers points, la procédure ”classique” de l'algorithme de KS reprend c'est à dire par le troisième point sera le point le plus éloigné du centre de gravité des deux points précédemment sélectionnés.

II.6.3.2 *Algorithme DUPLEX :*

Snee [36] propose l'algorithme DUPLEX qui est une modification de l'algorithme de Kennard et Stone. DUPLEX construit en parallèle deux sous-ensembles de points, sélectionnés par l'algorithme de Kennard et Stone (Algorithme 2). Cet algorithme est souvent utilisé en spectroscopie pour construire les ensembles de calibration et de validation utilisés dans la modélisation des données. Cet algorithme, décrit ci-dessous, débute par la sélection des deux points les plus éloignés dans l'ensemble des points candidats, qui seront assignés à l'ensemble de calibration, puis dans les points restants, les deux points les plus éloignés seront attribués à l'ensemble de validation. L'alternance entre l'ensemble de calibration et l'ensemble de validation est poursuivie jusqu'à ce que tous les points candidats soient assignés à l'un des deux sous-ensembles.

Algorithme DUPLEX (Algorithme 2):

Considérer un ensemble de Nc points candidats dans l’espace à D dimensions

Calculer la matrice des distances euclidiennes de l’ensemble des points candidats

$$d_{ij} = \left\| x_i - x_j \right\| = \sqrt{\sum_{r=1}^{D} (x_{ir} - x_{jr})^2}$$

Choisir les points I et J tels que : $d_{IJ} = \max(dij)$, qui appartiendront au sous-ensemble de calibration.

Parmi les points candidats restants, choisir les points K et L tels que : $d_{KL} = \max(d)$. Les points K et L seront affectés au sous-ensemble de validation.

JUSQU'A ce que $N = N_C - N$) points souhaités

- Calculer les distances des ($N_C - N$) points restants par rapport aux N points choisis du sous-ensemble de calibration:

$$\Delta_i(N) = \min\{d_{i1}, d_{i2}, d_{i3}, \ldots\ldots d_{iN}\}$$

- Pour le $(N+1)^{\text{ème}}$ point du sous-ensemble de calibration, choisir parmi les (Nc−N) points candidats restants celui pour lequel :

$$\Delta(N+1) = \max\{\Delta_i(N)\}$$

- Calculer les distances des (N_C −N) points restants par rapport aux N points choisis du sous-ensemble de validation :

$$\Delta_i(N) = \min\{d_{i1}, d_{i2}, d_{i3}, \ldots\ldots d_{iN}\}$$

- Pour le (N+1)`^{eme} point du sous-ensemble de calibration, choisir parmi les (Nc−N) points candidats restants celui pour lequel :

$$\Delta(N+1) = \max\{\Delta_i(N)\}$$

FIN

La méthode DUPLEX permet de construire en parallèle les sous-ensembles de calibration et de validation qui vont couvrir le domaine mais nous n'avons aucune information quant à la bonne répartition des points dans l'espace des variables et le principe de la méthode peut laisser supposer que les points des deux sous-ensembles seront proches.

II.6.3.3 *Algorithme OptiSim :*

La méthode de sélection OptiSim, proposée par Clark [37], repose sur une procédure de recyclage sur un ensemble de points choisis aléatoirement. Elle nécessite de définir au départ, le nombre N de points à sélectionner, la distance minimale entre deux points et S le nombre de points aléatoirement sélectionnés. Généralement, S est de l'ordre de 5% à 25% du nombre de points initial [35]. Quatre ensembles différents sont utilisés dans l'algorithme : la matrice X des Nc points candidats, la matrice X_N des N points sélectionnés, la matrice X_S des S points sélectionnés aléatoirement et Xr la matrice de recyclage.
L'algorithme OptiSim peut se résumer ainsi (Algorithme 3) :

Algorithme OptiSim (Algorithme 3) :

Définir les valeurs des paramètres : e, N et S.
Choisir le point initial O qui sera le point le plus proche du centre de gravité de l'ensemble des Nc points candidats.
Supprimer le point O de la matrice X.
JUSQU'À obtenir N points dans XN

− JUSQU'À obtenir S points dans X_S ou JUSQU' À ce qu'il n'y ait plus de points candidats

• Choisir aléatoirement un point ei parmi les points candidats de la matrice X.

• Calculer la distance d entre ei et le point *O*.

— SI

— d ≥ e, placer le point e_i dans la matrice X_S.

— SINON

— placer le point e_i dans la matrice Xr.

— FIN

− FIN

− Placer dans X_N le point de X_S associé à la plus grande distance d.

− Tous les points restants dans X_S sont alors transférés vers Xr.

• SI

— X_N ne contient pas N points et que la matrice candidate X est vide, les points de Xr sont transférés vers X

• FIN

FIN

Daszykowski *et al.* [35] proposent une valeur par défaut de ce qui est obtenue en considérant la fraction de points sélectionnés (N/Nc) et le volume V de l'hypersphère en D dimensions formée par le même nombre de points (Nc) que l'ensemble candidat mais uniformément distribués dans l'espace des variables. L'algorithme OptiSim repose sur un choix aléatoire des points à placer initialement dans Xs ou Xr, ce qui conduit à des solutions différentes pour des valeurs de paramètres identiques.

Les sous-ensembles obtenus par l'algorithme OptiSim pour une valeur de N fixée, peuvent présenter des répartitions de points différentes pour des valeurs de paramètres identiques.

II.6.4 Méthodes de sélection de points basées sur les clusters :

II.6.4.1 Méthode des k-means :

La méthode des k-means [38, 39, 40] est aussi une méthode de clustering c'est-à-dire une méthode destinée à former des clusters de points. Elle divise l'ensemble des données en k clusters, avec k fixé par l'utilisateur. Au début de l'algorithme, les objets sont aléatoirement rattachés aux k clusters.

Au cours des itérations les centres de gravité des clusters sont redistribués dans l'ensemble des données afin que les objets similaires appartiennent au même cluster (Algorithme 4)

Algorithme k-means (Algorithme 4):

Définir le nombre de clusters k à trouver

Affecter aléatoirement les points aux k clusters

JUSQU'A stabilisation du critère E

- Calculer le centre de gravité $\bar{}$ xj de chaque cluster,
- Réaffecter chaque point à son centre de gravité $\overline{x}_j$ le plus proche,
- Calculer le critère E défini par :

$$E = \sum_{j=1}^{k} \sum_{i=1}^{N_C} (x_i - \overline{x}_j)^2$$

avec xi les coordonnées du point i et $\overline{x}_j$ les coordonnées du centre de gravité de chaque cluster k

FIN

L'algorithme k-means permet de regrouper les points dans des clusters en fonction de leur proximité. Si nous souhaitons utiliser cet algorithme pour la sélection de points, nous proposons de sélectionner le point le plus proche du centre de gravité de chacun des clusters construits par l'algorithme k-means. L'utilisation de la méthode des k-means pour la sélection de points conduit à des sous-ensembles différents car l'affectation des points aux clusters à la première étape est aléatoire.

II.7 Sélection des descripteurs :

II.7.1 Introduction :

Un grand nombre de descripteurs peut être obtenu 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, des méthodes spécifiques doivent être utilisées afin de réduire le nombre de descripteurs aux descripteurs les plus informatifs. Plusieurs approches sont possibles pour résoudre ce problème :

- Réduire la dimension de l'espace des entrées,
- Remplacer les variables corrélées par de nouvelles variables synthétiques, obtenues à partir de leurs combinaisons,
- Sélectionner les variables les plus pertinentes.

Nous allons maintenant décrire brièvement les méthodes les plus fréquemment utilisées.

II.7.2 Analyse en composantes principales :

L'analyse en composantes principales (ou *ACP*) [41], est une technique d'analyse de données utilisée pour réduire la dimension de l'espace de représentation des données. Contrairement à d'autres méthodes de sélection, celle-ci porte uniquement sur les variables, indépendamment des grandeurs (propriété ou activité) que l'on cherche à modéliser. Les variables initiales sont remplacées par de nouvelles variables, appelées composantes principales, deux à deux non corrélées, et telles que les projections des données sur ces composantes soient de variance maximale. Elles peuvent être classées par ordre d'importance.
Considérons un ensemble de n observations, représentées chacune par p données. Ces observations forment un nuage de n points dans R^p. Le principe de l'ACP est d'obtenir une représentation approchée des variables dans un sous-espace de dimension k plus faible, par projection sur des axes bien choisis; ces axes principaux sont ceux qui maximisent l'inertie du nuage projeté, c'est-à-dire la moyenne pondérée des carrés des distances des points projetés à leur centre de gravité. La maximisation de l'inertie permet de préserver au mieux la répartition des points. Dès lors, les n composantes principales peuvent être représentées dans l'espace sous-tendu par ces axes, par une projection orthogonale des n vecteurs d'observations sur les k axes principaux. Puisque les composantes principales sont des combinaisons linéaires des variables initiales, l'interprétation du rôle de chacune de ces composantes reste possible. Il suffit en effet de déterminer quels descripteurs d'origine leur sont le plus fortement corrélés.
Les variables obtenues peuvent ensuite être utilisées en tant que nouvelles variables du modèle. . Par exemple, la régression sur composantes principales [42] (ou *PCR*) est une méthode de modélisation dont la première étape est une analyse en composantes principales, suivie d'une régression linéaire multiple.

II.7.2.1 La méthode de régression des moindres carrés partiels :

La régression des ***moindres carrés partiels*** [43,44] (MCP, ou PLS) est également une méthode statistique utilisée pour construire des modèles prédictifs lorsque le nombre de variables est élevé et que celles-ci sont fortement corrélées. Cette méthode utilise à la fois des principes de l'analyse en composantes principales et de la régression multilinéaire. Elle consiste à remplacer l'espace initial des variables par un espace de plus faible dimension, sous-tendu par un petit nombre de variables appelées « variable latentes », construites de façon itérative. Les variables retenues sont orthogonales (non corrélées), et sont des combinaisons linéaires des variables initiales. Les variables latentes sont obtenues à partir des variables initiales, mais en tenant compte de leur corrélation avec la variable modélisée, contrairement aux variables résultant de l'analyse en composantes principales. Elles doivent ainsi expliquer le mieux possible la covariance entre les entrées et la sortie. Elles sont alors les nouvelles variables explicatives d'un modèle de régression classique, telles que la régression linéaire multiple.

II.7.2.2 Les méthodes de sélection pas à pas

Ces méthodes de sélection de variables sont utilisées pour choisir le meilleur sous-ensemble de variables explicatives. Parmi ces méthodes, nous pouvons citer :

- la méthode d'élimination progressive ou ”backward selection” : l'algorithme débute en considérant toutes les variables dans le modèle. A chaque étape, la variable associée à la plus grande p-value est éliminée du modèle, si cette valeur est supérieure au seuil fixé a priori (en général 5%). La procédure s'arrête lorsque les variables restant dans le modèle ont toutes une p-value inférieure au seuil.
- la méthode d'introduction progressive ou ”forward selection”: cette méthode est l'inverse de la méthode ”backward”. Lors de la première étape, le modèle ne contient aucune variable. A chaque étape, la variable associée à la plus petite p-value est ajoutée au modèle, si cette valeur est inférieure au seuil fixé a priori. La procédure s'arrête lorsque toutes les variables sont introduites ou lorsque les variables restant dans le modèle ont toutes une p-value supérieure au seuil.
- la méthode de régression pas à pas ou ”stepwise regression” : à chaque étape de la procédure, on examine à la fois si une nouvelle variable doit être ajoutée selon un seuil d'entrée fixé, et si une des variables déjà incluses doit être éliminée selon un seuil de sortie fixé. Cette méthode permet de retirer du modèle d'éventuelles variables qui seraient devenues moins indispensables du fait de la présence de celles nouvellement

introduites. La procédure s'arrête lorsqu'aucune variable ne peut être rajoutée ou retirée du modèle en fonction des critères choisis.

II.7.2.3 Le facteur d'inflation K

Todeschini et al. [45], proposent le facteur d'inflation K (KIF pour K inflation factor) qui est une méthode de réduction de variables reposant sur l'indice de corrélation multivariée K. Le principe de cette méthode repose sur l'idée que la structure d'une base de données est le plus souvent conservée lors de la suppression de la variable q telle que les variables restantes présentent une corrélation multivariée minimale. Ainsi, la suppression de la variable q du jeu de données implique que la corrélation multivariée restante est minimale et qu'elle est issue des variables restantes. La valeur KIFj associée à la j-ème variable est le facteur d'inflation obtenu en considérant la corrélation multivariée totale notée Kp et Kp/j l'indice de corrélation multivariée calculé à partir des données en supprimant la j-ème variable. Les auteurs suggèrent de retenir toutes les variables avec une valeur de KIF supérieure à la limite proposée égale à 0.5.

II.7.2.4 Algorithmes génétiques

Les algorithmes génétiques, initiés dans les années 1970 par John Holland, sont des algorithmes d'optimisation s'appuyant sur des techniques dérivées de la génétique et des mécanismes d'évolution de la nature : croisement, mutation, sélection.

Les algorithmes génétiques (AG) sont des méthodes d'optimisation stochastiques qui font partie des algorithmes évolutionnaires. Les algorithmes évolutionnaires sont des méta-heuristiques qui s'inspirent des mécanismes d'évolution darwinienne des populations biologiques.

Les AGs ont été utilisés pour la première fois en 1950 par des biologistes dans le but de simuler l'évolution des organismes. La première adaptation aux problèmes d'optimisation combinatoire de ces algorithmes a été réalisée dans les années 70 par John Holland [46] en développant une analogie entre un individu dans une population et une solution d'un problème dans un ensemble de solutions potentielles. Des travaux de recherche ont été développés ensuite par Goldberg [47] et Michalewicz [48]. Ces travaux ont montré la possibilité d'appliquer les AGs à des problèmes concrets.

II.7.2.4.1 Principe

Les AGs sont des algorithmes itératifs basés sur la reproduction et l'évolution naturelle des individus en utilisant les principes de la survie des individus considérés comme les plus forts ou les mieux adaptés à l'environnement. Il s'agit alors de combiner les points forts de chaque individu pour en créer de nouveaux de manière à ce que leur efficacité soit meilleure.
Avec les AGs on cherche à optimiser une fonction (objectif) donnée dans un espace de recherche, celui des individus. Pour l'optimiser, on définit une fonction d'évaluation (fitness) reliée à cette fonction objective et appliquée sur chaque individu ou chromosome.
En général, le fonctionnement d'un AG est basé sur les phases suivantes **(Figure 3)** :

II.7.2.4.1.1 Initialisation :

Générer aléatoirement une population initiale de taille N chromosomes.

II.7.2.4.1.2 Evaluation :

Evaluer chaque individu de la population par la fonction d'évaluation appropriée au problème (fonction de fitness).

II.7.2.4.1.3 Reproduction :

Créer une nouvelle population de N chromosomes par l'utilisation d'une méthode de sélection appropriée et l'application d'opérateurs génétiques (croisement et mutation) sur certains chromosomes au sein de la population courante.

II.7.2.4.1.4 Retour

Retour à la phase 2 tant que la condition d'arrêt du problème n'est pas satisfaite.

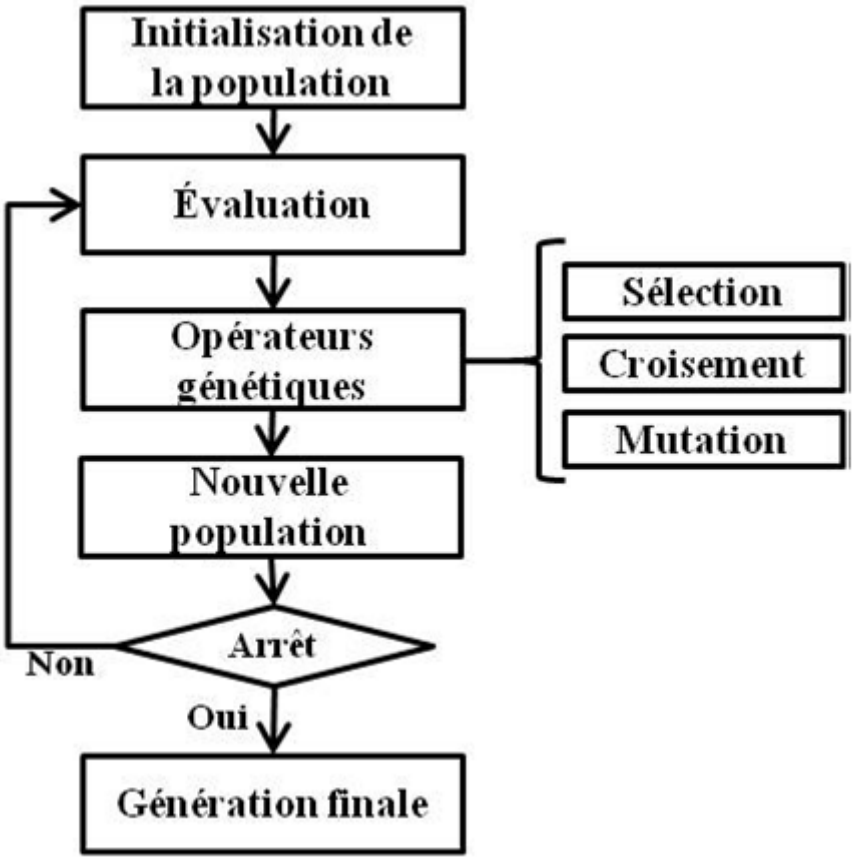


Figure 3: Architecture générale d'un algorithme génétique **[49]**

II.7.2.4.2 La méthode de remplacement (Replacement Method RM) :

Dans le cadre de sélection de descripteurs, il a été introduit une alternative méthode appelée « Méthode de Remplacement » [50]. Cette technique consiste à remplacer une variable choisie de l'ensemble par un autre qui minimise l'écart-type total et c'est la raison de sa désignation « Méthode de Remplacement (RM) »

Un grand ensemble de descripteurs D calculés par Dragon [17], l'objectif est de choisir un sous ensemble optimal de descripteurs $dm = \{Xm1, \text{Xm2}, \text{Xm3},.....\text{Xmd}\}$ avec écart type minimum S :

$$S = \frac{1}{(N - d - 1)} \sum_{i=1}^{N} res_i^2 \qquad (2.1)$$

Où N est le nombre de molécule dans l'ensemble d'apprentissage, et res_i les résidus de la molécule I (la différence entre la valeur expérimentale et la valeur prédite).

Notons que S(dn) est une distribution sur un discret espace de $D!/d!(D-d)$ des points désordonnées.

Le but est de calculer S(dn) sur tous ses points qui nous permettent d'arriver à cet écart type minimum.

La méthode de remplacement RM comprend les étapes suivantes :

- Choisir un ensemble initial de descripteurs d_k au hasard, remplacer l'un des descripteurs X_{ki} avec tous les descripteurs D-d restants, un par un en gardant l'ensemble avec la plus petite valeur de S.
- A partir de cet ensemble résultant, choisir le descripteur avec le plus grand écart type (nous ne considérons pas celui modifié précédemment) et remplacer tous les descripteurs D-d restants, un par un. Répéter cette procédure jusqu'à ce que l'ensemble reste non modifié. Dans chaque cycle, nous ne modifions pas le descripteur optimisé dans le précédent.
- Nous effectuons ce chemin ci-dessus pour tous les chemins possibles i=1,2,3……d et garder le point dm avec le plus petit écart type.

II.7.2.4.3 Conclusion :

Finalement, pour que les relations RQSA/RQSP ne soient pas statistiquement non significatives ou en cas d'erreur ponctuelles, il faut que le rapport composés/descripteurs doive être supérieur à 5 [51, 52].

II.8 DEVELOPPEMENT DES MODELES :

Une fois les descripteurs moléculaires utiles calculés et sélectionnés, nous allons passer au développement d'un modèle QSAR/SQPR. La figue 2 représente un schéma de la mise en place d'un modèle QSAR/QSPR. La recherche et l'optimisation des structures ne sont qu'une étape de la première partie du schéma. La seconde étape constitue le calcul des descripteurs. La deuxième partie du schéma « entraînement, domaine d'applicabilité, validation » explique comment ces descripteurs vont permettre d'obtenir un modèle prédictif.

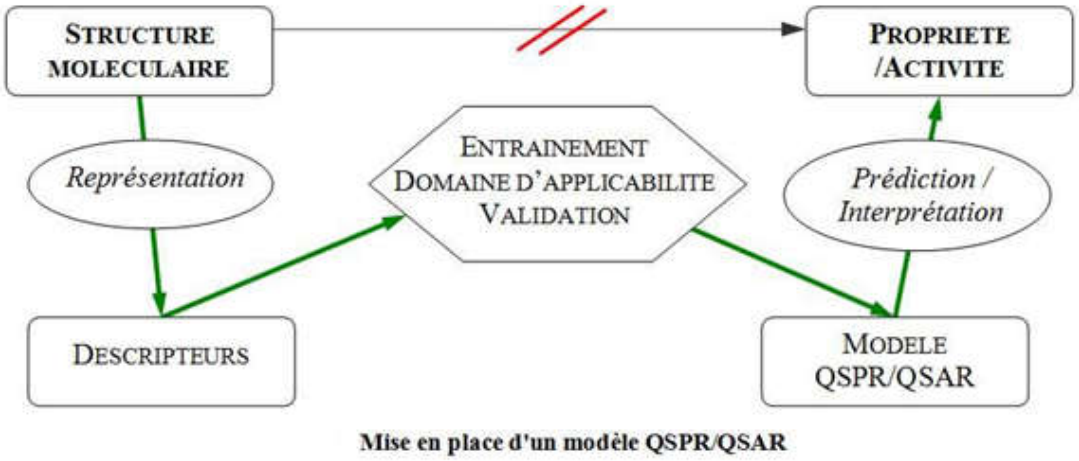


Figure 4: Mise en place d'un modèle QSPR/QSAR [33]

II.8.1 Méthodes statistiques

Faire de la statistique suppose que l'on étudie un ensemble d'objets équivalents sur lesquels on observe des caractéristiques appelées « variables ». Dans notre cas, les objets (ou individus) sont les molécules et les variables sont les descripteurs moléculaires précédemment décrits.

Après le recueil des descripteurs, la démarche statistique consiste à traiter et interpréter les informations recueillies sur ces molécules.

La modélisation vise à fournir un modèle qui soit non seulement ajusté aux données d'apprentissage, mais aussi capable de prédire la valeur de la sortie sur de nouveaux exemples, c'est-à-dire de généraliser.

Pour relier la structure des molécules à la propriété expérimentale, différentes méthodes sont utilisées pour développer des modèles (linéaires ou non linéaires, interprétables ou non), choisir les paramètres les plus pertinents, valider ces modèles (en interne ou en externe) et déterminer leurs domaines d'applicabilité.

Dans cette partie, les méthodes statistiques utilisées au cours de cette thèse seront détaillées (la régression multi-linéaire MLR et les réseaux de neurones artificiels RNA)

II.8.1.1 La régression multi-linéaire (MLR, pour Multiple Linear Regression) [53] :

est la méthode la plus simple et la plus communément employée pour le développement de modèles prédictifs. Elle repose sur l'hypothèse qu'il existe une relation linéaire entre une variable dépendante y (ici, la propriété) et une série de n variables indépendantes xi (ici, les descripteurs).

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

$$Y = a_0 + \sum_{i}^{N} a_i \, x_i \qquad (2.40)$$

Afin de déterminer la valeur des coefficients a_i, nous avons utilisé la méthode des moindres carrés. Elle a pour but de minimiser le carré des résidus ou encore RSS (*Residual Sum of Squared*) c'est-à-dire la somme des carrées des écarts entre les valeurs prédites et les valeurs réelles sur toute la base de données de p molécules (figure 3)

La taille de ces coefficients indique le degré d'influence des descripteurs moléculaires correspondants sur l'activité/propriété cible. Un coefficient positif indique que le descripteur moléculaire correspondant contribue positivement à la cible, tandis qu'un coefficient négatif suggère la contribution négative.

$$RSS = \sum_{i=1}^{p} (y_{\exp,i} - y_{calc,i})^2 \qquad (2.41)$$

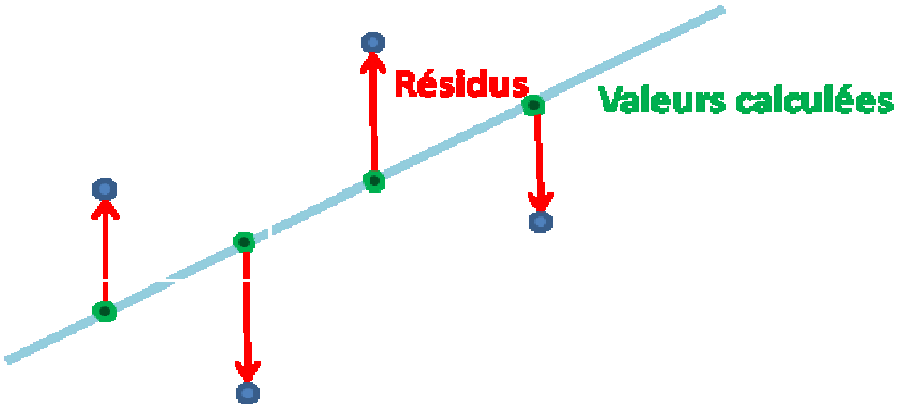


Figure 5: Représentation graphique des résidus [33]

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 RSS. C

$$A = (X^T X)^{-1} X^T Y \qquad (2.42)$$

Le système peut être résolu en utilisant une notation matricielle :

II.8.1.2 *Les* réseaux de neurones artificiels RNA

II.8.1.2.1 Introduction aux réseaux de neurones artificiels RNA :

Depuis une vingtaine d'années, l'utilisation des réseaux de neurones artificiels (RNA) s'est étendue dans des domaines très divers de l'industrie et des services. En géophysique les RNA ont été utilisés pour plusieurs méthodes, par exemples pour détecter les premières arrivées d'ondes sismiques [54 ,55] classifier les différents signaux par l'inversion [56,57] transformer un problème de déconvolution sismique dans un réseau de Hopefield afin de réduire le temps de calcul [58] pour caractériser la distribution de résistivité du sous-sol par l'inversion de données magnétotelluriques [59] et électromagnétiques [60]**.** Ils sont particulièrement utilisés pour résoudre des problèmes de classification, de prédiction, de reconnaissance des formes, de catégorisation, de mémoire associative et d'optimisation [61]**.** Par l'entrainement d'un système non linéaire de multiples variables, les RNA peuvent prédire la variable indépendante [62]. Par conséquent, les RNA constituent une technique d'approximation de systèmes complexes, qui sont difficiles à modéliser par les méthodes statistiques classiques

II.8.1.2.1.1 Principe :

L'approche par réseaux de neurones est analogue aux systèmes de neurones biologiques. Les neurones biologiques permettent de transmettre et de traiter des informations en faisant des signaux électriques dans un réseau constitué d'axones. L'information est propagée d'un neurone à d'autres neurones qui y sont connectés via les synapses.

En modélisation, un neurone possède des entrées par lesquels lui arrivent des données. A chacune de ces entrées est associé un poids w, qui est ajusté au cours de l'apprentissage. Le neurone renvoie un signal de sortie si la somme pondérée des entrées dépasse un certain seuil.

Un réseau de neurones est constitué de multiples couches, une couche d'entrée représentée par les descripteurs, une ou plusieurs couches cachées 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.

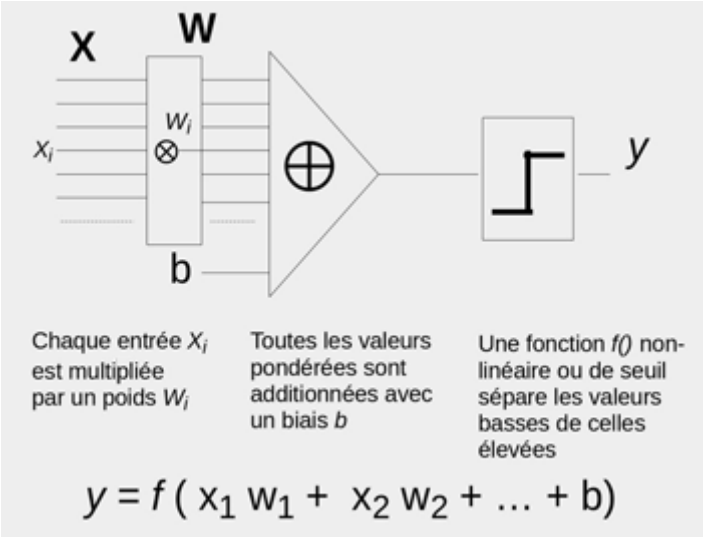


Figure 6: Schéma d'une architechture classique d'un réseau de neurones artificiels

La couche d'entrée comporte autant de neurones que de descripteurs pour le jeu d'apprentissage.

Chaque neurone de la couche cachée réalise des opérations de sommations pondérées, à l'issu desquelles le neurone peut être activé ou non. Chaque neurone de la couche d'entrée est relié par des synapses à chacun des neurones de la couche cachée, et au niveau de ces synapses virtuelles, se trouvent des poids $\mathbf{w_i}$ permettant de moduler l'importance relative de chacun des descripteurs.

II.8.1.2.1.1.1 La couche de sortie :

Comporte autant de neurones que de propriétés modélisées.

Dans notre travail une seule propriété a été modélisée.

Pendant la phase d'apprentissage du modèle par un réseau de neurones, les molécules sont présentées une par une aux neurones de la couche d'entrée. Les poids $\mathbf{w_i}$ associés aux neurones d'entrée sont ajustés itérativement, afin de minimiser l'erreur entre la propriété calculée et la propriété expérimentale.

Beaucoup d'études QSPR employant les réseaux de neurones sont publiées [63] et conduisent à des performances élevées. Pendant le travail de cette thèse, nous utiliserons les réseaux de neurones artificiels RNA dans MATLAB version 7.12 [64].

Matlab est un logiciel de calcul matriciel à syntaxe simple. Avec ses fonctions spécialisées, Matlab peut être aussi considéré comme un langage de programmation adapté pour les problèmes scientifiques. Il est organisé en boîte à outils (*Neural Network Toolbox*

[65]) spécialisés. Cette boîte offre de nombreuses architectures et fonctions d'apprentissage qui permettent de modéliser en toute simplicité des systèmes complexes non linéaires à l'aide de systèmes artificiels. Les applications de cette boite permettent de concevoir, d'effectuer l'apprentissage, de visualiser et simuler le réseau de manière interactive pour ensuite générer le code MATLAB équivalent et ainsi automatiser le processus.

II.8.1.2.1.2 Choix des paramètres du modèle RNA

Le choix des paramètres d'un réseau de neurones dépend principalement du problème à résoudre. Il n'existe donc pas de règle globale pour déterminer avec exactitude les paramètres à adopter pour résoudre un problème donné [66].

Le nombre de neurones en entrée et sortie est généralement déterminé par la taille du fichier à analyser [67,68] Le nombre de couches intermédiaires (couches cachées) variee en fonction de la complexité du problème et de l'objectif recherché (prédiction souhaiter). La distribution de ces données (distribution uniforme, non uniforme, aléatoire,....) dépend du problème à modéliser, toutefois la distribution uniforme est le plus retenu par les concepteurs en absence d'informations préalables des problèmes à modéliser [69,70].

II.9 DEVELOPPEMENT DU MODELE :

II.9.1 Introduction :

Afin d'évaluer l'importance des modèles QSAR/QSPR et, par conséquent, ces capacités de prédiction des activités/propriétés d'autres (nouveaux) composés, la validation des modèles QSAR/QSPR reste une étape très sensible dans les études statistiques. Un modèle étant le résultat d'une analyse statistique, son interprétation et son application doivent se faire dans le cadre très précis du domaine couvert par l'analyse [71].

Pour éviter les erreurs, tant au moment de la validation qu'au moment de l'exploitation, les limites du modèle doivent être clairement établies : le non-robuste du modèle doit être vérifié, les pouvoir de prévision interne et externe doivent être déterminés, et l'espace chimique de l'application du modèle doit être limité.

Dans le but de déterminer la qualité d'un modèle, différents indicateurs statistiques sont employés.

II.9.2 Coefficients et tests statistiques standards :

II.9.2.1 *La racine carrée de l'erreur quadratique RMSE (root mean square error) :*

entre les valeurs des données prédites et expérimentales qui se calcule avec la formule suivante :

$$RMSE = \sqrt{\frac{RSS}{n-p-1}} \qquad (2.43)$$

Où

RSS est le carré des résidus, n le nombre de données, p le nombre de paramètres (i.e. de descripteurs dans l'équation).

Ou encore, l'erreur type résiduel « S »

$$S = \sqrt{\frac{\sum(\hat{y}_1 - y_i)^2}{n-p-1}} \qquad (2.44)$$

II.9.2.2 *Coefficient de détermination R^2 :*

La corrélation peut aussi être mesurée par le coefficient de détermination R^2 qui évalue la part de la variance expliquée par le modèle.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \qquad (2.45)$$

Où

y_i est la valeur expérimentale de la propriété, $\hat{y}_i$ est celle prédite pour propriété, $\bar{y}$ est la moyenne des valeurs expérimentales et n le nombre de molécules.

Plus la valeur de R^2 sera proche de 1 (cas idéal) et plus les valeurs prédites et observées sont corrélées.

II.9.2.3 *Coefficient de détermination ajusté* R^2_{adj} :

Le coefficient de détermination ajusté R^2_{adj} tient compte du nombre de variables et au contraire du R^2, il n'augmente pas forcément lors de l'introduction de variables supplémentaires dans le modèle. On le définit ainsi:

$$R^2_{aj} = 1 - \left[\frac{n-1}{n-p-1} \left(1 - R^2\right) \right] \tag{2.46}$$

II.9.2.4 *L'erreur moyenne absolue (Mean absolute error MAE) :*

est une autre mesure de l'ajustement des valeurs des données calculées avec celles des données expérimentales qui peut également être donnée en pourcentage,

$$MAE = \frac{\sum_{i=1}^{n} \left| y_i - \widehat{y_i} \right|}{n} \tag{2.47}$$

II.9.2.5 *L'indice de Fisher F (test de Fisher) :*

est également couramment employé afin de mesurer le niveau de signifiance statistique du modèle à « x% » (le niveau usuel est 95%), c'est-à-dire la qualité du choix du jeu de paramètres.

$$F = \frac{ESS / p}{RSS / (n-p-1)} \tag{2.48}$$

Où n est le nombre de molécules, p le nombre de descripteurs et ESS est la somme des carrés des résidus due à la régression calculée par l'équation suivante.

$$ESS = \sum_{i=1}^{n} (\hat{y}_i - \overline{y})^2 \tag{2.49}$$

Avec :

F est l'indice de Fisher ; y_i et $\hat{y}_i$ sont, respectivement, les valeurs observées et calculées de la variable dépendante ; $\bar{y}$ est la valeur moyenne des valeurs prédites ; n est le nombre des observations (les molécules) ; p est le nombre de variables dépendantes (les descripteurs).

Plus la valeur de F est grande, plus la probabilité que l'équation soit pertinente augmente. L'équation est considérée comme significative si la valeur F est supérieure à celle tabulée à 95% pour un nombre de degrés de liberté (n-p-1).

II.9.2.6 *Le coefficient du test de Student (t-test) :*

En ce qui concerne la pertinence des descripteurs dans le modèle, elle est également évaluée par le t-test de Student.

Le coefficient du test de Student (t-test) est calculé avec un niveau de confiance de 95% sur les descripteurs obtenus dans l'équation afin de valider leur pertinence. Le descripteur est considéré comme significatif si la valeur t est supérieure à celle tabulée à 95% pour un nombre de degrés de liberté (n-p-1). Le descripteur ayant la valeur absolue de t-test la plus élevée est le plus pertinent.

$$t = \frac{b_i}{S_{b_i}} \qquad (2.50)$$

Où

b_i est le coefficient du descripteur dans l'équation et S_{b_i} la déviation standard du descripteur.

II.9.2.7 *Le facteur d'inflation de la variance VIF :*

C'est un paramètre qui permet de détecter la colinéarité entre les descripteurs utilisés dans un modèle statistique, il est défini par :

$$VIF = \frac{1}{1 - r_i^2} \qquad (2.51)$$

Avec :

r_i^2 est le coefficient de détermination de la régression de la variable x_i sur les autres variables. Plus x_i est linéairement proche des autres variables, plus r_i^2 est proche de 1 et le VIF est grand. L'avantage du VIF par rapport à la matrice de corrélation est qu'il prend en compte des corrélations multiples.

II.10 VALIDATION DU MODELE:

La validation des modèles est une partie intégrante et importante du développement d'un modèle. La validation sert à démontrer que le modèle est prédictif et que les bonnes performances mesurées jusque là ne sont pas dues au sur-apprentissage ou à la chance. Les modèles sont créés sur une partie de la base de données, appelée jeu d'entraînement. Mais avant de pouvoir envisager d'utiliser un modèle, il faut le valider. Il existe plusieurs types de validation qui se complètent : la validation dite interne qui utilise le jeu d'entraînement et la validation externe réalisée sur le jeu de validation.

II.11 Les méthodes de validation :

II.11.1 Validation croisée (validation interne):

La technique la plus employée pour déterminer la stabilité du modèle prédictif est de tester l'influence de chaque échantillon sur le modèle final. Pour ce faire, on emploie une technique de validation croisée (*cross validation* ou CV).

La validation interne ou validation croisée (*cross validation* ou CV) mesure la robustesse du modèle c'est-à-dire sa capacité à rester corrélé à la propriété quand on modifie légèrement les données (suppression d'une ou plusieurs données).

Ce processus consiste à extraire un certain nombre *n* de molécules du jeu initial à *k* molécules et à construire un nouveau modèle avec les *n-k* molécules restantes à l'aide des descripteurs choisis (seules les constantes de la régression changent). Ce nouveau modèle est alors utilisé pour la phase de prédiction sur les *n* molécules retirées. Ce processus est ensuite réitéré pour retirer et prédire les valeurs de toutes les molécules du jeu d'entraînement.

En fonction du nombre de molécules retirées à chaque itération, on parlera de *Leave-One-Out* (LOO) ou de *Leave-Many-Out* (LMO) [72] selon qu'une ou plusieurs molécules est (sont) retirée(s). Dans ce dernier cas, on parle de partitionnement *N-fold* pour indiquer qu'une portion (1/N) du jeu de données est exclue à chaque itération. Ainsi le LOO correspond à une *cross validation* en *k-fold*.

D'une manière générale, ces techniques de validation interne permettent l'évaluation de la robustesse du modèle, autrement dit la stabilité des paramètres du modèle QSAR/QSPR vis-à-vis des molécules du jeu d'entraînement. Cela dit, elles ne permettent en aucun cas de démontrer le pouvoir prédictif des modèles [71,73].

Un coefficient de prédiction LOO ou LMO, désigné par Q^2 ou R^2_{CV}, est calculé à partir de la dispersion des estimations :

$$Q^2 = 1 - \frac{\sum_{i=1}^{n} \left(y_i - \hat{y}_{i/i} \right)^2}{\sum_{i=1}^{n} \left(y_i - \bar{y} \right)^2} \qquad (2.52)$$

$\hat{y}_{i/i}$ désigne la réponse du i^{ème} objet en utilisant un modèle obtenu sans faire intervenir cet i^{ème} objet ;

y_i et $\overline{y}$ représentent, respectivement, la valeur de la i^{ème} observation et la valeur moyenne des n observations ; la sommation porte sur l'ensemble des composés de calibration.

Une faible valeur de Q² implique que le modèle n'est pas robuste et ne sera pas prédictif, mais la réciproque n'est pas nécessairement vraie [73].

En effet, le modèle est considéré comme robuste quand les différents coefficients de validation Q² ont des valeurs très proches et quand la différence entre les Q² et le R² est faible.

Le bootstrap [74,76] est une autre méthode de validation interne qui été utilisée au cours de cette thèse.

Dans la technique de validation par bootstrop on simule de nouveaux échantillons, de taille(n) , par tirages aléatoires avec remise. De cette façon l'ensemble de calibration, qui conserve sa taille initiale(n), se compose, en général, d'objets répétés, l'ensemble d'évaluation rassemblant les objets exclus [77]. Le modèle est calculé sur l'ensemble de calibration et les réponses prédites pour l'ensemble d'évaluation. Cette procédure de construction des ensembles de calibration et d'évaluation est répétée (50 à 2000) fois, et permet d'évaluer l'intervalle de confiance et des estimations statistiques tels que la variance.

Dans le but d'établir que le modèle n'est pas dû au hasard nous avons appliqué le test de randomisation de **Y** (Y-scrambling) [78]. Le 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. Ce procédé est répété plusieurs fois. Si un score élevé est obtenu, le modèle original n'est pas acceptable.

Un bon modèle doit se distinguer, aucun des R² obtenus avec des données mélangées ne doit être supérieur au R² du modèle développé. Les modèles obtenus doivent avoir des performances nettement inférieures à celles du modèle initial. Selon Rücker [79] pour que la probabilité que le modèle ne soit pas dû au hasard soit de 1%, il faut que :

$$R^2 - R_{YS}^2 > \sigma_{YS} \qquad (2.53)$$

Avec R_{YS}^2 la moyenne des R² des modèles fortuits et σ_{YS} l'écart type.

II.11.2 Validation externe ou prédictivité

Les valeurs élevées des paramètres de la validation interne quoique nécessaires sont encore insuffisantes pour valider la qualité d'un modèle. Cependant il faut toujours faire appel à la validation externe.

Afin de tester de manière fiable le pouvoir prédictif d'un modèle, l'utilisation d'un jeu de validation externe, non employé pour le développement, est nécessaire. Pour peu que le jeu de données initial soit suffisamment large, ce dernier peut être aisément divisé en deux : un jeu d'entraînement sur lequel le modèle est développé et un jeu de validation utilisé pour caractériser son pouvoir prédictif.

La mesure de la prédictivité la plus utilisée est le R²ext coefficient de détermination externe [80], ce paramètre est associé à la prédiction de la propriété pour les données du jeu de validation qui n'ont pas été utilisées pour le développement du modèle.

$$R^2_{EXT} = 1 - \frac{\sum_{i=1}^{n_{EXT}} (y - \hat{y}_i)^2}{\sum_{i=1}^{n_{EXT}} (y_i - \bar{y})^2} \qquad (2.54)$$

L'article de Chirico et Gramatica [81] répertorie différents coefficients de calcul de la prédictivité

Le coefficient $Q^2_{F_1}$ coefficient de corrélation de l'ensemble de calibrage (training set) a été proposé par Tropsha [82,73,71], donné par l'équation suivante :

$$Q^2_{F_1} = 1 - \frac{\sum_{i=1}^{n_{EXT}} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n_{EXT}} (y_i - \bar{y}_{TR})^2} \qquad (2.55)$$

Avec

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

En 2008, une autre mesure de la prédictivité, proposée par Schüürmann [83], coefficient de corrélation prédictive de l'ensemble test (test set) $Q^2_{F_2}$, qui se différencie de $Q^2_{F_1}$ par le fait que la moyenne utilisée au dénomination 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 $Q^2_{F_1}$ est plus optimiste [81] car supérieur ou égal à $Q^2_{F_2}$ et par conséquent accepte plus facilement les modèles.

$$Q^2_{F_2} = 1 - \frac{\sum_{i=1}^{n_{EXT}} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n_{EXT}} (y_i - \bar{y}_{EXT})} \quad (2.56)$$

Avec

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

En 2009, le coefficient $Q^2_{F_3}$ capacité prédictive externe, a été proposé par Consonni [84]

$$Q^2_{F_3} = 1 - \frac{\sum_{i=1}^{n_{EXT}} (y_i - \hat{y}_i)^2 / n_{EXT}}{\sum_{i=1}^{n_{EXT}} (y_i - \bar{y}_{TR}) / n_{TR}} \quad (2.57)$$

Avec

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

Un autre paramètre *CCC* [85,86] coefficient de corrélation de concordance, qui mesure à la fois la précision et la justesse (c'est-à-dire à quel point la ligne de la régression dévie de la droite x=y)

$$CCC = \frac{2\sum_{i=1}^{n_{EXT}} (y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})}{\sum_{i=1}^{n_{EXT}} (y_i - \bar{y})^2 + \sum_{i=1}^{n_{EXT}} (\hat{y}_i - \bar{\hat{y}})^2 + n_{EXT}(\bar{y} - \bar{\hat{y}})^2} \quad (2.58)$$

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

II.11.3 Critère de validation :

Parmi les critères de validation les plus utilisés, Tropsha [71] propose d'accéder à la prédictivité du modèle en mesurant les coefficients de détermination lorsque la ligne de régression passe par zéro. Un modèle QSPR possède une capacité de prévision acceptable s'il vérifie les conditions suivantes :

$Q^2_{EXT} > 0,5$

$r^2 > 0,6$

$\left(r^2 - r_0^2\right) / r^2 < 0,1$ ou $\left(r^2 - r_0'^2\right) / r^2 < 0,1$

$0,85 \le k \le 1,15$ ou $0,85 \le k' \le 1,15$

r est le coefficient de corrélation entre les valeurs calculées et expérimentales de l'ensemble de test ; r_0^2 (valeurs calculées en fonction de celles observées) et $r_0'^2$ (valeurs observées en fonction de celles calculées) sont les coefficients de détermination ; k et k' sont les pentes des droites de régression passant par l'origine, respectivement des valeurs calculées en fonction de celles observées, et des valeurs observées en fonction de celles calculées.

II.11.4 Domaine d'application :

Un modèle QSAR s'applique uniquement à des composés similaires à ceux avec lesquels le modèle a été développé. Le domaine d'application [87, 88, 89, 90] du modèle (AD) est l'espace chimique dans lequel le modèle est fiable et peut être interpolé. Il permet de déterminer si le modèle peut être utilisé pour prédire la propriété pour une nouvelle molécule.

En effet, un modèle QSAR n'est pas destiné à être employé en dehors de son domaine d'applicabilité, c'est-à-dire en dehors de l'espace chimique couvert par son jeu d'entraînement. La détermination des AD est donc d'une grande importance surtout si le modèle QSAR devait etre utilisé dans le domaine de la réglementation [91].

Le domaine d'application (AD) est une région théorique de l'espace définie par les descripteurs du modèle et la réponse modélisée, pour lequel un modèle QSAR donné devrait conduire à des prédictions fiables. Cette région qui dépend de la nature des composés de l'ensemble de calibration peut être caractérisée de différentes façons.

Dans ce travail la structure de AD a été vérifiée par l'approche des leviers, définis par les éléments diagonaux de la matrice $\mathbf{H}$ qui permet, par simple multiplication, de passer du vecteur $\mathbf{y}$ au vecteur $\hat{\mathbf{y}}$.

L'élément diagonal h_{ii} est défini par:

$$\boldsymbol{h_{ii}} = \boldsymbol{x_i}\left(\boldsymbol{X^T X}\right)^{-1} \boldsymbol{x_i^T}$$

où

$\boldsymbol{x_i}$ est le vecteur-ligne des descripteurs du composé i, et $\mathbf{X}$ la matrice du modèle déduite des valeurs des descripteurs de l'ensemble de calibration ; l'exposant T désignant le vecteur (ou la matrice) transposé(e).

L'élément h_{ii} détermine l'influence de l'observation i sur les estimateurs obtenus par la méthode des moindres carrés. Un point levier est une observation qui influence considérablement les estimateurs. Dans la pratique, une observation i est considérée comme un point levier si :

$$h_{ii} > h^* = 3\left(\sum_i h_{ii}\right)\Big/ n = 3(p+1)/n$$

II.12 Conclusion :

D'une manière générale, les coefficients R^2 et Q^2 doivent avoir des valeurs proches de 1 (de préférence supérieure 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 R^2) et de l'ordre de la 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.

Références bibliographiques :

[1] A.F.A Cros. *Action de l'alcool amylique sur l'organisme*. Thèse de Doctorat, Faculté de Médecine, Université de Strasbourg, Strasbourg, **1865**.

[2] A.C. C. Brown, T.R. Fraser.*Earth and EnvironmentalScience Transactions of the Royal Society of Edinburgh.*,**1868**,25, 151–203; *Earth andEnvironmental Science Transactions of the Royal Society of Edinburgh.*, **1869**, 25, 693–739.

[3] M.C. Richet.*Comptes rendus des séances de la Société de biologie et de ses filiales*, *Paris,* **1893**, 45, 775–776.

[4] H. Meyer.*Archiv für experimentelle Pathologie und Pharmakologie.*, **1899**, 42, 109–118.

[5] E. Overton.*Studien über die Narkose zugleich ein Beitrag zur allgemeinen Pharmakologie*.Ed. G. Fischer, Jena*,* **1901**.

[6] a) R. L. Lipnick.*Trends in Pharmacological Sciences.*, **1986**, 7, 161–164. b) R. L. Lipnick.*Trends inPharmacological Sciences.*, **1989**, 10, 265–269.

[7] H. Fühner, E. Neubauer.*Archiv fürexperimentelle Pathologie und Pharmakologie.*, **1907**, 56, 333–345.

[8] O. R. Hansen.*Acta Chem. Scand.*, **1962**, 16, 1593–1600.

[9] C. Hansch, T. Fujita.*J. Am. Chem. Soc.*, **1964**, 86, 1616–1626.

[10] S. M. Free, J.W. Wilson.*J.Med. Chem.*, **1964**, 7, 395–399.

[11] C. Hansch, E. J. Lien. *J. Med. Chem.*, **1971**, 14, 653–670.

[12] S. Y. Tham, S. A. Kustrin.*J. Pharm. Biomed. Anal.*, **2002**, 28, 581-590.

[13] R.D. Cramer, D.E. Patterson, J.D. Bunce.*J. Am. Chem. Soc.*, **1988**, 110, 5959-5967.

[14] G. Klebe, U. Abraham, T. Mietzner.*J. Med. Chem.*, **1994**, 37, 4130-4146.

[15] S.Chtita. *Modélisation des molécules organiques hétérocycles biologiquement actives par des méthodes QSAR/QSPR. Recherche de nouveaux médicaments*. Thèse de Doctorat en Chimie, Faculté des Sciences, Université Moulay Ismaïl, **2017.**

[16]R. Todeschini. *Handbook of Molecular Descriptors*. Wiley-VCH: Weinheim, New York, **2000**.

[17] Talete srl. *DRAGON (Software for Molecular Descriptor Calculation)*; Milano, Italy, **2012.**

[18] *CodessaPro*; **2002.**

[19] A. Varnek,D. Fourches,D. Horvath,O. Klimchuk,C. Gaudin,P. Vayer,V. Solov'ev,F. Hoonakker,I. V. Tetko,G. Marcou. *Curr. Comput. Aided Drug Des*.,**2008**, 4, 191–198.

[20] A.Goulon-Sigwalt-Abram. *Une nouvelle méthode d'apprentissage de donnéesstructurées : applications à l'aide à la découverte de médicaments*.Thèse de Doctorat, Paris : Université Pierre et Marie Curie (Paris 6), **2008**.

[21] F.Bonachera. *Les triplets pharmacophoriques flous : développement et applications*. Thèse de Doctorat, Lille : Université Lille1 sciences et téchnologies, **2011**.

[22] M.V.Diudea, I. Gutman, L.Jantschi. *Molecular Topology*. Science publishers, **1999**.

[23] H. Wiener. *J. Chem. Inform. and Comput. Sci.*, **1947**, 69, 17-20.

[24] M. Randić. *J. Am. Chem. Soc.*, **1975**, 97, 6609-6614.

[25]L.B. Kier, L.H. Hall.*Molecular connectivity in chemistry and drug research*. New-York : Academic Press, **1976.**

[26] A.T. Balaban. *Chem.Phys. Lett.*, **1982**, 89, 399-404.

[27] T.W. Heritage*,* Al. Eva. *Perspect. Drug Discov. Des.*, **1998**, 9, 381-398.

[28] J. H. Schuur, P. Selzer,J. Gasteiger. *J. Chem.Inform.Comput. Sci.*, **1996**, 36, 334-344.

[29] P. Peduzzi,J. Concato,A. R.Feinstein, T. R. Holford.*J. Clin. Epidemiol.,***1995**, 48, 1503–1510.

[30] M. A.Babyak. *Psychosom. Med.,***2004**, 66, 411–421.

[31]D. M. Hawkins. *J. Chem. Inf. Model.,***2004**, 44, 1–12.

[32] T. Puzyn,A. M. Szlichtyng,A. Gajewicz,M. Skrzyński, A. P.Worth. J. *Struct. Chem.,***2011**, 22, 795–804.

[33]V. Prana. *Approches Structure-Propriété par le prédiction des propriétés physico-chimiques des substances chimiques*. Thèse de Doctorat, Ecole Doctorale Chimie physique et Analytique de Paris IV, Université Pierre et Marie Curie,**2013**.

[34] R. W. Kennard, L. A. Stone.*Technometrics*, **1969**, 11, 137–148.

[35]M. Daszykowski, B. Walczak, D. L. Massart.Anal. Chim.Acta.,**2002**, 468, 91–103.

[36] R. Snee.*Technometrics*, **1977**, 19, 415–428.

[37] R. Clark.*J. Chem. Inform. Comput. Sci.*, **1997**, 37, 1181–1188.

[38] J. MacQueen. *Some methods for classification and analysis of multivariate observations*. The Regents of the university of California, **1967**.

[39] D. Massart, L. Kaufman.*The interpretation of analytical chemical data by the use of cluster analysis*. Wiley, **1983**.

[40] W. Vogt, D. Nagel, H. Sator.*Cluster analysis in clinical chemistry : a model*. Wiley, **1987**.

[41] I.T.Jolliffe. *Principal Component Analysis*. New-York, NY., Springer, 2ème,Ed., **2002**.

[42] H. Martens, T. Næs. *Multivariate calibration*. Chichester : Wiley, **1989**.

[43] H. Wold. *Estimation of principal components and related models by iterative least squares, in Multivariate Analysis*. Krishnaiaah, P.R., Editor., New York : Academic Press,**1966**, pp. 391-420.

[44] A. Höskuldson. *J. Chemom.*, **1988**, 2, 211-228.

[45] R. Todeschini, V. Consonni, A. Maiocchi.*Chemom. Intell. Lab. Syst.*, **1999**, 46, 13–29.

[46] J. H. Holland.*Adaptation in natural and artificial systems*. The university ofMichigan Press, Ann Arbor, Michigan,**1975**.

[47] D. E. Goldberg. *Genetic algorithms in search, optimization and machinelearning*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 1ère, Ed, **1989**.

[48] Z. Michalewicz. *Genetic algorithms + data structures = evolution programs*. Springer, **1992**

[49] H.Chouaib. *Sélection de caractéristiques : méthodes et application*. Thèse de Doctorat, Université Paris Descartes, **2011**.

[50] P.R.Duchowicz, F.M. Fernandez, E.A. Castro. *Math.Comput. Chem*.,**2006**, 55, 179-192.

[51] P. P. Roy, S. Paul, I. Mitra, K. Roy.*Molecules.*,**2009**, 14, 1660–1701.

[52] J. G. Topliss, R.P. Edwards. *J. Med. Chem.*, **1979**, 22, 1238–1244.

[53] M. Lejeune.*Statistiques : la théorie et ses applications*.Springer-Verlag, Paris, **2004**.

[54] M. Murat, A. Rudman. *Geophys. Prosp*.,**1992**, 40,587-604.

[55] M. McCormack, D. Zaucha,D. *Geophysics*, **1993**, 34, 255-270.

[56]G. Roth, A. Tarantola. *J.Geophys. Res*., **1994**, 99, 811-822.

[57] H. Langer,G. Nunnari, L. Occhipinti. *J. Geophysics*. *Res*., **1996**, 101, 20109-20118.

[58] L. X. Wang.J. M. Mendel. Adaptive m1mmum prediction error deconvolution and source wavelet estimation using Hopfield neural networks. *Geophysics*.,**1992**, 57, 670-679.

[59] Y. Zhang, K. V. Paulson. *Geophys.Prosp*.,**1997**, 45, 725-743.

[60] M. Poulton, K. Sternbeg, C. Glass. *J. Appl. Geophys.*, **1992**, 29, 21-36.

[61] P. J. Drew, J. R. T. Monson.*Surgery*.,**2000**, 127: 3-11.

[62] Z. Huang, J. Shimeld,M. Williamson, J. Katsube. *Geophysics*.,**1996**, 61, 422-436.

[63] N.V. Artemenko, II. Baskin, V. A. Palyulin.*Russian chem. Bull.*, **2003**, 52, 20-29.

[64] MATLAB 7.9.0 (R2009b) and Statistics Toolbox Release, ―The Math Works‖, *Inc., Natick, Massachusetts, United States,* **2011**.

[65] H. Demuth, M. Beale, M. Hagan. *Neural Network Toolbox™ 6 ser's Guide*, Available at:https://filer.case.edu/pjt9/b378s10/nnet.pdf

[66] P. Coulibaly,F. Anctil, B. Bobée. *Real time neural network- based forecasting system for hydropower reservoirs*. Proceedings of the First International Conference on New

Information Technologies for Decision Making in Civil Engineering, (Montreal, 1998), **1998**, 2. École de Technologie Supérieure, Montréal, 1001-1011.

[67] D. E. Rumelhart,E. Hinton, J. Williams, J.*Learning internai representation by error propagation*. Dans Parallel distributed processing. **1986**, l. MIT Press, Cambridge, Massachusetts, 318-362.

[68] M.H. Hassoum. *Fundamentals of artificial neural networks*. MIT Press, Cambridge, **1995**.

[69]Q. J. Zhang, K. C. GUPTA. *Neural Networks for RF and Microwave Design*. Norwood, MA: Artech House, **2000**.

[70] V. K. Devabhaktuni, B. Chattaraj,M. C. E. Yagoub,Q. J. Zhang.*IEEE Trans. Microw. Theory Tech.,***2002**, 51, 1822-1833.

[71] A. Tropsha, P. Gramatica, V.K. Gombar.*QSAR Comb. Sci.*, **2003**, 22, 69–77.

[72] P. Gramatica. *QSAR Comb. Sci.*,**2007**, *26*, 694–701.

[73] A. Golbraikh, A. Tropsha.*J. Mol. Graph. Model.*,**2002**, 20, 269–276.

[74] B. Efron,R. J. Tibshirani. *An Introduction to the Bootstrap*; 1ère,Ed.; Chapman and Hall/CRC, **1994**.

[75] R. Wehrens,H. Putter,L. M. Buydens. *Chemom. Intell. Lab. Syst.,***2000**, 54, 35–52.

[76] B. Efron.*Ann. Stat*.,**1979**, 7, 1–26.

[77] Efron B., Tibshirani R.J., **1993**. An introduction to the bootstrap. Chapman and Hall.456p.

[78] S. Wold, L. Eriksson. *Statistical validation of QSAR results*. In: H. Van de Waterbeemd Ed. Chemometrics methods in molecular design. VCH, New York,**1995**, 2, pp.309-318.

[79] C. Rücker,G. Rücker,M. Y. Meringer.*J. Chem. Inf. Model.,***2007**, 47, 2345–2357.

[80] R. Kiralj,M. M. C.Ferreira. *J. Braz. Chem. Soc*., **2009**, 20, 770-787.

[81] N. Chirico,P. Gramatica. *J. Chem. Inf. Model.,***2011**, 51, 2320–2335.

[82] A. Golbraikh, M. Shen, Z. Y. Xiao,Y. D.Xiao, K. H. Lee. *J. Comput. Aided Mol. Des*.,**2003**, 17, 241–253.

[83] G. Schüürmann,R. U.Ebert, J. Chen,B. Wang,R. Kühne. *J. Chem. Inf. Model*.,**2008**, 48, 2140–2145.

[84]V. Consonni, D. Ballabio, R. Todeschini. *J. Chem. Inf. Model.,***2009**, 49, 1669–1678.

[85] L. I. K.Lin. *Biometrics*.,**1989**, 45, 255–268.

[86] L. I. K. Lin. *Biometrics*.,**1992**, 48, 599–604.

[87] B. Bhhatarai, P. Gramatica. *Environ. Sci. Technol.,***2011**, 45, 8120–8128.

[88] P. P. Roy,S. Kovarich,P. Gramatica. *J. Comput. Chem.,***2011**, 32, 2386–2396.

[89] J. Jaworska, N. N. Jeliazkova, T. Aldenberg, T. *Altern. Lab. Anim.,***2005**, 33, 445–459.

[90] I. I.Baskin, N. Kireeva, A. Varnek. *Mol. Inform.,***2010**, 29, 581–587.

[91] Oecd principles for the validation, for regulatory purposes, of (quantitative) structure-activity relationship models. Document pdf URL : http://www.oecd.org/chemicalsafety /risk-assessment/37849783.pdf(Accédé le 01/02/2018)

Partie II
Les dérivés benzéniques: Etude QSRR et QSPR

Chapitre III
Aperçu sur les molécules étudiées

III LES HYDROCARBURES

Les hydrocarbures, au sens chimique du terme, sont des composés organiques exclusivement constitués d'hydrogène et de carbone, associés sous la forme de molécules d'une grande diversité, de la plus simple, le méthane (CH_4), constituant principal du gaz naturel, aux plus complexes que l'on trouve dans les fractions lourdes des pétroles bruts et dans les schistes bitumineux [1]. Les hydrocarbures font partie de la vie quotidienne, ils sont présents de façon ubiquiste dans l'environnement en raison de la multiplicité de leurs origines à la fois naturelles et anthropiques [2,3]. Leur devenir dans le milieu marin est soumis à un ensemble de processus physico-chimiques et biologiques.

III.1 Les différentes familles d'hydrocarbures

Les hydrocarbures peuvent être regroupés en trois familles : ***(i) les hydrocarbures saturés***, qui ne possèdent pas de double ou de triple liaison ; ils peuvent être linéaires (*n*-alcanes ou paraffines), ramifiés (isoparaffines) ou cycliques (cycloalcanes, appelés aussi naphtènes); ***(ii) les hydrocarbures insaturés***, qui possèdent une à plusieurs double(s) liaisons(s) (alcènes1) ou triple(s) liaisons (alcynes) ; les hydrocarbures (poly)insaturés, alcènes, sont principalement d'origine biogénique, synthétisés par des organismes vivants ; à titre d'exemple, les phytadiènes se retrouvent dans le zooplancton, les mollusques lamellibranches et les poissons [4,5] ; chez le requin, le squalène ($C_{30}H_{50}$), est fabriqué et stocké dans le foie [6,8]; ***(iii) les hydrocarbures aromatiques***, qui comportent un cycle aromatique (composés mono-aromatiques) ou plusieurs cycles aromatiques (composés polyaromatiques) ; une description plus précise des hydrocarbures aromatiques monocycliques sera présentée ultérieurement.

III.1.1 Les hydrocarbures aromatiques monocycliques (CAV ou BTEX)

Les hydrocarbures aromatiques sont d'origine naturelle et leur présence dans l'environnement n'est pas nécessairement d'origine anthropogène [1]. Les BTEX (Benzène, Toluène, Ethylbenzène et Xylène) sont des hydrocarbures monoaromatiques.

Les principales sources naturelles des composés monoaromatiques dans l'environnement sont la dégradation de la lignine [1] et le pétrole brut [9]**.** Par contre, il y a aussi des quantités considérables d'hydrocarbures monoaromatiques qui sont d'origine pétrolière ou pétrochimique [1]. La conséquence de la présence de ces composés dans l'environnement est la pollution des sols et des eaux souterraines engendrée par l'activité humaine en plus de celle dans l'air et dans les eaux de surface.

III.1.2 Définition :

Les hydrocarbures aromatiques volatils constituent une famille de composés contenant un cycle benzénique C6, sur lequel se branche une large variété de substituant. On distingue les substituants suivants :

- Radicaux hydrocarbures aliphatiques (méthyle, éthyle,ect…), qui forment l'essentiel de ce groupe par exemple le toluène.
- Chlore et autres halogènes, donnant notamment la famille des chlorobenzènes
- Alcool OH, qui forment le groupe des phénols
- Nitrate, qui confère des propriétés explosives

III.1.3 Origine :

Les dérivés pétroliers contiennent une petite proportion de dérivés benzéniques. Toutefois, les sources les plus importantes d'hydrocarbures aromatiques sont la carbochimie (cokéfaction et la carbochimie de synthèse) et la chimie organique de base. Les dérivés benzéniques sont très fréquents dans toute l'industrie chimique synthèse, l'industrie des plastiques et celles des colorants. Les propriétés solvantes de ces composés expliquent leur utilisation abondante dans l'industrie mécanique, le traitement de surface, la plasturgie.

III.1.4 Les propriétés physico-chimiques :

Les hydrocarbures monoaromatiques ont certaines propriétés physiques semblables. Ils sont composés uniquement de carbone et d'hydrogène et renferment un anneau de benzène. Les BTEX sont tous liquides, très volatils et sont très inflammables [10]. Ils ont une faible solubilité dans l'eau et une grande solubilité dans les huiles et dans la plupart des solvants organiques [11]**.** Leur solubilité leur confère une bonne mobilité dans les eaux et dans les sols où ils ont des importants impacts environnementaux. Ils sont facilement accessibles aux micro-organismes sous forme solubilisée [1]**.** Les BTEX sont moyennement adsorbés par la phase organique du sol. Leur valeur de coefficient de partage octanol/eau (log Koe) est comprise entre 2 et 4. Si la valeur du coefficient de partage est supérieure à 1, cela signifie que la substance est plus facilement soluble dans les graisses que dans l'eau, tandis que si cette valeur est inférieure à 1, la substance sera plus soluble dans l'eau que dans les graisses [11]**.**

En général les hydrocarbures aromatiques monocycliques, et en particulier les BTEX s'accumulent à la surface des eaux avant d'être partiellement solubilisés. Ils sont volatils et

leur densité de vapeur est plus importante que celle de l'air [9]**.** Les hydrocarbures monoaromatiques sont toxiques pour l'organisme humain et le benzène est cancérogène [1]**.**

III.1.4.1 *LE BENZENE :*

Le benzène (Numéro de registre CAS : 71-43-2) est un composé organique cyclique simple dont la formule moléculaire est C_6H_6. Il est un liquide transparent, volatil, inflammable et incolore à la température ambiante avec une odeur aromatique [12]. Le benzène se mélange avec la plupart des solvants organiques ordinaires. Sa tension de vapeur est de 10,1 à 13,2 kPa à 25 °C et sa solubilité dans l'eau est de 820 à 2 167 mg/L à 25 °C. Le benzène est le plus hydrophile des BTEX avec une valeur de logarithme de son coefficient de partage octano1/eau de 1,56 à 2,69 [13]. Le benzène n'absorbe pas la lumière de façon appréciable à des longueurs d'onde supérieures à 260 nm [14]. Il est mobile dans les sols et est entraîné dans les eaux souterraines par lixiviation [9].

On donne le nom de benzène à l'hydrocarbure pur, réservant le terme de benzols à des mélanges d'hydrocarbures aromatiques (benzène, toluène, xylènes) riches en benzène.

Le benzène (C_6H_6) est le premier terme des hydrocarbures aromatiques (arènes) : composés organiques cycliques insaturés constitués uniquement de carbone et d'hydrogène. C'est un liquide incolore, volatil, à odeur agréable dite aromatique, perceptible dans l'air selon les individus entre 1 et 12 ppm soit un seuil olfactif entre 3.2 et 39 mg/m3.

C'est un des composants des mélanges complexes issus du craquage ou du reformage catalytique d'hydrocarbures pétroliers.

La distillation de ces mélanges permet d'obtenir les composants pratiquement purs et en particulier le benzène.

Les emplois du benzène restent multiples :

- dans l'industrie chimique il sert de matière première de synthèse organique pour la fabrication de nombreux produits d'importance industrielle : plastiques, fibres synthétiques, caoutchouc de synthèse, résines polyesters, solvants, pesticides, colorants, …
- il entre dans la composition des carburants, grâce à ses propriétés antidétonantes susceptibles d'améliorer l'indice d'octane,
- il entre dans la composition de solvants ou diluants,
- il peut être occasionnellement utilisé comme solvant d'extraction (parfumerie par exemple), mais seulement en circuit fermé.

III.1.4.2 *Toluène*

Le toluène (numéro de registre CAS : 108-88-3) est un liquide transparent et incolore; il dégage une odeur sucrée et piquante. Il s'agit d'un composé aromatique monocyclique dont un hydrogène du cycle benzénique a été remplacé par un groupe méthyle (formule moléculaire :
$C_6H_5CH_3$). Le toluène est un liquide volatil qui est inflammable et explosif; il présente une tension de vapeur relativement élevée (3,7 kPa à 25 °C). Le toluène est modérément soluble dans l'eau (535 mg/L à 25 °C) et il est miscible avec la plupart des solvants organiques [15]. Le toluène est moyennement mobile dans les sols et se volatilise rapidement à partir de l'eau ou de la surface du sol [9].
Le toluène est un produit de substitution du benzène depuis de nombreuses années principalement à cause du pouvoir cancérogène de ce dernier. Le toluène peut réagir vivement avec les agents oxydants forts et il peut attaquer certains caoutchoucs et matières plastiques.

Une décomposition thermique peut mener à un dégagement de monoxyde de carbone, de dioxyde de carbone, d'aldéhydes (acétaldéhyde, etc), d'acides carboxyliques (acide acétique, etc) et d'autres composés organiques [16].

III.1.4.3 *Éthylbenzène*

L'éthylbenzène (numéro de registre CAS : 100-41-4) est un liquide clair incolore avec une odeur aromatique caractéristique. Il s'agit d'un composé aromatique monocyclique dont un hydrogène du cycle benzénique a été remplacé par un groupe éthyle (formule moléculaire :
$C_6H_5CH_2CH_3$). Il a la propriété de flotter sur l'eau en raison de sa densité inférieure à celle de l'eau (la densité est 0,87 g/mL), et de sa faible solubilité dans l'eau. Sa tension de vapeur est de 1,273 kPa à 25 °C et sa solubilité dans l'eau est de 161,2 mg/L à 25 °C. L'éthylbenzène est un liquide inflammable. Il s'enflamme facilement en présence de chaleur et d'une source d'ignition. Les vapeurs d'éthylbenzène sont plus lourdes que l'air et peuvent parcourir une grande distance vers une source d'ignition. Ils peuvent, aussi, former un mélange explosif avec l'air [17]
L'éthylbenzène est très mobile dans l'environnement [18]. Il peut se propager par évaporation depuis le sol vers l'atmosphère (en fonction de facteurs comme la température et l'humidité) et par sa fixation rapide dans les sols riches en matières organiques.
Il peut aussi, être entraîné dans les eaux souterraines [18,19]. L'éthylbenzène est très mobile dans l'environnement [18]. Il peut se propager par évaporation depuis le sol vers l'atmosphère (en fonction de facteurs comme la température et l'humidité) et par sa fixation rapide dans les

sols riches en matières organiques. Il peut aussi, être entraîné dans les eaux souterraines [18,19].

III.1.4.4 *Xylènes*

Les xylènes (numéro de registre CAS : 1330-20-7) sont des composés aromatiques monocycliques constitués de deux groupes méthyles liés à un cycle benzénique (formule : $C_6H_4(CH_3)_2$). Trois isomères existent, à savoir : l'*ortho*- ou *o*-xylène (1,2-diméthylbenzène); le *méta*- ou *m*-xylène (1,3-diméthylbenzène); le *para*- ou *p*-xylène (1,4-diméthylbenzène).

Liquides volatils, incolores et transparents et dégageant une odeur aromatique marquée, les xylènes possèdent une pression de vapeur relativement élevée (8,8 à 11,6 kPa à 25 °C), une solubilité modérée dans l'eau (122 à 223 mg/L à 25 °C) et un coefficient de partage octanol eau assez faible (log Koe de 3,08 à 3,29) [13]. Ils sont très inflammables et les propriétés chimiques diffèrent peu d'un isomère à l'autre. Les xylènes sont hydrophobes et solubles dans les solvants non polaires comme l'acétone et l'éthylène. Ils se bioconcentrent peu dans les organismes vivants, car ils ont tendance à s'adsorber aux sols, aux matières en suspension et aux sédiments.

III.1.5 Toxicité :

Le benzène est cancérigène, il induit la formation de lymphomes et de carcinomes suite à une ingestion ou à une inhalation de celui-ci [12]. Le benzène s'accumule dans les tissus adipeux et forme lors de sa dégradation des métabolites toxiques tels que le benzoquinone et I'hexa-2,4diénedial [12]. Le benzène pourrait jouer un rôle dans le déclenchement de la leucémie chez l'homme en plus d'avoir des effets génotoxiques à faible dose [20]. Il peut également avoir un effet embryotoxique à des concentrations de 150 $mg.m^3$ chez le rat. L'inhalation du ben&ne (32 mgm^3) réduit également la réponse immunitaire et induit une dépression du système hématopoïétique [12]. Puisqu'il est jugé toxique pour la santé humaine, le benzène figure sur la liste canadienne des substances toxiques [12]. Quant au Quant au toluène, il a des effets neurologiques réversibles sur les humains à des concentrations variant de 150 à 375 $mg.m^{-3}$ [21]**.** Les xylènes ont des effets similaires à des concentrations supérieures à 1400 $m.gm^{-3}$ [22]. Le toluène et les xylènes n'ont pas été placés sur la liste canadienne des substances toxiques [21,22].

Le benzène à des concentrations de 100 $mg.L^{-1}$ en phase aqueuse est toxique pour la biomasse présente dans les aquifères [23]**.** À de fortes concentrations (200 $mg\text{-}L^{-}$), le toluène peut affecter ***Escherichia coli*** de façon létale [24]. À 30 $mg.L^{-1}$, il inhiberait certaines enzymes, entre autres, chez ***Pselldomonas fluorescens*** [24]. Dans le sol, le toluène aurait des effets

significatifs sur la respiration et l'arnmonification microbiennes des concentrations variant entre 100 et 1300 $mg.kg^{-1}$ [21]**.** Les BTEX et autres solvants dissolvent la membrane cellulaire modifiant l'intégrité cellulaire et affectant la perméabilité et le transport des molécules [25]. Il semble que l'effet toxique du toluène et du xylène augmenterait selon l'âge et l'état des cellules [26].

III.2 Dangers environnementaux des hydrocarbures monoaromatiques

Une substance qui est toxique à de basses concentrations chimiques peut être dangereuse mais cela ne suffit pas pour dire qu'elle présente un danger écotoxicologique [27].

Afin d'estimer le danger écotoxicologique d'une substance, il importe de comparer sa concentration chimique maximale observée dans l'environnement et sa concentration toxique minimale mesurée dans des bioessais [27]. Lorsque la concentration maximale dans l'environnement est supérieure à la concentration toxique minimale, il y a présence d'un danger écotoxicologique. Trois types de dangers doivent être envisagés : les dangers létaux, sublétaux et insidieux qui sont regroupés en deux catégories, les dangers à court terme et les dangers à long terme.

III.2.1 Dangers à court terme

Un danger est létal lorsque la concentration chimique maximale d'un composé dans l'environnement dépasse sa concentration létale aiguë minimale pour des organismes aquatiques ou terrestres, soit sa CL50 en 24 à 96 heures ou sa DL50 lorsqu'il est question de danger pour l'humain [27].

Un danger sublétal est présent lorsque les concentrations environnementales maximales d'une substance sont supérieures à ses concentrations sublétales minimales, soit sa CMEO, CE25, CE50 ou CI50 [27]. Ce danger n'entraîne pas la mort mais a un effet néfaste sur l'organisme exposé. Il peut s'agir d'une inhibition (ralentissement ou blocage d'un processus chimique ou physiologique) ou de tout autre effet.

III.2.2 Dangers à long terme

Un danger à long terme est un danger chronique ou insidieux qui peut se développer lorsqu'une substance chimique s'avère être persistante et peut être bioaccumulée et (ou) lorsque cette substance induit des effets cumulatifs tels que la génotoxicité ou la cancérogénicité [27]. Ce danger est présent lorsque le «toxique dure» ou lorsque «l'effet toxique dure» [27]**.**

Dans le premier cas, il n'y a pas de bioaccumulation des quatre substances considérées car, bien qu'elles puissent être rapidement absorbées en fonction de leur Koe, elles sont vite biotransformées et éliminées. Dans le seconde cas, le tableau 1 expose les effets génotoxiques et cancérogènes des hydrocarbures monoaromatiques pour une exposition au benzène, toluène, éthylbenzène et xylènes. Des dangers toxicologiques marginaux peuvent devenir des risques lors qu'il y a de longues ou fréquentes expositions [27]**.**

Tableau 1: les effets génotoxiques et cancérogènes des hydrocarbures monoaromatiques pour une exposition au benzène, toluène, éthylbenzène et xylènes. **[27]**

Substances chimiques	Effets génotoxiques	Effets cancérogènes	Effets sur la reproduction
Benzène	Chez animal : génotoxique et induit des aberrations chromosomiques et des micronoyaux *in vivo*. Les effets sont établis sur les cellules somatiques et sur les cellules germinales (INRS, 2007a). Chez humain : aucune relation ne peut être actuellement établie (INRS, 2007a).	Chez animal : cancérogène par voie orale et inhalatoire (INRS, 2007a). Les organes cibles sont le système hématopoïétique et différents tissus d'origine épithéliale. Chez humain : cancérogène, groupe 1 des agents cancérogènes (INRS, 2007a).	Chez animal : n'est pas toxique pour le développement. Les données animales montrent des dommages testiculaires (INRS, 2007a). Chez humain : transfert placentaire lors de la grossesse (INRS, 2007a)
Toluène	Chez animal : génotoxique avec des résultats variables in vitro et négatifs in vivo (INRS, 2008). Chez humain : n'est pas génotoxique (INRS, 2008).	Chez animal : n'est pas cancérogène (INRS, 2008). Chez humain : cancérogène, groupe 3 des agents inclassables (CIRC cité par INRS, 2008).	Chez animal : n'altère pas la fertilité. Il est toxique pour le développement à des concentrations non toxiques pour les mères (INRS, 2008). Chez humain : produit pouvant avoir un risque possible sur la fonction de reproduction (INRS, 2008).
Éthylbenzène	Chez animal : n'est pas génotoxique dans la plupart des études *in vitro* et dans toutes les études effectuées *in vivo*. (INRS, 2007b). Chez humain : études insuffisantes, mais il peut causer une augmentation de la génotoxicité (ATSDR, 2007b)	Chez animal et humain : cancérogène dans le groupe 2B des agents qui peuvent être cancérogènes pour l'homme (CIRC cité par INRS, 2007b).	Chez animal : n'est pas toxique pour la fertilité. Il est fœtotoxique à des concentrations toxiques pour les mères (INRS, 2007b). Chez humain : études insuffisantes (INRS, 2007b).
Xylène	Chez animal et humain: n'est pas génotoxique (INRS, 2009)	Chez animal et humain : cancérogène dans le groupe 3 des agents inclassables (CIRC cité par INRS, 2009)	Chez animal : n'est pas toxique pour le développement (INRS, 2009). Chez humain : transfert placentaire lors de la grossesse (INRS, 2009)

Références bibliographiques :

[1] J.P.Vandecasteele. *Microbiologie pétrolière. Concepts, implications environnementales, applications industrielles.* Edition Technip, Paris, **2005**, pp. 796.

[2] M. G. Commendatore.J. L. Esteves. *Mar. Pollut.Bull*., **2004**, 48, 910-918.

[3] J. Volkman, D. Holdsworth, G. Neill, H. Bavor. *Sci. Total Environ*.,**1992***,* 112, 203-219.

[4] M. Blumer, D. W. Thomas.*Science*., **1965**,147, 1148-1149.

[5] V. Grossi,M. Baas,N. Schogt, W.C.M. Klein Breteler, J.W.De Leeuv, J. F.Rontani.*Org. Geochem*.;**1996**, 24, 833-839.

[6] M.J.Bakes, P.D.Nichols. *Comp.Biochem.Physiol*.,**1995**,110B, 267-275.

[7] P. P. Deprez, J. K. Volkman, S. R. Davenport. *Aust. J. Mar. Freshw.Res*.,**1990**, 41, 375-387.

[8] Ó. Fernández, L. Vázquez, G. Reglero, C. F. Torres. *Food Chem.*, **2013**,136, 464-471.

[9] A. Saada, C. Nowak, N. Coquereau. *État des connaissances surl'atténuation naturelle des hydrocarbures*, Rapport intermédiaire, Résultat de la phase 1.Étude réalisée dans le cadre des opérations de Service public du Bureau de recherchesgéologiques et minières (BRGM) 2004 POL A16, **2005**, pp. 110.

[10] Environnement Canada. *Canadian soil quality guidelines for the protection of environnemental and human health : Toluene, Ethylbenzene and Xylenes (TEX)*. Scientific Supporting Documents, Ecosystem Health : Science-based Solutions, National Guidelines and Standards Office, Water Policy and Coordination Directorate, Environment Canada, Ottawa, Report No. 1-9, **2005**, pp. 84.

[11] C. H. Walker, S. P. Hopkin, R. M.Sibly, D. B. Peakall. *Principles of Ecotoxicology*. Taylor & Francis Group, Boca Raton, FL. **2006**, pp. 315.

[12] Gouvernement du Canada. *Loi canadienne sur la protection de l'environnement, Liste des substances d'intérêt prioritaire*. Rapport d'évaluation benzène, Ottawa, **1993**, pp. 40,http://www.hc-sc.gc.ca/ewhsemt/pubs/contaminants/psl1-lsp1/benzene/index-fra.php(Page consultée le 5 février 2010).

[13] D. Mackay, W.Y.Shiu, K.C. MA. *Illustrated handbook of physical-chemical proprieties and environmental fate for organic chemicals*. Vol. 1, Lewis Publishers, Boca Raton, Florida,**1992**.

[14] D. B. Smith,A. Gilbert. *Tetrahedron.*, **1976**, 32, 1309-1326.

[15] Gouvernementdu Canada. *Loi canadienne sur la protection de l'environnement, Liste des substances d'intérêt prioritaire*, Rapport d'évaluation, toluène,Ottawa, **1992**, pp.

29.http://www.hc-sc.gc.ca/ewh-semt/pubs/contaminants/psl1-lsp1/toulene/index-fra.php (Page consultée le 14 février 2010).

[16] CSST (Commission de la santé et de la sécurité du travail). Service du répertoiretoxicologique Toluène, Numéro CAS : 108-88-3, **2004**. http://www.reptox.csst.qc.ca/ Produit.asp?no_produit=1545&nom=Toluene#653 (Page consultée le 25 février 2010).

[17] CSST (Commission de la santé et de la sécurité du travail). Service du répertoire toxicologique, Éthylbenzène, Numéro CAS : 100-41-4,**2007a**.http://www.reptox.csst.qc.ca/Produit.asp?no_produit=3749&nom=%C9thylbenz%E8ne(Page consultée le 7 mars 2010).

[18] Santé Canada. *L'éthylbenzène et la santé*. **2007**. http://www.hc-sc.gc.ca/ewhsemt/pubs/contaminants/ethylbenzene-fra.php (Page consultée le 7 mars 2010).

[19] Environnement*Canada. Recommandation canadienne pour la qualité des sols, l'éthylbenzène*.**2004a**. http://ceqg-rcqe.ccme.ca/download/fr/179/ (Page consultée le 7 mars 2010).

[20] V. Karacic,L. Skender, B. Bosnercucancic, A. Bogadisaee. *Am. J.Ind. Med.*, **1995**, 27, 379-388.

[21] Gouvernement duCanada. *Toluène.* Rapport d'évaluation #4. Loi canadienne sur la proteqtion de l'environnement. **1992**, pp. 29.

[22] Gouvernement du Canada. *Xylènes*.Rapport d'évaluation. LoiCanadienne sur la htection de l'Environnement. **1993b**, pp. 36.

[23] P. J. J Alvarez,P .J. Anid,L. Cohen.*Bidemadation*., **1991**, 2, 43-51.

[24] Environnement Canada. *Le toluène*. Direction des services techniques,Sevice de la protection de l'environnement. **1984**, p. 89.

[25] H.J. Heipieper,F. J. Weber,J. Sikkema, H. Keweloh, J. Debont. *Trends Biotechnol*.,**1994**, 12, 409-415.

[26] S.H.Kong, D.L.Johnstone. *Biotechnol. Lett*.,**1994**,16, 1217- 1220**.**

[27] R. V. Coillie. *Analyse de risques écotoxicologiques* (ENV-789). Recueil de notes, Centre de formation universitaire en environnement, Université de Sherbrooke, Sherbrooke, **2007**, pp. 408.

Chapitre IV
Modélisation des indices de rétention

IV La chromatographie

La chromatographie est une méthode physique d'analyse basée sur la séparation de constituants d'un mélange; ces derniers appelés solutés sont séparés et entraînés par un fluide (un liquide ou gaz) que l'on appelle phase mobile; La séparation est basée sur l'entraînement différentiel des constituants du mélange. Ces derniers parcourent la phase stationnaire avec des temps proportionnels à leurs propriétés intrinsèques (taille, structure, ...) ou à leur affinité avec la phase stationnaire (polarité, ...).

- **Optimisation d'une analyse chromatographique**

La résolution et le temps d'élution sont les deux variables dépendantes les plus importantes à considérer. Dans toute optimisation, le but est de réussir une séparation suffisante des composés intéressants en un minimum de temps. Dans la pratique, on s'efforcera d'abord de choisir les conditions chimiques de la séparation (nature et composition chimique de la phase stationnaire, nature et composition chimique de l'éluant, modification éventuelle des produits à séparer) pour que le facteur de sélectivité ne soit pas trop proche de 1 et pour que les facteurs de capacité soient compris entre 1 et 10.

- **La chromatographie en phase gazeuse :**

La CPG est une méthode d'analyse par séparation qui s'applique aux composés gazeux ou susceptibles d'être vaporisés par chauffage sans décomposition [1]. Cette technique s'applique donc aux molécules de bas poids moléculaires (PM < 500 g. mol-1) et aux composés stables avec la température. Pour les composés thermolabiles ou peu volatils, l'analyse ne sera possible qu4après des réactions de transformation (**dérivatisation**). Dans cette technique chromatographique :

• La phase stationnaire est un liquide à haut point d'ébullition (**G/L**), ou solide dans ce cas, la phase stationnaire est un polymère poreux (**G/S**). • La phase mobile est un gaz qui balaie en permanence la colonne et qui est encore appelé gaz vecteur ou gaz porteur. Cette chromatographie permet la séparation des produits légers et en particulier des gaz.

• La phase mobile est un gaz inerte (N_2, He, Ar,...)

• Les solutés à analyser sont injectés à l'état pur ou en solution. Les quantités de produit injectées sont de l'ordre de quelques µl (2 à 3 µl).

Chaque constituant est caractérisé par des indices calculés à partir d'une gamme d'alcanes ou plus rarement d'esters méthyliques linéaires, à température constante (indice de Kováts) [2] ou en programmation de température (indices de rétention) [3]. Les temps de rétention, bien que

spécifiques d'un composé, ont tendance à varier d'une analyse à l'autre, notamment du fait du vieillissement des colonnes.

Ces derniers caractérisent la rétention d'un composé d'une famille chimique déterminée, avec une phase stationnaire donnée, en isotherme ou gradient linéaire.

Les indices de Kovats sont déterminés soit graphiquement (en traçant le logarithme des temps de rétention d'une série homologue en fonction du nombre de carbone), soit par le calcul **(droite et formule de Kovats pour le calcul des indices de rétention).**

$$I_k = 100\left[n + \frac{t_R(x) - t_R(C_n)}{t_R(C_{n+1}) - t_R(C_n)} \right]$$

n : nombre d'atomes de carbone de la paraffine éluée avant le produit inconnu x.

$t_R(x)$: temps de rétention du produit inconnu x.

$t_{R(Cn)}$: temps de rétention de la paraffine à n atomes de carbone éluée avant le produit x.

$t_{R(Cn+1)}$: temps de rétention de la paraffine à n+1 atomes de carbone éluée après le produit x.

Les indices de rétention polaire (Ir p) et apolaire (Ir a) sont comparés à ceux d'échantillons authentiques contenus dans des bibliothèques de référence élaborées au laboratoire, dans des bibliothèques commerciales [4,7] ou répertoriés dans la littérature.

Cependant, une reproductibilité parfaite des indices de rétention est difficile à obtenir et ne peut être observée que sur des chromatogrammes réalisés sur une période courte avec des conditions expérimentales rigoureusement identiques. Les variations les plus importantes sont observées lorsqu'on compare les indices de rétention obtenus au laboratoire avec ceux de la littérature, particulièrement pour ce qui concerne la colonne polaire [8].

IV.1 METHODOLOGIE :

IV.1.1 Données expérimentales :

IV.1.1.1 *Mesure des indices de rétention [9] :*

Les indices de rétention des solutés (Tableau 2 ; figure 7) ont été mesurés sur une colonne en acier inoxydable (longueur : 2 m ; diamètre intérieur 1.8 mm) garnie de Chromosorb W (100-120 mesh) imprégné à 10%, par rapport au support d'Apiezon MH.

Les indices de rétention des dérivés benzéniques monosubstitués ont été ramenés à 150°C en traitant les variations des indices de rétention en fonction de la température par la méthode des moindres carrés. Les indices de rétention des autres composés ont été mésurés dans le

domaine de température allant de 90°C (fluorotoluène) à 180°C (bromochlorobenzène).
Les indices de rétention des 38 dérivés du benzène étudiés varient entre 664.1 et 1287.7 unités d'indice (u.i), pour un indice de rétention moyen égal à 965.2 u.i.

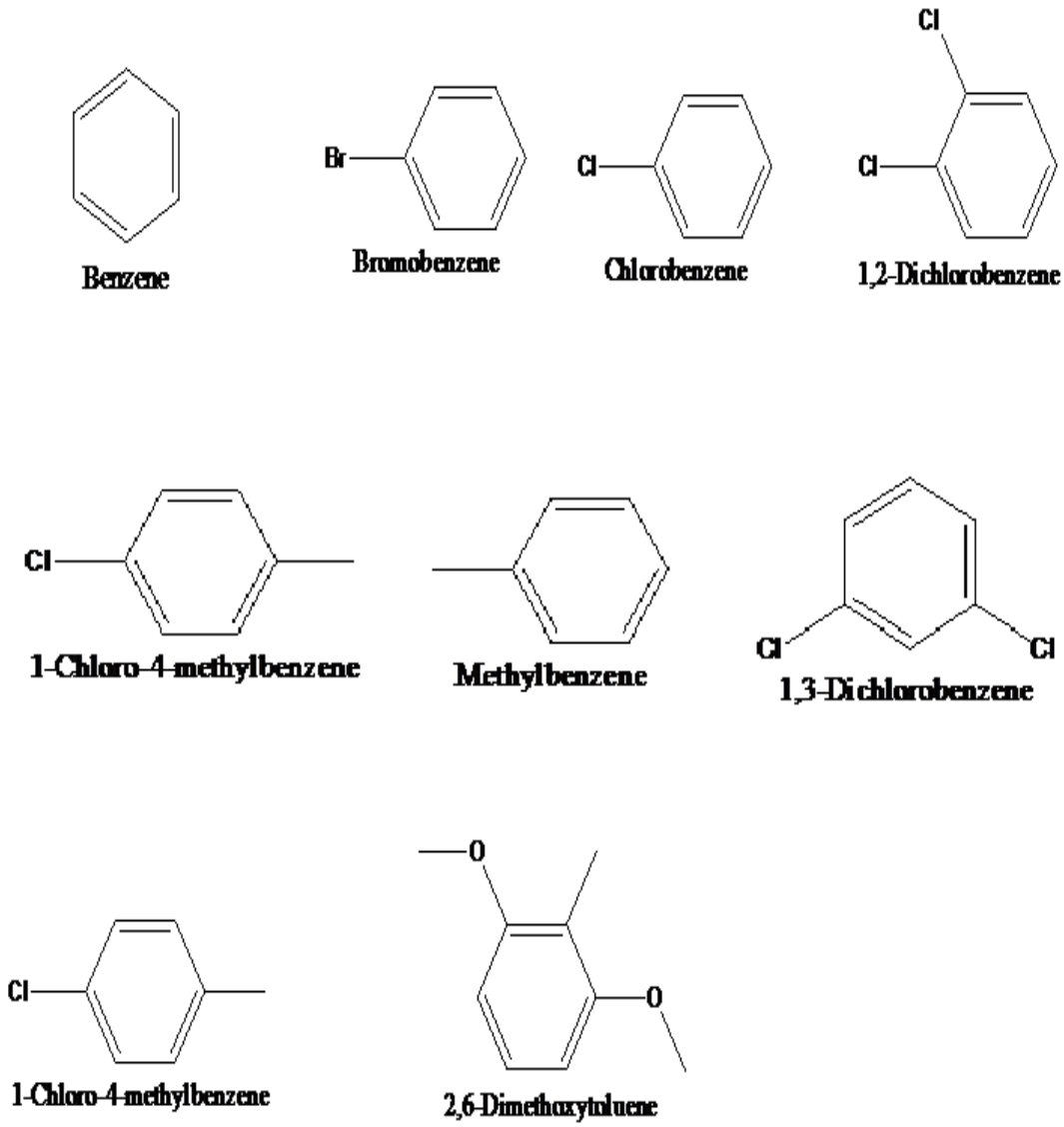


Figure 7: Quelques molécules citées dans le tableau 2

IV.1.2 Calcul des descripteurs.

Nous avons utilisé le logiciel HYPERCHEM [10] pour représenter chaque molécule dont la géométrie est d'abord pré-optimisée par des calculs de mécanique moléculaire, puis, pour chacune d'elles, les coordonnées atomiques (x,y,z) correspondant à la conformation de plus basse énergie sont déterminées par la méthode PM3. Tous les calculs ont été menés dans le cadre du formalisme de Hartree-Fock avec contrainte de spin (ou RHF: pour Restricted Hartree-Fock) sans interaction de configuration.

Les structures moléculaires ont été optimisées à l'aide de l'algorithme Polack-Ribiere avec pour critère d'arrêt une racine du carré moyen du gradient égale à 0,001 kcal/mol. Les géométries ainsi optimisées ont été transférées dans le logiciel informatique Dragon version 5.3 [11] pour le calcul de 1664 descripteurs appartenant à 20 classes différentes. Les descripteurs d'un même groupe à valeur constante (écarts types inférieurs à 0,0001) qui n'apportent aucune information sont éliminés ; pareillement, de deux descripteurs hautement corrélés $r \geq 0,92$ qui véhiculent une information redondante, on exclut automatiquement celui qui est corrélé avec le plus grand nombre de descripteurs. Le pool initial de 1664 descripteurs est ainsi réduit à 203.

IV.1.3 Sélection des points (les molécules):

En utilisant l'algorithme de Kennard et Stone [12] l'ensemble complet a été divisé en deux sous-ensembles : un ensemble de calibration de 28 composés, et un ensemble de test comprenant les 10 composés restants.

- Le jeu d'entrainement qui va permettre la construction du modèle, est composé de 28 molécules
- Le jeu de validation qui va servir à calculer la prédictivité du modèle est constitué de 10 molécules.

IV.2 TECHNIQUES DE SELECTION DES MODELES :

La sélection de sous-ensembles de variables (VSS) est réalisée par algorithme génétique (GA-VSS) en maximisant le coefficient de prédiction Q^2_{LOO}. Les algorithmes génétiques sont des algorithmes d'optimisation s'appuyant sur des techniques dérivées de la génétique et des mécanismes d'évolution de la nature : croisement (ou crossover) et mutation qui sont responsables de la génération de nouveaux individus.

Dans le logiciel MobyDigs [13] de tels processus sont contrôlés par un paramètre T variant entre 0 et 1, dont la valeur, choisie par l'utilisateur, renseigne sur le niveau du compromis croisement/mutation.

Pour éviter des modèles présentant des problèmes de colinéarité et sans réelle capacité de prédiction, nous avons appliqué la règle QUIK (Q Under Influence of K) [14] basée sur l'indice de corrélation multivariable K [14], défini par :

$$K = \frac{\sum_J \left| \frac{\lambda_j}{\sum_j \lambda_j} - \frac{1}{p} \right|}{2(p-1)/p}; j = 1,..., p \quad \text{et} \quad 0 \le k \le 1$$

Les λ_j sont les valeurs propres de la matrice de corrélation de l'ensemble des données $(n \times p)$, n étant le nombre d'objets et p le nombre de variables.

Cette règle est déduite de l'hypothèse que la corrélation totale dans l'ensemble formé par les prédicteurs **X** du modèle plus la réponse **Y**, (K_{xy}) doit toujours être plus grande que celle uniquement mesurée dans l'ensemble des prédicteurs (K_{xx}). Le calcul de (K_{xy}) est réalisé en considérant la réponse **Y** comme une variable **x** et en calculant la matrice de corrélation correspondante. En général [15], on s'accorde à rejeter les modèles qui ne vérifient pas la relation :

$$D(K) = K_{xy} - K_{xx} > 0,05$$

Nous rechercherons un modèle avec le minimum de variables explicatives (règle de parcimonie), dont l'interprétation pourra être reliée au phénomène de rétention sur la phase stationnaire apolaire Apiezon MH utilisée pour les analyses.

Tableau 2: Indices de rétention observés et valeurs des descripteurs optimaux sélectionnés.

N°	**Composés**	**Ensemble**	IR_{exp}	$X\,0Av$	$X\,1sol\,2$
01	Benzène	Training	681,3	0,577	9,0000
02	Fluorobenzène	Training	664,1	0,538	9,0000
03	Chlorobenzène	Training	877,9	0,646	13,5645
04	Bromobenzène	Training	979,6	0,764	15,7688
05	Toluène	Training	788,2	0,627	11,5192
06	Anisole	Training	923,6	0,599	15,4606
07	p-Chloroanisole	Training	1131,7	0,650	21,2890
08	p-Xylène	Training	889,2	0,664	14,3489
09	p-Fluorotoluène	Training	777,7	0,586	11,5192
10	p-Bromotoluène	Training	1096,3	0,784	19,0532
11	p-Bromofluorobenzène	Training	940,9	0,706	15,7688
12	p-Chlorobromobenzène	Training	1174,4	0,801	21,6597
13	m-Chloroanisole	Training	1126,0	0,650	21,2890
14	m-Méthylanisole	Training	1029,6	0,635	18,7143
15	m-Xylène	Training	892,0	0,664	14,3489
16	m-Chlorobromobenzène	Training	1179,0	0,801	21,6597
17	m-Bromotoluène	Training	1100,0	0,784	19,0532
18	m-Fluorotoluène	Training	778,0	0,586	11,5192
19	m-Dibromobenzène	Training	1287,7	0,905	24,4234
20	o-Méthylanisole	Training	1013,5	0,635	18,8616

Tableau 2: *Suite et fin*

N°	**Composés**	**Ensemble**	IR_{exp}	$X\,0Av$	$X\,1sol\,2$
21	o-Chloroanisole	Training	1135,6	0,650	21,4462
22	o-Bromofluorobenzène	Training	959,6	0,706	15,7688
23	o-Xylène	Training	916,2	0,664	14,4780
24	o-Bromochlorobenzène	Training	1197,6	0,801	21,8182
25	p-Méthylanisole	Training	1029,5	0,635	18,7143
26	o-Bromotoluène	Training	1095,7	0,784	19,2019
27	m-Fluoroanisole	Training	908,5	0,566	15,4606
28	p-Chlorotoluène	Training	989,2	0,680	16,6138
29	p-Fluoroanisole	Test	910,6	0,566	15,4606
30	m-Chlorotoluène	Test	990,9	0,680	16,6138
31	m-Chlorofluorobenzène	Test	835,4	0,603	13,5645
32	m-Dichlorobenzène	Test	1060,5	0,697	19,0532
33	o-Fluorotoluène	Test	777,4	0,586	11,5192
34	o-Fluoroanisole	Test	919,7	0,566	15,4606
35	o-Chlorofluorobenzène	Test	862,0	0,603	13,5645
36	p-Chlorofluorobenzène	Test	840,5	0,603	13,5645
37	o-Chlorotoluène	Test	986,3	0,680	16,7526
38	m-Bromofluorobenzène	Test	932,8	0,706	15,7688

IV.3 Résultats et discussion :

Des modèles ont été générés en utilisant la méthode de régression multilinéaire MLR, les valeurs des IR, comme variable dépendante, $X\,0Av$ et $X\,1sol\,2$ sont utilisés comme variables

indépendantes.

Plusieurs modèles impliquant un, deux et 4 descripteurs sont élaborés.

Les modèles sont évalués par plussieurs paramètres statistiques (R^2, R^2_{adj} , Q^2_{LOO} , F, S, DK).

Le tableau 3 indique que l'indice de rétention est corrélé linéairement avec le descripteur $X1sol$, ou mieux avec son carré $X1sol2$.

Tableau 3: Comparaison des paramètres statistiques des différents modèles.

N	Descripteurs	R^2	R^2_{aj}	Q^2_{LOO}	Q^2_{BOOT}	F	S	DK
28	$X1sol$	98,00	97,7	97,64	97,64	1247,37	23,16	-
28	$X1sol2$	98,19	98,12	97,95	97,76	1409,32	21,82	-
28	$X0Av, X1sol2$	99,59	99,56	99,47	99,4	3068,88	10,53	0,123
32	$NOCH_3, XV0, VOL, DIMO$	99,63	99,57	99,46	99,36	1814,17	10,12	0,126

Nous avons adopté le modèle à 2 descripteurs $X0Av$ et $X1sol2$, dont l'équation, calculée en utilisant les valeurs centrées réduites est :

$$IR = 0,168\ X\ 0Av + 0,872\ X\ 1sol2 \qquad (1)$$

où

$X0Av$ désigne l'indice de connectivité de valence d'ordre zéro moyen,

$X1sol$ l'indice de connectivité de solvatation du 1er ordre [16].

L'association de ces 2 descripteurs permet d'améliorer tous les paramètres statistiques (qui sont comparés dans le tableau 3) en particulier l'erreur standard qui est divisée par plus de 2 (21,82→10,53) est de l'ordre de grandeur (10,12) de celle obtenue avec le modèle à 4 descripteurs publié [9].

Notons également un DK (=0,123) supérieur à la limite 0,05 choisie.

Les paramètres statistiques montrent que le modèle (eq. 1) établit une forte corrélation entre les 2 variables choisies et la propriété étudiée, caractérisée par un excellent coefficient de détermination $\left(R^2 = 99,59\%\right)$ qui explique autour de 99,60% de la variation des données, en plus d'une très grande valeur du F de Fisher (=3068,88), qui indique l'excellence du

modèle dans la prédiction des valeurs IR, et une erreur standard acceptable (S=10,53). L'équation (1) présente un R^2_{aj} de 99,56, ce qui indique un excellent accord entre la corrélation et la variation des données. La petite différence entre R^2 et Q^2_{LOO} renseigne sur la robustesse du modèle. La valeur de $Q^2_{Boot}(=99,4)$ confirme à la fois la prédictivité interne et la stabilité du modèle.

La figure 8 qui représente le graphe des coefficients statistiques Q^2 et R^2 permet de comparer les résultats obtenus pour les modèles randomisés (cercles) au modèle de départ (losange). Il est clair que les statistiques obtenues pour les vecteurs modifiés des indices de rétention sont plus petites que celles du modèle QSRR réel, ce qui permet d'assurer qu'une relation réelle structure/rétention a été établie.

Les paramètres statistiques suivants, obtenus pour l'ensemble de test externe, démontrent le pouvoir de prédiction du présent modèle.

$$Q^2_{EXT} = 0,9869 > 0,5 \qquad r^2 = 0,9765 > 0,6$$

$$\left(r^2 - r_0^2\right)/r^2 = \left(0,9765 - 1,000\right)/0,9765 = -0,0240 < 0,1$$

Ou $$\left(r^2 - r_0'^2\right)/r^2 = \left(0,9765 - 1,000\right)/0,9765 = -0,0240 < 0,1$$

$$0,85 < k = 1,0002 < 1,15 \quad ou \quad 0,85 < k' = 09996 < 1,15$$

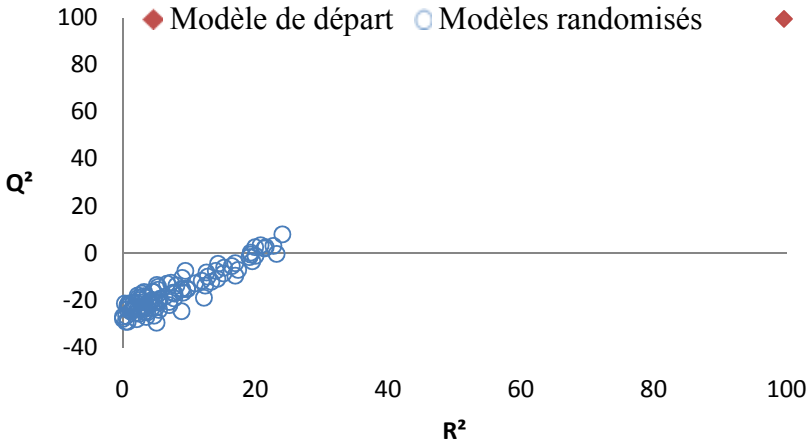


Figure 8: Représentation graphique du test de randomisation**.**

- **Domaine d'application.** La figure 9 compare les diagrammes de Williams obtenus pour notre modèle à 2 descripteurs, et celui à 4 descripteurs publié [9]. Dans les 2 cas, les valeurs des leviers de tous les composés de calibration et de test sont inférieures aux valeurs critiques h^*correspondantes (respectivement 0,321et 0,468) et, dans les deux cas, aucun des composés n'est influent.

Par ailleurs, pour le modèle à 2 descripteurs développé (figure 9-A) tous les composés de calibration et de test présentent des résidus de prédiction standardisés inférieurs, en valeur absolue, à 3 unités d'écart-type (3σ), ce qui montre qu'il n'y a pas de données aberrantes.

Il n'en est pas de même pour le modèle à 4 descripteurs publié (figure 9-B) pour lequel nous notons deux données aberrantes, l'une de l'ensemble de calibration (composé 30: o-xylène) et l'autre de l'ensemble de test (composé 38: m-Bromofluorobenzène).

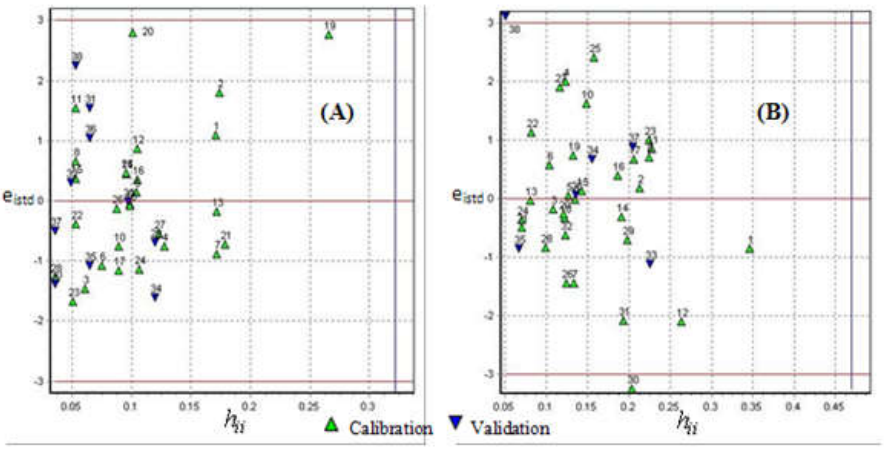


Figure 9: Diagramme de Williams du modèle à 2 descripteurs proposé (A) et du modèle à 4 descripteurs publié (B).

- **Interprétation du modèle.** Le descripteur $X1sol2$ qui est très corrélé avec IR régit notablement celui-ci comme le montrent les valeurs des coefficients des 2 descripteurs du modèle, et des t de Student associés qui valent 48,33 (pour $X1sol2$) et 9,31 (pour $X0Av$).

Les indices de connectivité de solvatation sont définis [17] pour un graphe moléculaire dépourvu des atomes d'hydrogène et en ne tenant pas compte des atomes de fluor, ces 2 atomes ayant des dimensions proches.

L'indice de connectivité de solvatation du premier ordre est calculé à partir de l'équation :

$$X\,1sol = \frac{1}{4}\sum_{b=1}^{B}\frac{\left(L_i\,L_j\right)_b}{\left(\delta_i\,\delta_j\right)_b^{0,5}} \qquad (2)$$

où b porte sur toutes les liaisons au nombre de B, L_i et L_j sont les nombres quantiques principaux des 2 sommets (atomes) incidents à la liaison considérée ; δ_i et δ_j représentent les degrés (valences) des sommets correspondants.

Les indices de connectivité de solvatation permettent de modéliser l'entropie de solvatation et de décrire les interactions de dispersion en solution qui jouent un rôle décisif dans le phénomène de rétention.

L'indice de connectivité de valence d'ordre zéro moyen ($X0Av$) est obtenu en divisant l'indice de connectivité de valence d'ordre zéro ($X0V$) par le nombre B d'arêtes (liaisons) du graphe de la molécule dépourvue des atomes d'hydrogène. $X0V$ est défini [18,19] par :

$$X0v = \sum_{i=1}^{N}\left(\delta_i^v\right)^{-0,5} \qquad (3)$$

N étant le nombre de sommets du graphe, c'est-à-dire le nombre d'atomes de la molécule autres que l'hydrogène.

δ_i^v est calculé pour l'atome i, à partir de l'expression :

$$\delta_i^V = \frac{Z_i^V - H_i}{Z_i - Z_i^V - 1} \qquad (4)$$

Z_i et Z_i^V représentant, respectivement, le numéro atomique et le nombre d'électrons de valence de l'atome i, alors que H_i désigne le nombre d'atomes d'hydrogène liés à l'atome considéré.

Outre qu'il introduit des corrections relatives aux différences entre les types d'halogènes contenus dans une molécule donnée, le descripteur $X0v$ est lié à la taille et au degré de ramification des molécules, qui peuvent jouer un rôle non négligeable dans le processus de distribution du soluté entre les 2 phases (mobile/stationnaire) chromatographiques.

IV.4 Conclusion :

Les indices de rétention des 38 dérivés benzéniques séparés par chromatographie en phase gazeuse, ont été corrélés avec 2 descripteurs théoriques calculés à partir de la structure des molécules, et sélectionnés par algorithme génétique parmi 1664 descripteurs moléculaires obtenus à partir du logiciel DRAGON.

Le modèle QSRR présenté est robuste, avec de bonnes capacités prédictives internes et externes et une bonne qualité de l'ajustement.

Références bibliographiques :

[1] P. Arpino, A. Prévôt, J. Serpinet, J. Tranchant, A. Vergnol, P. Witier. *Manuel pratique de chromatographie en phase gazeuse*. Masson, Paris, **1995**.

[2] E. Kováts. Gas Chromatographic characterization of organic substances in the retention index system, in *Advances in Chromatography*, Chap. 7, **1965**, 229-247.

[3] H. Van Den Dool, P. D. Kratz. *J. Chromatogr.*, **1963**, 11, 463-471.

[4] D. Joulain, W. A. König. *The atlas of spectral data of sesquiterpene hydrocarbons.* Ed. E.B.-Verlag, Hambourg, **1998**.

[5] W.A. König, D. H. Hochmuth, D. Joulain. *Terpenoids and related constituents of essential oils*. University of Hamburg, Institute of organic chemistry, Hamburg, Germany, **2001**.

[6] W. Jennings, T. Shibamoto, *Qualitative analysis of flavour and fragrance volatiles by glass-capillary gas chromatography*. Ed. Jovanovitch H.B., Academic press, New-York, **1980**.

[7] R. P. Adams. *Identification of essential oil components by gas chromatography/mass spectroscopy*. Allured publishing corporation, Carol stream, Illinois, **1995**.

[8] F. Grundschober. *Z. Lebensm. Unters. Forsh.*,**1991**, 192, 530-534.

[9] M. J. Heravi, Z. G. Nejad. *J. Chromatogr.,***1993**, 648, 389-393.

[10] Hyperchem 6.03, (Hypercube), http://www.hyper.com.

[11] Dragon 5.4, http://www.disat.unimib.it

[12] R. W. Kennard, L. A. Stone L.A. *Technometrics*.,**1969**, 11, 137-148.

[13] MobyDigs 1.1, http://www.disat.unimib.it

[14] R. Todeschini. *Anal. Chim. Acta.*,**1997**, 348, 419-430.

[15] R. Todeschini, V. Consonni, A. Mauri, M. Pavan. *Anal. Chim. Acta.,***2004**, 515, 199-208.

[16] R. Todeschini, V. Consonni. *Molecular descriptors for chemoinformatics*. Second, Revised and Enlarged Edition.Vol. I: Alphabetical listing. Series Editors: Mannhold R., Kubinyi H., Folkers G., Wiley-VCH Verlag GmbH CO. KGaA. **2009**, pp. 967.

[17] N.S. Zefirov, V.A. Polyulin. *J. Chem. Inf. Comput. Sci.*, **2001**, 41, 1022-1027.

[18] L. B. Kier, L. H. Hall. *J. Pharm. Sci.*, **1981**, 70, 583-589.

[19] L. B. Kier, L. H. Hall. *J. Pharm. Sci.*, **1983**, 72, 1170-1173.

Chapitre V
Modélisation de la toxicté

V Introduction :

Au début des années 1960 Corwin Hansch a proposé un modèle mathématique pour corréler l'activité biologique et la structure chimique [1], cette date est considérée comme étant la naissance des méthodes QSAR. Depuis, l'utilisation des QSAR en toxicologie n'a pas cessé d'évoluer.

Des modèles QSAR sont maintenant mis au point en utilisant une variété d'approches, de méthodes d'analyse de données et de paramètres [2].

Des modèles QSAR basés sur des indices topologiques pour l'étude de la toxicité ont fait l'objet de nombreux travaux (Burden [3], Gramatica *et al.* [4], Grodnitzky et Coast [5], Huuskonen [6], Rose et Hall [7]).

Un grand nombre d'études QSAR de toxicité et en particulier la toxicité aigüe ont été puliées dans la littérature. La plupart des données de toxicité pour l'environnement ont été obtenues en utilisant des espèces aquatiques en l'occurrence les poissons, telle que *PIMEPHALES PROMELAS.*

Figure 10: L'espèce aquatique utilisée pour l'étude de la toxicité.

V.1 METHODOLOGIE :

V.1.1 Données expérimentales :

Les travaux de Lemon B. Kier and Lowell H. Hal [8], ont permis de rassembler les données de la toxicité vis-à-vis de *PIMEPHALES PROMELAS* également appelé FATHEAD MINNOW ou tête de boule de 141 dérivés benzénques comportant 9 différents groupements fonctionnels. Les valeurs de la toxicité correspondant à la concentration du composé pour laquelle 50% des animaux meurent en 96 heures, cette toxicité est notée « 96h LC50».
Cette variable est exprimée par le rapport logarithmique $p^{DL}{}_{50}$ ($Log1/DL_{50}$).
La toxicité expérimentale exprimée par $p^{DL}{}_{50}$ des 141 dérivés benzéniques sont numérotés dans le tableau 3.

Tableau 4: Données expérimentales de la toxicité des dérivés benzéniques vis-à vis PIMEPHALES PROMELAS.

N°	Composé	Toxicité observée pDL_{50}
01	Benzene	3,40
02	Bromobenzene	3,89
03	Chlorobenzene	3,77
04	1,3,5-Tricholoro-2-hydroxybenzene	4,40
05	1,3-Dichlorobenzene	4,62
06	1,4-Dichlorobenzene	4,02
07	1-Chloro-2 hydroxybenzene	3,84
08	1-Methyl-2,3,6-trinitrobenzene	3,04
09	1-Chloro-4-methylbenzene	3,21
10	1-Aminobenzene	3,58
11	1-Hydroxy-3-methylbenzene	3,07
12	1-Hydroxy-4-nitrobenzene	4,21

Tableau 4: Suite

N°	Composé	Toxicité observée pDL_{50}
13	1,4-Dimethoxybenzene	3,57
14	1,2-Dimethylbenzene	3,76
15	1,4-Dimethylbenzene	4,38
16	1-Methyl-3-nitrobenzene	3,24
17	1-Aldehydo-2-nitro-5-hydroxybenzene	3,35
18	1-Methyl-4-nitrobenzene	3,80
19	1,3-Dinitrobenzene	3,80
20	1-Amino-2-methyl-3-nitrobenzene	3,79
21	1-Amino-2-methyl-5-nitrobenzene	4,89
22	1-Aldehydo-3-methoxy-4-hydroxy-5-bromobenzene	4,74
23	1,2,3-trichlorobenzene	3,86
24	1,2,4-trichlorobenzene	3,75
25	1,3,5-trichlorobenzene	4,04
26	1,2-Dichloro-4-methylbenzene	5,01
27	1-Aldehydo-3-methoxy-4-hydroxybenzene (vaniline)	3,99
28	1-Hydroxy-2,6-dimethylbenzene	3,91
29	1-Hydroxy-3,4-dimethylbenzene	4,12
30	1-Hydroxy-2,4-dinitrobenzene	5,34
31	1,2,4-Trimethylbenzene	4,26
32	1-Methyl-2,3-dinitrobenzene	4,21

Tableau 4: Suite

N°	Composé	Toxicité observée pDL_{50}
33	1-Aldehydo-2-hydroxy-3,5-dibromobenzene	4,18
34	1-Methyl-2,6-dinitrobenzene	4,70
35	1-Methyl-3,5-dinitrobenzene	5,85
36	1-Amino-2-methyl-3,5-dinitrobenzene	4,88
37	1-Amino-2-methyl-3,6-dinitrobenzene	3,56
38	1-Amino-2,6-dinitro-3-methylbenzene	4,07
39	1-Amino-2,6-dinitro-4-methylbenzene	3,60
40	1-Amino-3,5-dinitro-4-methylbenzene	3,93
41	1,3,5-Tribromo-2-hydroxybenzene	4,74
42	2-Allylphenol	6,09
43	1,2,3,4-Tetrachlorobenzene	5,93
44	1-Methyl-2,4,6-trinitrobenzene	3,73
45	1-Hydroxy-2,3,4,5,6-pentachlorobenzene	4,00
46	1-Amino-4-bromobenzene (4-bromoaniline)	3,42
47	4-Butylphenol	4,72
48	1-Amino-2,4-dinitrobenzene	4,81
49	1-Amino-2-chloro-4-methylbenzene	4,02
50	1-Amino-2-chloro-4-nitrobenzene	4,14
51	4-Butylphenol	3,92
52	1-Amino-2,3,4-trichlorobenzene	5,19

Tableau 4: Suite

N°	Composé	Toxicité observée pDL_{50}
53	1,3,5-Trichloro-2,4-dinitrobenzene	5,31
54	1-Amino-2,3,5,6-tetrachlorobenzene	4,73
55	1-Cyano-3,5-dibromo-4-hydroxybenzene	4,02
56	1-Cyano-2-amino-6-methylbenzene	4,83
57	1-Cyano-2-methylbenzene	3,69
58	4-Pentylphenol	3,70
59	1-Aldehydo-2-chloro-5-nitrobenzene	3,82
60	1-Aldehydo-2,4—dichlorobenzene	5,25
61	1-Aldehydo-4-chlorobenzene	4,23
62	1-Aldehydo-2-dinitrobenzene	4,56
63	4-Nonylphenol	2,87
64	1-Aldehydo-2,4-dimethylbenzene	3,39
65	1-Aldehydo-2-hydroxy-5-bromobenzene	3,79
66	1-Aldehydo-2-hydroxy-5-chlorobenzaldehyde	3,34
67	2,4-dichlorophenol,	3,77
68	1-Aldehydo-3-methoxy-4-hydroxybenzene (vaniline)	3,52
69	1-Aldehyde-2-hydroxy-4,6-dimethoxybenzene	3,51
70	1-Fluoro-4-nitrobenzene	4,12
71	2,3,4,5-Tetrachlorophenol	4,21
72	1-Amino-4-fluorobenzene	4,26

Tableau 4: Suite

N°	Composé	Toxicité observée pDL_{50}
73	1-Aldehydo-2,3,4,5,6-pentafluorobenzene	4,46
74	1-Acyl-4-chloro-3-nitrobenzene	5,29
75	1-Acyl-2,4-dichlorobenzene	2,60
76	4-t-pentylphenol	4,06
77	3-Nitrobenzonitrile	5,00
78	4-Nitrobenzonitrile	4,16
79	2-Phenylphenol	3,32
80	4-Amino-2-nitrotoluene	5,07
81	3-Methyl-2-nitrophenol	5,07
82	5-Methyl-2-nitrophenol	3,65
83	1,5-Dimethyl-2,4-dinitrobenzene	4,27
84	4-Phenylazophenol	5,15
85	2,4-Dinitro-5-methylphenol	3,60
86	1-Naphtol	3,59
87	1,3,5-Trinitrobenzene	3,36
88	2-Phenoxyethanol	3,26
89	Benzophenone	4,96
90	2,3,4-Trichloroacetophenone	3,93
91	1-Chloronaphtalene	4,46
92	2,6-Dimethoxytoluene	5,18

Tableau 4: Suite

N°	Composé	Toxicité observée pDL_{50}
93	Diphenylether	6,20
94	p-Nitrophenylphenylether	4,30
95	1,2-Dinitrobenzene	4,45
96	1,4-Dinitrotoluene	5,26
97	1-Amino-3-nitro-4-hydroxybenzene	4,53
98	1-Methyl-2,5-dinitrobenzene	4,91
99	Hyroxybenzene	3,51
100	1,2-Dichlorobenzene	4,30
101	1-Chloro-3-hydroxybenzene	4,33
102	1,3-Dihydroxybenzene	3,77
103	1-Hydroxy-3-methoxybenzene	3,29
104	1-Hydroxy-2-methylbenzene	3,36
105	1-Hydroxy-4-methylbnzene	3,48
106	1-Amino-2-chlorobenzene	3,63
107	1-Methyl-2-nitrobenzene	3,48
108	1-Amino-2-methyl-4-nitrobenzene	3,77
109	1-Amino-2-methyl-6-nitrobenzene	5,00
110	1-Amino-3-methyl-6-nitrobenzene	4,30
111	1-Amino-2-nitro-4-methylbenzene	4,74
112	1-Amino-3-nitro-4-methylbenzene	4,54

Tableau 4: Suite

N°	Composé	Toxicité observée pDL_{50}
113	1-Aldehydo-2-methylbenzene	3,90
114	1,3-dichloro-4-hydroxybenzene	4,21
115	1,3-dichloro-4-methylbenzene	3,75
116	1-Hydroxy-2,4-dimethylbenzene	5,08
117	1-Methyl-2,4-dinitrobenzene	4,46
118	1-Methyl-3,4-dinitrobenzene	5,43
119	1-Aldehydo-2-fluorobenzene(o-luorobenzaldehyde)	6,06
120	1-Amino-2,4-dinitro-3-methylbenzene	4,33
121	1,2,3,5-Tetrachlorobenzene	4,38
122	1-Amino-3,4-dichlorobenzene	4,99
123	1-Cyano-2-amino-5-chlorobenzene	4,81
124	1-Aldehydobenzene (benzaldehyde)	4,21
125	1-Aldehydo-2-hydroxybenzene	3,48
126	1-Aldehydo-2-methoxy-4-hydroxybenzene	3,80
127	1-Amino-2,3,4,5,6-pentafluorobenzene	4,39
128	1-Aldehydo-2-chloro-6-fluorobenzene	4,92
129	1-Acylbenzene	2,87
130	2-Amino-4-nitrotoluene	3,88
131	2-Amino-6-nitrotoluene	4,63
132	3-Amino-4-nitrotoluene	4,91

Tableau 4: Suite et fin

N°	Composé	Toxicité observée pDL_{50}
133	2-Amino-4,6-dinitrotoluene	4,33
134	3-Amino-2,4-dinitrotoluene	6,37
135	3-Amino-2,6-dinitrotoluene	3,56
136	4-Amino-2,6-dinitrotoluene	4,34
137	Acetophenone	5,52
138	2,4-Dichloroacetophenone	4,47
139	5-Amino-2,4-dinitro-1-methylbenzene	5,72
140	Methylbenzene	4,82
141	1-Chloro-2-methyl-4-hydroxybenzene	4,85

Lors d'une étude antérieure réalisée par ces auteurs, une relation structure –toxicité des 141 dérivés benzéniques a été établie.

L'ensemble des molécules a été divisé en 2 sous ensembles : un sous ensemble d'apprentissage constitué de 118 composés (environ 83.7%) utilisé pour la construction du modèle QSAR et un sous ensemble de 23 composés (environ 16.3%) réservé pour la validation externe.

A cet effet la méthode utilisée pour relier la structure de ces molécules à la propriété expérimentale est les réseaux de neurones artificiels (RNA).

L'architecture de ce réseau est la suivante (17 :4 :1) qui est composé de 3 couches :

- Une couche d'entrée comprenant 17 neurones (les 17 descripteurs sélectionnés)
- Une couche intermédiaire, appelée couche cachée comprenant 4 neurones
- Une couche de sortie comprenant une neurone $p^{DL}{}_{50}$.

Les paramètres statistiques du modèle sont listés dans le tabeau 5.

Tableau 5: Les paramètres statistiques du modèle 118/23.

	R^2	MAE
Sous ensemble d'apprentissage	0.86	0.19
Sous ensemble de prédiction	0.89	0.24

Cependant la discussion établie par les auteurs comporte des anomalies à signaler.

Anomalies:

1- Les 2 résidus les plus importants dans l'ensemble d'entraînement sont 1.01 pour le composé 1-aldéhydo-3-methyl-4-hydroxybenzene et 0,72 pour le composé 56 1-methyl-2,4,6-trinitrobenzene.

   En se réferant à leur tableau des résultats les composés ayant les valeurs de 1.01 et 0.72 sont respectivement 1-cyano-2-amino-3,5-dinitrobenzene, 1-amino-2-methyl-3,5-dinitrobenzene.

2- Pour l'ensemble test la valeur la plus importante qui a été signalée est de 0.61 pour le composé 4-tert-pentylphenol.

   En se réferant au tableau des résultats le composé 118 appartient à l'ensemble d'entraînement d'une part et d'autre part la valeur la plus importante des résidus dans l'ensemble de test est de 0,66 pour le composé 3-amino-2,4-dinitrotoluene.

3- Le composé 1-tert-pentylphenol est dans l'ensemble de l'entraînement et a un résidu ei=-0,61.

   Ce qui nous oblige à ne pas comparer les résultats de cette thèse avec le travail cité.

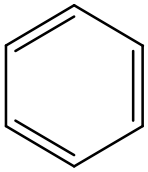
V.2 Utilisation des données pour une modélisation:

V.2.1 Source des données:

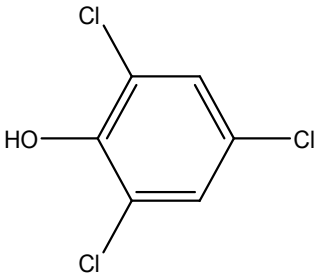
L'agence de protection de l'environnement des Etats-Unis a établi un programme pour la production de données de haute qualité sur la toxicité vis-à vis des poissons. Ces données ont été publiées dans une série de volumes (Brooke *et al.*, 1984) et ont fourni la base de plusieurs modèles QSAR (Hall et Kier, 1984, 1986, Hall *et al.*, 1985, 1989).

Dans notre étude, l'ensemble à modéliser est constitué de 141 dérivés benzéniques utilisés dans le travail de Lemond B. Kier and Lowell H. Hall et prélevé de la base de données citée.

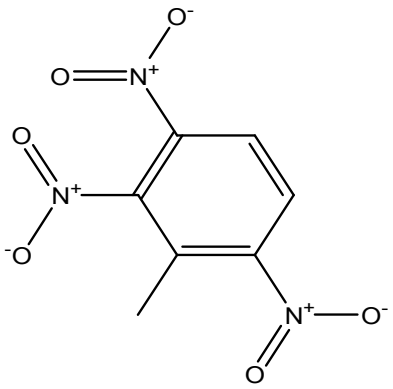
Les structures moléculaires de quelques dérivés benzéniques sont utilisées dans la figure 11



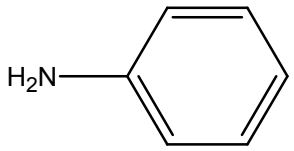
benzene



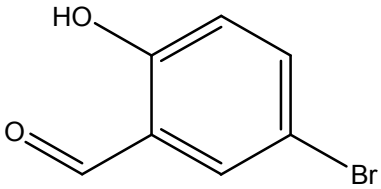
1,3,5-trichloro-2-hydroxybenzene



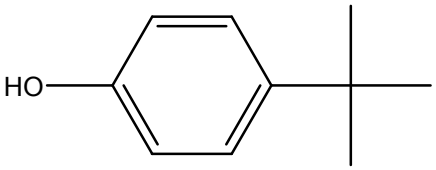
1-Methyl -2,3,6-trinitrobenzene



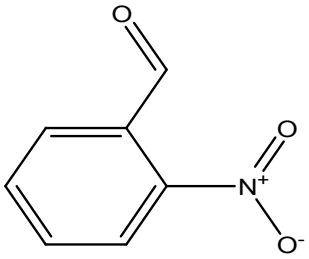
1-Aminobenzene



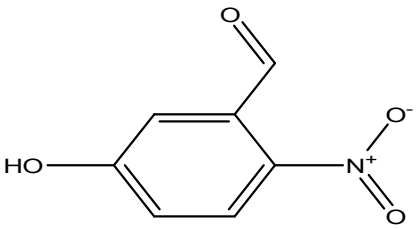
2-Hydroxy-5-bromobenzaldehyde



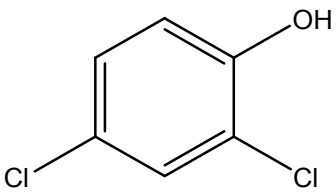
4-t-Butylphenol



2-nitrobenzaldehyde



5-Hydroxy-2-nitrobenzaldehyde



2,4-dichlorophenol

Figure 11: Quelques molécules étudiées

V.2.2 Logiciels utilisés dans nos études QSTR:

Il existe plusieurs logiciels libres ou commerciaux disponibles dans les études QSAR/QSPR. Ceux-ci comprennent des logiciels spécialisés pour dessiner les structures chimiques, générant des structures 3D, le calcul des descripteurs moléculaires et le développement de modèles RQSA/RQSP.

Les logiciels utilisés dans nos travaux sont:

- Le dessin des molécules a été fait par ChemOffice, ChemSketch et Marvin Sketch [9,11].
- L'optimisation des molécules a été réalisée par le logiciel Hyperchem
- Les descripteurs ont été calculés par le logiciel Dragon.

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

- Pour la régression linéaire multiple, nous avons utilisé le logiciel QSARIN [12]
- Pour les réseaux de neurones artificiels RNA, nous avons utilisé MATLAB [13].

V.2.3 Dessin et optimisation des structures :

Les structures des molécules ont été dessinées et optimisées à l'ide du logiciel Hyperchem 6.03(2000) en utilisant la méthode semi empirique PM3 [14].

Tous les calculs ont été exécutés dans le cadre du formalisme de Hartree-Fock avec contraite de spin (ou RHF pour restricted Hatree-Fock) sans interaction de configuration [15]. Les structures moléculaires ont été pré-optimisées à l'aide de l'algorithme Polak-Ribiere avec pour crière une racine du carré moyen du gradient égale à 0.001 $kcal.mol^{-1}$.

V.2.4 Calcul des descripteurs:

Après l'optimisation des géométries des molécules, nous avons procédé au calcul de 1664 descripteurs sur 22 différents blocs à l'aide du logiciel Dragon 5.5.

Les descripteurs à valeurs constantes (écarts types < à 0,0001) et ceux largement corrélés (R>0.95) sont exclus.

V.2.5 Sélection des points (les molécules):

Nous avons partagé la base de données expérimentale en 2 jeux distincts :

- Le jeu d'entraînement qui va permettre la construction du modèle, est composé de 98 molécules.
- Le jeu de validation qui va servir à calculer la prédictivité du modèle est constitué de 43 molécules.

Pour ce faire nous avons fait appel à l'algorithme Kennard et Stone (KS).

V.2.6 Sélection des variables (les descripteurs):

Comme nous l'avons rappelé, un grand nombre de descripteurs différents sont collectés pour la modélisation de la toxicité. Cependant, les descripteurs envisagés n'ont pas tous une influence significative sur cette grandeur modélisée, et les variables ne sont pas toujours mutuellement indépendantes. De plus, le nombre de descripteurs, c'est-à-dire la dimension du vecteur d'entrée, détermine la dimension du vecteur des paramètres à ajuster. Si cette dimension est trop importante par rapport au nombre d'exemples de la base d'apprentissage, le modèle risque d'être surajusté à ces exemples, et incapable de prédire la grandeur modélisée sur de nouvelles observations. Il est donc nécessaire de réduire la dimensionnalité des variables d'entrée.

A cet effet, le choix des descripteurs optimaux pour le calcul du meilleur modèle, a été réalisé selon 3 différentes méthodes de sélection de variables : régresssion stepwise progressive qui est une méthode mixte (c'est une combinaison entre la méthode descendante et la méthode ascendante, algorithme génétique dans la version MobiDigs de Todeschini et dans la version QSARINS de Grammatica [12] et méthode de remplacement proposée par A.H.Morales *et al.*, 2006 [16].

Les six descripteurs moléculaires sélectionnés par les 3 méthodes de sélection sont rassemblés dans le tableau 6

Tableau 6: Les descripteurs moléculaires sélectionnés par les 3 méthodes de sélection de variables.

Méthode	**Descripteurs**					
RM	Me	EEig10x	Mor30u	SEigv	BEHe6	Mor07m
AG	EEig11x	ESpm07u	Mor14u	Mor30u	Mor23m	$R1e^+$
SW	RARS	PJI2	DECC	HOMA	BEHv2	Mor17u

V.3 Résultats et discussion :

V.3.1 Analyse statistique :

V.3.1.1 Technique de sélection de modèles :

Pour chaque méthode de sélection de variables (Méthode de remplacement RM, Méthode Stepwise SW et algorithme génétique AG), des modèles structure-toxicité ont été générés en utilisant la régression linéaire multiple MLR et la méthode des réseaux de neurones RNA.

V.3.1.1.1 La méthode de régression linéaire multiple :

Un modèle QSAR linéaire pour chaque méthode de sélection de variables est élaboré.

Le nombre de descripteurs est fixé à 6 descripteurs.

Les paramètres statistiques sont rassemblés dans le tableau 7.

Tableau 7: les paramètres statistiques en appliquant la regression MLR pour les 3 méthodes de sélection de variables

Méthodes	R^2	R^2_{aj}	S	F	K_{xx}	MAE_{tr}	Q^2_{LOO}	MAE_{CV}	MAE_{ext}	R^2_{ext}
RM	0,139	0,274	0,618	7,11	0,441	0,488	0,21	0,5255	0,5895	0,0422
AG	0,290	0,243	0,631	6,20	0,230	0,480	0,58	0,5158	0,6369	0,042
SW	0,289	0,242	0,631	6,17	0,441	0,471	0,16	0,5108	0,6001	0,125

Malgré les résultats peu concluants, la méthode stepwise s'avère la meilleure méthode de sélection des descripteurs.

En se basant sur le tableau ci-dessus, et les mauvais paramètres statistiques obtenus, nous avons essayé une autre méthode qui est les réseaux de neurones artificiels (RNA)

V.3.1.1.2 La méthode des réseaux de neurones artificiels

Pour les réseaux de neurones artificiels, nous avons, nous avons utilisé dans l'ensemble de nos travaux le logiciel MATLAB version 7.12.

La méthode a été appliquée pour les mêmes descripteurs du tableau 6 sélectionnés par les 3 différents méthodes (RM, AG, et SW).

- Une couche d'entrée dont les neurones reçoivent l'information présentée au réseau et qui est constituée de 6 neurones (nombre de descripteurs).

- Une couche de sortie qui fournit les résultats de traitement réalisé par le réseau artificiel et contient une couche représentant la toxicité p^{DL50}.
- Une couche intermédiaire appelée couche cachée, elle contient un nombre variable de neurones à déterminer.

Afin de déterminer le nombre de neurones dans une couche cachée, une méthode pratique a été utilisée qui est la procédure essai-erreur.

Nous avons optimisé le nombre de neurones cachés en commençant par un seul neurone en arrivant jusqu'à 12 neurones.

Les résultats obtenus pour chaque méthode de sélection de variables sont rassemblés dans les tableaux et figures suivants :

V.3.1.1.3 La méthode d'algorithme génétique AG :

Tableau 8: Variation des paramètres en fonction du nombre de neurones de la couche cachée.

N	12N	11N	10N	9N	8N	7N	6N	5N	4N	3N	2N	1N
R^2	90,251	91,596	86,657	84,751	81,2158	66,42	60,78	72,53	36,77	48,065	35,494	31,52
S	0,2266	0,2104	0,265	0,2834	0,3145	0,421	0,454	0,3803	0,577	0,5229	0,5828	0,601
$RMSE_{Tr}$	0,2254	0,2093	0,2637	0,2819	0,3129	0,418	0,452	0,3784	0,5741	0,5203	0,5798	0,597
$RMSE_{Test}$	0,298	0,2054	0,2851	0,2731	0,32	0,372	0,509	0,4579	0,5462	0,5988	0,5759	0,557
$RMSE_{Val}$	1,0993	1,549	1,4552	1,1273	1,0348	0,892	1,107	0,9579	0,8013	0,8929	0,8605	0,854

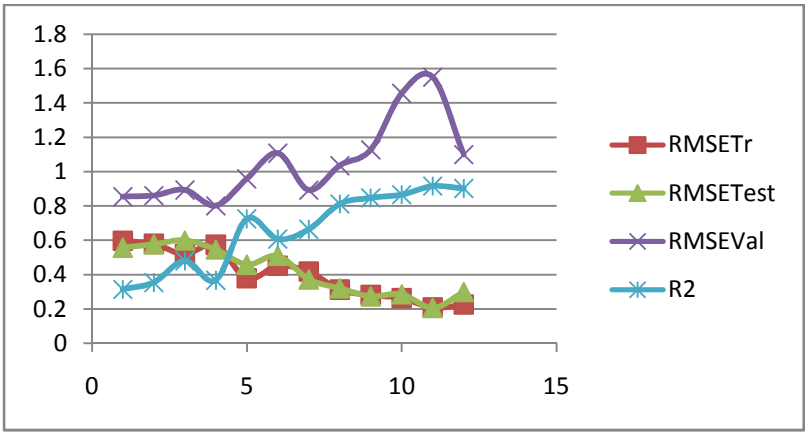


Figure 12: Variation des RMSE en fonction du nombre de neurone pour la méthode d'algorithme génétique.

V.3.1.1.4 La méthode de remplacement :

Tableau 9: Variation des paramètres en fonction du nombre de neurones de la couche cachée par la méthode de remplacement.

	12N	11N	10N	9N	8N	7N	6N	5N	4N	3N	2N	1N
R^2	99,64	97,39	95,99	92,52	93,65	89,08	55,6075	73,282	67,025	59,863	52,82	45,392
S	0,043	0,117	0,145	0,198	0,183	0,24	0,4835	0,3751	0,4167	0,4597	0,498	0,5362
$RMSE_{Tr}$	0,043	0,117	0,145	0,197	0,182	0,239	0,481	0,3732	0,4146	0,4574	0,496	0,5335
$RMSE_{Test}$	0,047	0,145	0,117	0,249	0,194	0,174	0,4637	0,3102	0,3775	0,4392	0,485	0,5228
$RMSE_{Val}$	1,105	1,599	0,895	0,939	1,02	1,115	0,8563	0,8826	0,85	1,0066	0,918	0,9299

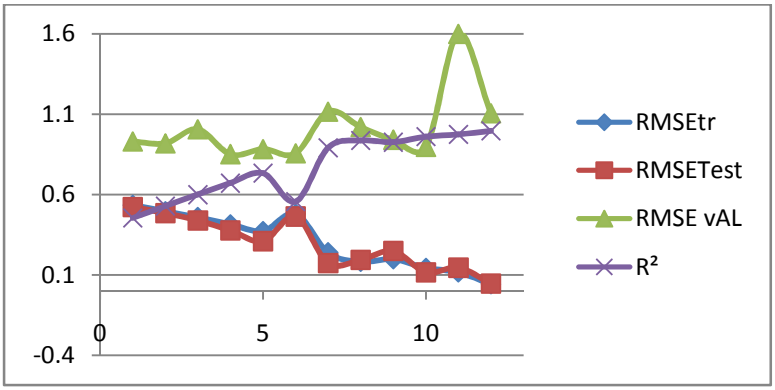


Figure 13: Variation des RMSE en fonction du nombre de neurones pour la méthode de remplacement

V.3.1.1.5 La méthode stepwise (SW):

Tableau 10: Variation des paramètres en fonction du nombre de neurones de la couche cachée par la méthode stepwise.

	12N	11N	10N	9N	8N	7N	6N	5N	4N	3N	2N	1N
R^2	98,82	98,457	96,50	93,22	91,10	90,69	78,11	70,29	65,30	36,585	35,459	32,49
S	0,078	0,090	0,136	0,188	0,216	0,221	0,339	0,395	0,427	0,577	0,583	0,596
$RMSE_{Tr}$	0,078	0,089	0,135	0,188	0,215	0,220	0,337	0,393	0,425	0,574	0,58	0,593
$RMSE_{Test}$	0,067	0,067	0,153	0,227	0,276	0,240	0,393	0,356	0,390	0,593	0,643	0,641
$RMSE_{Val}$	2,321	1,532	1,504	1,358	1,778	1,205	1,137	1,129	0,921	0,727	8,167	0,741

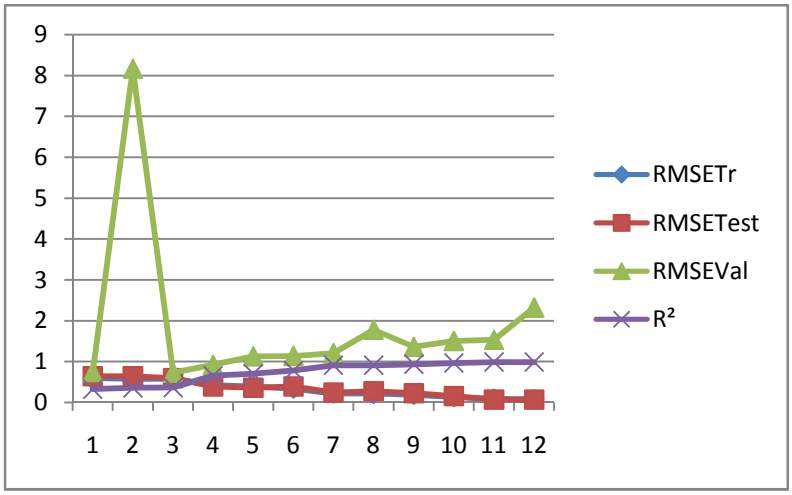


Figure 14: Variation des RMSE en fonction du nombre de neurones pour la méthode de Stepwise.

V.4 Discussion:

Selon les figures et les tableaux précédents, nous constatons que quelques soit la méthode de sélection des variables explicatives utilisée (RM, SW et AG) et pour n'importe quelle architecture du réseau de neurones, les performances des modèles en prédiction externes mesurées en $RMSE_{Ext}$ sont faibles. Par exemple :

- Pour la RNA basée sur les descripteurs sélectionnés par AG, le meilleur R^2 est 91,59% avec un $RMSE_{Test}$ = 0,2054 alors que celui de validation est de 1,549.
- Pour la RNA basée sur les descripteurs sélectionnés par SW, le meilleur R^2 est 98,83% avec un $RMSE_{Test}$ = 0,0679 alors que celui de validation est de 2,3219.
- Pour la RNA basée sur les descripteurs sélectionnés par RM, le meilleur R^2 est 99.64% avec un $RMSE_{Test}$ = 0.047 alors que celui de validation est de 1.105.

On peut conclure que la modélisation sur l'ensemble des 141 composés n'est pas possible avec un nombre réduit de descripteurs comme préconisé par différents chercheurs dans le domaine QSAR/QSPR.

V.5 Modélisation des 52 dérivés benzéniques :

A partir des résultats précédents, nous avons classé les molécules étudiées selon les fonctions chimiques. Les dérivés azotés représentent le plus grand pourcentage (37%).

Le tableau 11 représente la nouvelle base de données sélectionnée qui ne compte que les 52 composés azotés.

Nous avons partagé les données expérimentales en 2 jeux distincts :

- Le jeu d'entraînement qui va permettre la construction du modèle, est composé de 36 molécules.
- Le jeu de validation qui va servir à calculer la prédictivité du modèle est constitué de 16 molécules.

Pour ce faire nous avons fait appel à l'algorithme de Kennard et Stone (KS).

Tableau 11: Données expérimentales sélectionnées (52 dérivés azotés)

N°	Composé	Toxicité observée pDL_{50}
01	1-Methyl-2,3,6-trinitrobenzene	3.04
02	1-Aminobenzene	3.58
03	1-Amino-2-chlorobenzene	3.63
04	1-Methyl-2-nitrobenzene	3.48
05	1-Methyl-3-nitrobenzene	3.24
06	1-Methyl-4-nitrobenzene	3.8
07	1,3-Dinitrobenzene	3.8
08	1-Amino-2-methyl-3-nitrobenzene	3.79
09	1-Amino-3-nitro-4-methylbenzene	4.54
10	1-Methyl-2,3-dinitrobenzene	4.21
11	1-Methyl-2,6-dinitrobenzene	4.7
12	1-Methyl-3,4-dinitrobenzene	5.43
13	1-Methyl-3,5-dinitrobenzene	5.85
14	1-Amino-2,4-dinitro-3-methylbenzene	4.33
15	1-Amino-3,5-dinitro-4-methylbenzene	3.93

Tableau 11: suite

N°	Composé	Toxicité observée pDL_{50}
16	1-Methyl-2,4,6-trinitrobenzene	3.73
18	1-Amino-2,4-dinitrobenzene	4.81
19	1-Amino-2-chloro-4-methylbenzene	4.02
20	1-Amino-2-chloro-4-nirobenzene	4.14
21	1-Amino-2,3,4-trichlorobenzene	5.19
22	1,3,5-trichloro-2,4-trinitrobenzene	5.31
23	1-Amino-2,3,5,6-tetrachlorobenzene	4.73
24	1-Cyano-2-amino-5-chlorobenzene	4.81
25	1-cyano-2-amino-6-methylbenzene	4.83
26	1-Cyano-2-methylbenzene	3.69
27	1-Amino-2,3,4,5,6-pentafluorobenzene	4.39
28	1-Fluoro-4-nitrobenzene	4.12
29	3-Nitrobenzonitrile	5
30	4-Nitrobenzonitrile	4.16
31	2-Amino-4-nitrotoluene	3.88
32	1,5-Dimethyl-2,4-dinitrobenzene	4.27
33	1,3,5-trinitrobenzene	3.36
34	5-Amino-2,4-dinitro-1-methylbenzene	5.72
35	1,4-dinitrotoluene	5.26
36	1-Methyl-2,5-dinitrobenzene	4.91

Tableau 11: suite et fin

N°	Composé	Toxicité observée pDL_{50}
37	1-Amino-2-methyl-4-nitrobenzene	3.77
38	1-Amino-2-methyl-5-nitrobenzene	4.89
39	1-Amino-2-methyl-6-nitrobenzene	5
40	1-Amino-3-methyl-6-nitrobenzene	4.3
41	1-Amino-2-nitro-4-methylbenzene	4.74
42	1-Methyl-2,4-dinitrobenzene	4.46
43	1-Amino-2-methyl-3,5-dinitrobenzene	4.88
44	1-Amino-2,6-dinitro-3-methylbenzene	4.07
45	1-Amino-2,6-dinitro-4-methylbenzene	3.6
46	1-Amino-4-fluorobenzene	4.26
47	3-Amino-4-nitrotoluene	4.91
48	4- Amino-2-nitrotoluene	5.07
49	2-Amino-4,6-dinitrotoluene	4.33
50	3-Amino-2,6-dinitrotoluene	3.56
51	4-Amino-2,6-dinitrotoluene	4.34
52	1-Amino-2-methyl-3,6-dinitrobenzene	3.56

Pour relier la structure des molécules à la propriété expérimentale –logDL50, nous avons utilisé en premier la régression linéaire multiple.

Le nombre de descripteurs est fixé à 6 descripteurs comme précédemment sélectionnés par la méthode de remplacement qui s'avère la meilleure.

L'analyse de regression linéaire multiple a été réalisée avec les logiciels MobyDigs et QSARINS

V.6 Résultats et discussion:

Modèle QSAR sélectionné à 6 descripteurs est le suivant :

$$-\log(DL50) = 3.27 + 0.971\ G2s - 0.140\ RDF025u + 5.76\ R5e+ - 2.48\ R5m - 1.51\ EEig11x + 0.265\ Am$$

Les symboles et la signalisation des descripteurs optimaux sélectionnés sont les suivants:

- G2s : l'indice WHIM 2st composante de symétrie directionnelle pondérée par les états électrotopologiques atomiques. C'est parmi les descripteurs WHIM.

- RDF025u correspond à la fonction de distribution radicale -2.5/non pondéré. C'est parmi les descripteurs RDF

- R5e+ correspond à l'autocorrélation maximale R de distance topologique 5/pondérée par les électronégativités atomiques de Sanderson. Il est parmi les descripteurs GETAWAY.

- R5m correspond à l'autocorrélation R de distance topologique 5/ pondérée par les masses atomiques. Il appartient aux descripteurs GETAWAY.

- EEig11x désigne la valeur propre 11 de la matrice d'adjacence pondérée par les degrés de bord. C'est parmi les indices d'adjacences de bord.

- Am est l'indice de taille totale /pondéré par des masses atomiques. C'est parmi les descripteurs WHIM.

Les résultats obtenus sont rassemblés dans le tableau 12.

Tableau 12: Paramètres statistiques obtenus par la méthode Régression multi-linéaire (MLR) sur les 52 dérivés azotés.

Méthode	R^2	R^2_{aj}	S	F	K_{xx}	$RMSE_{tr}$	Q^2_{LOO}	$RMSE_{CV}$	$RMSE_{ext}$	$Q^2_{F_3}$
RM	0.718	0.660	0.4188	12.32	0.377	0.3759	0.579	0.4595	0.5183	0.4643

Les résulats statistiques réunis dans le tableau 12 permettent de faire des comparaisons et de tirer plusieurs conclusions.

Les valeurs assez élevées de R^2 et de R^2_{aj} montrent, la qualité de l'ajustement, le modèle est hautement significatif vu la valeur élevée de la statistique F de Fisher qui est 12,32 alors que que d'après les tables cette valeur est limitée à 2,41.

La validation interne ou validation croisée (*cross validation* ou CV) mesure la robustesse du modèle c'est-à-dire sa capacité à rester corrélé à la propriété quand on modifie légèrement les données (suppression d'une ou plusieurs données).

Leave –One-Out (LOO) correspond à une cross validation en K-fold.

Ce paramètre permet d'évaluer la robustesse du modèle c'est-à-dire la stabilité du modèle QSAR.

Ce dernier est représenté par la valeur Q^2_{LOO} qui est égale à 0.579.

Pour la validation externe qui est représentée dans le tableau par le facteur$Q^2_{F_3}$, cette valeur de 0.4643 n'est vraiment significative malgré l'amélioration de tous les paramètres statistiques par rapport au modèle précédent utilisant l'ensemble des 141 dérivés benzéniques.

V.6.1 La méthode réseaux de neurones artificiels pour la modélisation des 52 dérivés azotés :

La méthode a été appliquée pour les mêmes déscripteurs sélectionnés par la méthode RM.

Le réseau utilisé est un réseau de neurones multicouches, constitué de 3 couches:

- Une couche d'entrée dont les neurones reçoivent l'information présentée au réseau et qui est constituée de 6 neurones (le nombre de descripteurs).
- Une couche de sortie qui fournit les résultats de traitement réalisé par le réseau artificiel et contient une couche représentant la toxicité p^{DL50}.
- Une couche intermédiaire appelée couche cachée, elle contient un nombre variable de neurones à determiner.

Afin de déterminer le nombre de neurones dans une couche cachée, nous avons appliqué la méthode utilisée précédemment pour l'ensemble des 141 dérivés benzéniques qui est la procédure essai-erreur.

Nous avons optimisé le nombre de neurones cachés en commençant par un seul neurone en arrivant jusqu'à 6 neurones.

Les résultats obtenus sont rassemblés dans les tableaux et figures suivants :

Tableau 13 : Variation des RMSE en fonction du nombre de neurones pour les 52 dérivés azotés.

N	1	2	**3**	4	5	6
R^2	0,5365	0,8029	**0,8776**	0,8437	0,9254	0,9416
Q^2	0,5365	0,8029	**0,8776**	0,8437	0,9254	0,9416
Q^2_{ext}	0,4928	0,7789	**0,8527**	0,8094	0,7144	0,7952
SDEC	0,4821	0,3144	**0,2477**	0,2800	0,1934	0,1711
$SDEP_{ext}$	0,5043	0,3330	**0,2718**	0,3092	0,3785	0,3205
S	0,5371	0,3502	**0,2760**	0,3119	0,2155	0,1907

Dans ce premier cas nous avons fixé le nombre d'itération à 60 cycles ensuite nous avons varié le nombre de neurones afin de déterminer la couche intermédiaire.

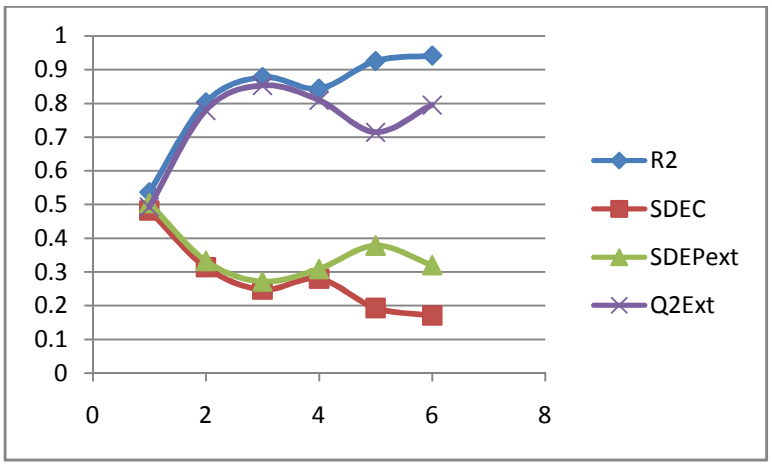


Figure 15: Variation des paramètres en fonction du nombre de neurones pour les 52 dérivés azotés.

A partir de la figure 15 et le tableau 13 précédents, nous déterminons le nombre de neurones de la couche cachée qui est 3.

Afin de confirmer le nombre d'itération, nous procédons à la variation des itérations de 10 à 100 cycles en fixant le nombre de neurones de la couche cachée à 3 (qui a été détreminée)

Tableau 14 : Variation des paramètres en fonction du nombre d'itération pour les 52 dérivés azotés

I	10	20	30	40	50	60	70	80	90	100
R2	0,8077	0,817	0,8558	0,8577	0,8691	**0,8776**	0,8668	0,899	0,876	0,8859
$Q2_{ext}$	0,6201	0,6778	0,7138	0,7037	0,7739	**0,8527**	0,8016	0,6485	0,6808	0,7257
SDEC	0,3105	0,3029	0,2689	0,2672	0,2562	**0,2477**	0,2584	0,225	0,2494	0,2392
$SDEP_{ext}$	0,4365	0,4019	0,3789	0,3855	0,3368	**0,2718**	0,3154	0,4199	0,4001	0,3709
S	0,346	0,3375	0,2996	0,2977	0,2854	**0,276**	0,2879	0,2507	0,2779	0,2665

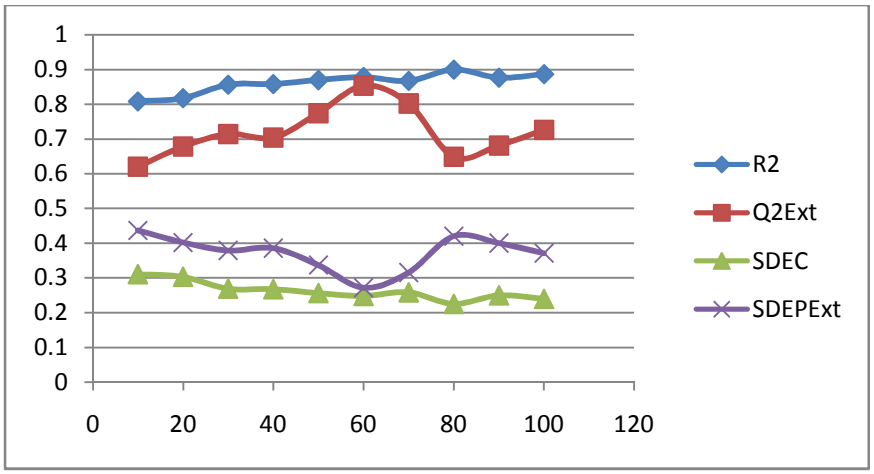


Figure 16 : Variation des paramètres en fonction du nombre d'itérations pour les 52 dérivés azotés.

D'après le tableau 14 et la figure 16, nous confirmons l'architecture du réseau de neurones qui est : (6,3,1) c'est-à-dire 6 descripteurs, 3 neurones dans la couche cachée et 1 neurone dans la couche de sortie avec 60 cycles (60 itérations)

Le modèle sélectionné a pour paramètres statistiques :

Tableau 15: valeurs de la toxicité observée, calculée et les résidus pour le modèle sélectionné.

Composé	Ensemble	-logpDL$_{50}$	Tox préd	Résidus ei
01	Training	3,04	2,9166	0,1234
02	Training	3,58	3,577	0,003
03	Training	3,63	3,5789	0,0511
04	Training	3,48	3,4865	0,0065
05	Training	3,24	3,2795	0,0395
06	Training	3,8	3,9997	0,1997
07	Training	3,8	3,8848	0,0848
08	Training	3,79	3,7793	0,0107
09	Training	4,54	4,5325	0,0075
10	Training	4,21	4,3315	0,1215
11	Training	4,7	4,7387	0,0387
12	Training	5,43	5,4928	0,0628
13	Training	5,85	5,2929	0,5571
14	Training	4,33	3,9575	0,3725
15	Training	3,93	3,9898	0,0598
16	Training	3,73	3,672	0,058
17	Training	4,99	4,8395	0,1505
18	Training	4,81	4,7946	0,0154
19	Training	4,02	4,0685	0,0485
20	Training	4,14	4,3135	0,1735

Tableau 15: Suite

Composé	Ensemble	-logpDL$_{50}$	Tox préd	Résidus ei
21	Training	5,19	5,1939	0,0039
22	Training	5,31	5,6097	0,2997
23	Training	4,73	4,7298	0,0002
24	Training	4,81	4,6569	0,1531
25	Training	4,83	4,8217	0,0083
26	Training	3,69	3,7139	0,0239
27	Training	4,39	4,5135	0,1235
28	Training	4,12	4,0119	0,1081
29	Training	5	4,5926	0,4074
30	Training	4,16	4,1837	0,0237
31	Training	3,88	3,9928	0,1128
32	Training	4,27	5,1805	0,9105
33	Training	3,36	3,6638	0,3038
34	Training	5,72	5,1123	0,6077
35	Training	5,26	5,2601	0,0001
36	Training	4,91	4,907	0,003
37	Test	3,77	3,8501	0,0801
38	Test	4,89	5,0125	0,1225
39	Test	5	5,1151	0,1151
40	Test	4,3	4,7357	0,4357

Tableau 15: Suite et fin

Composé	Ensemble	-logpDL$_{50}$	Tox préd	Résidus ei
41	Test	4,74	4,6034	-0,1366
42	Test	4,46	4,7024	0,2424
43	Test	4,88	4,6501	-0,2299
44	Test	4,07	3,8796	-0,1904
45	Test	3,6	3,2015	-0,3985
46	Test	4,26	3,9025	-0,3575
47	Test	4,91	4,7344	-0,1756
48	Test	5,07	5,1706	0,1006
49	Test	4,33	4,4509	0,1209
50	Test	3,56	3,9588	0,3988
51	Test	4,34	3,9887	-0,3513
52	Test	3,56	3,9677	0,4077

En se basant sur les résultats statistiques donnés par les deux modèles RNA et MLR, nous observons que le modèle RNA a une fiabilité et une capacité prédictive nettement supérieures à celles données par le modèle MLR. Les erreurs types de calcul sont plus faibles, les coefficients de détermination des ensembles d'apprentissage et de prédiction sont plus élevés avec le modèle RNA qu'avec le modèle MLR.

Tableau 16: Paramètres statistiques obtenus par la méthode de Réseaux de neurones artificiels (RNA) sur les 52 dérivés azotés.

R^2	Q^2_{ext}	SDEC	$SDEP_{ext}$	S	MAE_{tr}	MAE_{val}	R^2_{Ext}
0,8776	0,8527	0,2477	0,2718	0,276	0,1465	0,2415	0,8779

La valeur de R^2 montre la qualité de l'ajustement alors la faible différence entre le SDEC et le $SDEP_{ext}$ renseigne sur la robustesse du modèle.

Les validations statistiques externes Q^2_{ext} et R^2_{Ext} attestent la bonne capacité prédictive.

V.7 Conclusion :

L'étude QSAR sur la base de données des 141 dérivés benzéniques a donné des résultats insatisfaisants. Ce qui nous a incités à réduire notre ensemble à modéliser en se limitant aux 52 dérivés benzéniques.

La toxicité –logDL50 des 52 dérivés azotés a été corrélée avec les descripteurs théoriques calculés et sélectionnés par la méthode de remplacement parmi les 1664 descripteurs moléculaires obtenus à l'aide du logiciel DRAGON 5.4.

Les toxicités de 16 dérivés azotés ont été choisies pour former l'ensemble de validation externe.

Le modèle QSAR obtenu par la méthode des réseaux de neurones artificiels RNA présenté, est robuste, avec une bonne capacité prédictive interne et externe, et une bonne qualité de l'ajustement.

Références bibliographiques:

[1] C. Hansch, P.P. Maloney, T. Fujita, R.M. Muir.*Nature*.,**1962**, 194, 178-180.

[2] I. Lessigiarska, A.P. Worth, T. I. Netzeva.*Comparative review of QSARs for acute toxicity*. EUR report No. 21559 EN. EC Joint Research Centre, Ispra, Italy, **2005b**.

[3] F. R. Burden. *J. Chem. Inf. Comp. Sci*.,**2001**,41, 830-835.

[4] P.Gramatica, M.Vighi, F. Consolaro, R. Todeschini, A. Finizio, M. Faust. *Chemosphere*.,**2001**, 42, 873-883.

[5] J.A.Grodnitzky,J.R. *J. Agr. Food. Chem*., **2002**, 50,4576-4580.

[6] J. Huuskonen.*Chemosphere*.,**2003**, 20,949-953.

[7] K. Rose, L. H. Hall. *SAR. QSAR. Environ. Res*., **2003**, 14, 113-129.

[8] L. B. KIER,L.H. HALL. *Molecular structure sescription: The electrotopological state*.Academic Press, California, **1999**.

[9] ChemOffice, *PerkinElmer Informatics*, **2010**.

[10] ACDLABS 10, *Advanced Chemistry Development Inc., Toronto, ON, Canada,* **2015**.

[11] Marvin Sketch 5.11.4, *Chem Axon*, **2012**.

[12] QSARINS 2.2, 2015 University of Insubria, Varese, Italy, http://www.qsar.it

[13] MATLAB 7.9.0 (R2009b) and Statistics Toolbox Release, ―The Math Works‖, *Inc., Natick,Massachusetts, United States,* **2011**.

[14] A.A. Toropov, T. W. Schultz. *J.Chem. Inf. Comput.Sci*., **2003**, 43,560-567.

[15] I. N.Levine.*Quantum Chemistry*. 5ème, Ed, New Jersey: Prentice-Hall, **2000**.

[16] A. H. Morales, M. A. C. Pérez, R. D. Combesc, M. P. Gonzàlez.Quantitative structure activity relationship for the computationalprediction of nitrocompounds carcinogenicity. *Toxicology*.,**2006**, 220, 51–62.

Conclusion générale

Le besoin de mesurer l'impact des polluants sur l'environnement et la santé humaine est en constante augmentation. Ce qui exige de nombreuses mesures, nécessitant diverses techniques analytiques qui peuvent s'avérer contraignantes (temps et coût).

Une alternative à la mesure systématique, est le recours aux méthodes théoriques.

L'objectif de cette thèse était de développer des modèles QSAR/QSRR pour la prédiction d'une propriété et activité biologique d'une famille de composés aromatiques monocycliques « les dérivés benzéniques ».

Un grand nombre de descripteurs a été calculé (descripteurs constitutionnels, électroniques, topologiques, géométriques, physico-chimiques……) à l'aide du logiociel DRAGON.

Deux méthodes statistiques ont été utilisées dans la construction des modèles (RML, RNA). Les principales techniques de validation ont été utilisées (les tests statistiques standards, la validation interne, la validation externe, les domaines d'applicabilité…..).

Nous avons présenté deux applications que nous avons accomplies.

Dans la première application, nous avons établi un modèle QSRR, reliant 2 descripteurs moléculaires avec l'indice de rétention de 38 dérivés benzéniques séparés par chromatographie en phase gazeuse, prélevés dans la littérature.

Les résultats obtenus montrent que le modèle QSRP présenté est robuste, avec de bonnes capacités prédictives internes et externes et une bonne qualité d'ajustement.

Dans la deuxième application nous avons utilisé la RLM et la RNA pour établir des modèles QSAR reliant des descripteurs moléculaires calculés avec la toxicité des 141 dérivés benzénique correspondent à la concentration du composé pour laquelle 50% des animaux meurent en 96 heures.

La mavaise prédiction sur l'ensemble des 141 dérivés benzéniques, qui est du à l'hétérogéité des composés, nous a incités à réduire l'ensemble à modéliser en se limitant aux 52 dérivés azotés qui forment un ensemble homogène.

D'après les résultats obtenus sur cette dernière base de données, le RNA a une capacité prédictive nettement meilleure que la RLM.

Le modèle construit par la méthode RNA est robuste, avec une bonne capacité prédictive interne et externe.

Les problèmes rencontrés lors de ce travail pour la modélisation des 141 dérivés benzéniques, nous incitent à utiliser d'autres méthodes plus puissantes pour l'optimisation de la géométrie moléculaires telles que DFT, Ab intio.

Nous pourrions aussi user de moyens de régression non linéaires plus récentes et peut être plus adéquats à la modélisation d'un tel ensemble hétérogène de composés (SVM ?).

ISSN: 0975-8585

Research Journal of Pharmaceutical, Biological and Chemical Sciences

Quantitative Structure/Retention Relationship Study of Benzene Derivatives.

Karima DJELLOUL MOKRANI[1,2]* , Hamza HADDAG[1], and Djelloul MESSADI[1].

[1]Environmental and Food Safety Laboratory, Badji Mokhtar - Annaba University, Algeria
[2]Functional and Evolutionary Ecology Laboratory, University Chadli Bendjedid El Tarf. Algeria

ABSTRACT

Retention indices of 38 benzene derivatives, separated by gas chromatography were correlated with 2 connectivity indices, using PM3 semi-empirical calculation method and hybrid genetic algorithms/ multiple linear regression approach. For the sake of external validation, the available set of chemicals was separated using Kennard and Stone algorithm into training set of 28 compounds and an external set of 10 compounds. The proposed hybrid model was validated using different criterions, and its predictive capability meets the conditions defined by Golbraikh et al. In comparison to the previously published model, our model exhibits a large enhancement and its mechanistic interpretation was attempted to connect the selected variables to the retention phenomenon.

Keywords: Benzene derivatives- Kováts index- QSRR- Internal and external predictivity validation-Chemical applicability domain.

Corresponding author

ISSN: 0975-8585

INTRODUCTION

The mono-aromatic hydrocarbons, which are often present in the urban environments, constitute an important source of pollution and health hazard[1].

Hence, the development of reliable structural identification and quantification of these substances is imperative. Gas chromatography coupled with the mass spectrometry or infrared analysis by Fourier transform is largely used to this aim. Nevertheless, the measurement of their retention indices constitutes, even nowadays, a simple means of effective, sensitive and affordable identification.

Advantageously, any parameter of retention can be derived a priori from the molecular structure of the considered compound. The prediction of the retention indices of a set of 38 benzene derivatives separated by isothermal gas chromatography was obtained by Jalali-Heravi and Garkani-Nejad [2] who adopted a QSRR approach [3](Quantitative Structure-Retention/Relationship). Based on a training set of 32 randomly selected compounds, linear models were developed following a stepwise involving successive additions of 58 variables (topological, geometrical and electronic) along with previously characterized physical properties. Although widely used, the disadvantage of this approach is that it cannot account for combined effects since each variable is considered separately.

Genetic algorithms [4,5], based on the stochastic search, constitute an alternative method of choice for the selection of variables subsets (VSS: Variable Subset Selection).

The optimization of the molecules geometry, necessary to the calculation of certain descriptors was conducted by applying the MNDO (Modified Neglect of Diatomic Overlap) semi-empirical method [6] while MNDO is known to be ineffective when calculating the molecular structures and the heats of formation of the molecules containing fluorine [7].

The model includes four descriptors (XV0: valence connectivity index of order zero; $NOCH_3$: number of methyl groups in the molecule; VOL: Van der Waals volume of the molecule; DIMO: Dipole moment of the molecule) [2] is validated using the following set of parameters: the coefficient of determination R^2, Fisher parameter F and the standard deviation S,while the application domain of this model is not defined.

In addition, for models including more than two descriptors, low coefficients of correlation cannot positively ensure the complete independence of the descriptors. This aspect was not assessed by Jalali-Heravi and Garkani-Nejad [2]. Finally, the predictive capability of the proposed model was tested by calculating the retention indices of the six compounds not retained for its construction.

In this work, we proposed a statistical linear model using the same database by calculating the molecular descriptors with the software Dragon [8]. This statistical linear model is justified using different criteria and its prediction capability is assessed following Golbraikh et al.'s conditions [9,10].

Finally, the applicability domain (AD)is discussed using the Williams plot [11,12] that represents the standardized residual of predictions versus the leverage values (h_{ii}).

The semi-empirical method PM3 (Parametric Method 3) [13] was useful for optimization of the geometry of the molecules. It consists in re-parameterization of AM1 method (Austin Model1) [14] that is itself an improved version of the MNDO method.

It is important to define rationally the training set during the construction of the model and, for its assessment an external test set comprising 15 to 40% of the available data. The available set of chemicals was preliminary separated using Kennard and Stone algorithm [15].The hybrid approach Genetic Algorithm Multiple Linear Regression(GA-MLR) was adopted in our work.

ISSN: 0975-8585

MATERIALS AND METHODS

Database:

The Kováts indices of the 38 benzene derivatives (table1) were extracted from reference [2]that also provides a detailed description of the conditions of the chromatographic separation.

The extreme values are 664.1 and 1287.7 index units (iu) with an average of 965.2 iu.

Table1: Retention indices (experimental and calculated) and values of the used descriptors

N°	Compounds	CAS Number	$RI_{Experimental}$	$RI_{Calculated}$	$X0Av$	$X1sol2$
01	Benzene	71-43-2	681.3	689.94	0.577	9.0000
02	Fluorobenzene	462-06-6	664.1	678.31	0.538	9.0000
03	Chlorobenzene	108-90-7	877.9	863.89	0.646	13.5645
04	Bromobenzene	108-86-1	979.6	973.14	0.764	15.7688
05	Toluene	108-88-3	788.2	789.50	0.627	11.5192
06	Anisole	100-66-3	923.6	913.61	0.599	15.4606
07	p-Chloroanisole	623-12-1	1131.7	1124.67	0.650	21.2890
08	p-Xylene	106-42-3	889.2	895.62	0.664	14.3489
09	p-Fluorotoluene	352-32-9	777.7	777.28	0.586	11.5192
10	p-Bromotoluene	106-38-7	1096.3	1089.48	0.784	19.0532
11	p-Bromofluorobenzene	460-00-4	940.9	955.86	0.706	15.7688
12	p-Chlorobromobenzene	106-39-8	1174.4	1182.14	0.801	21.6597
13	m-Chloroanisole	2845-89-8	1126.0	1124.67	0.650	21.2890
14	m-Methylanisole	100-84-5	1029.6	1033.68	0.635	18.7143
15	m-Xylene	108-38-3	892.0	895.62	0.664	14.3489
16	m-	108-37-2	1179.0	1182.14	0.801	21.6597
17	m-Bromotoluene	591-17-3	1100.0	1089.48	0.784	19.0532
18	m-Fluorotoluene	352-70-5	778.0	777.28	0.586	11.5192
19	m-Dibromobenzene	108-36-1	1287.7	1306.01	0.905	24.4234
20	o-Methylanisole	578-58-5	1013.5	1038.63	0.635	18.8616
21	o-Chloroanisole	766-51-8	1135.6	1129.96	0.650	21.4462
22	o-Bromofluorobenzene	1072-85-1	959.6	955.86	0.706	15.7688
23	o-Xylene	95-47-6	916.2	899.96	0.664	14.4780
24	o-Bromochlorobenzene	694-80-4	1197.6	1187.46	0.801	21.8182
25	p-Methylanisole	104-93-8	1029.5	1033.68	0.635	18.7143
26	o-Bromotoluene	95-46-5	1095.7	1094.47	0.784	19.2019
27	m-Fluoroanisole	456-49-5	908.5	903.77	0.566	15.4606
28	p-Chlorotoluene	106-43-4	989.2	976.50	0.680	16.6138
29	p-Fluoroanisole	459-60-9	910.6	903.77	0.566	15.4606
30	m-Chlorotoluene	108-41-8	990.9	976.50	0.680	16.6138
31	m-Chlorofluorobenzene	625-98-9	835.4	851.08	0.603	13.5645
32	m-Dichlorobenzene	541-73-1	1060.5	1063.55	0.697	19.0532
33	o-Fluorotoluene	95-52-3	777.4	777.28	0.586	11.5192
34	o-Fluoroanisole	321-28-8	919.7	903.77	0.566	15.4606
35	o-Chlorofluorobenzene	348-51-6	862.0	851.08	0.603	13.5645
36	p-Chlorofluorobenzene	352-33-0	840.5	851.08	0.603	13.5645
37	o-Chlorotoluene	95-49-8	986.3	981.17	0.680	16.7526
38	m-Bromofluorobenzene	1073-06-9	932.8	955.86	0.706	15.7688

Descriptors calculation:

We have used the Hyperchem[16] to represent each molecule, whose geometry is initially pre-optimized by molecular mechanics calculation. Then for each molecule we have determined its (x,y,z) atomic coordinates corresponding to the conformation of lowest energy determined by the PM3 method. All calculations were carried in the frame of the Hartree Fock formalism with spin constraint (or RHF: for Restricted Hartree-Fock) without confirmation interaction.

The molecular structures were optimized; according to the Polack-Ribiere algorithm adopting a stopping criterion corresponding to a mean square root of the gradient of 0.001 kcal/mol. Following this optimization, the molecule geometries were transferred to Dragon software[8] for the calculation of 1664 descriptors belongingto20 different classes. The descriptors of the same group exhibiting constant values(standard deviation lower than 0.0001)provide no information and thus, are removed from subsequent analysis. Similarly, two highly correlated descriptors $r \geq 0,92$ conveying redundant information automatically exclude one that is correlated with the greatest number of descriptors. Consequently, the initial pool of 1664 descriptors was reduced to 203 elements.

Kennard and Stone algorithm [15]:

It is a sequential technique that maximizes the Euclidean distances between new selected samples and previous analyzed samples. It starts by locating the two most distant samples, which are removed from the original data set and assigned to the training set.

For each sample (sample i) not selected previously, the algorithm calculates its distance to each sample; and assigns to (sample i) the smallest distances.

The sample (sample i) associated with the greatest distance is the furthest of all the samples already selected. The procedure is repeated until the target number of training samples is reached.

This technique has two significant advantages. Selecting the most distant samples from each other introduces diversity across the training set. Obtaining a uniform distribution is another advantage of this technique.

As a result, using the algorithm of Kennard and Stone, the complete data set was divided into two subsets: the training subset containing 28 compounds and validation subset including the 10 remaining compounds.

Model validation development:

The variable subsets selection (VSS) is realized using the genetic algorithm (GA-VSS) by maximizing the prediction coefficient Q^2_{LOO}.

Genetic algorithms are optimization algorithms based on technique derived from genetic and natural evolution mechanisms: i.e, crossing (or crossover) and mutation that are responsible for the generation of new individuals.

In the MobyDigs software [17] such processes are controlled by a user-defined parameter T varying between zero and one, defining the relative extent of crossing and mutation.

In the terminology of genetic algorithms, the binary vector I, called chromosome, is a vector of dimension p where each position (a gene) corresponds to a variable (1: if it appears in the model; zero 0: otherwise. Each chromosome is a model with a subset of variables [4,5].

The genetic algorithm parameters have been defined as follow:

- Model population: Pop = 100

- Maximum number of variables in the model: L = 5, so as to associate a minimum of five compounds to each descriptor; the minimum number is arbitrarily 1
- T value: chosen equal to 0.5 to balance the effects of crossover and mutation.

To avoid models with co-linearity lacking high prediction capacibilities, we have applied the QUIK (Q Under Influence of K) rule [18] based on multivariable correlation index [19] defined as follow:

$$K = \frac{\sum_J \left| \frac{\lambda_j}{\sum_j \lambda_j} - \frac{1}{p} \right|}{2(p-1)/p}; j = 1,...,p \quad \text{et} \quad 0 \le k \le 1 \qquad (1)$$

λ_j :are the eigenvalues of the correlation matrix of the data set $(n \times p)$.

n : The number of objects and p the number of variables.

This rule derives from the assumption that the total correlation in the set formed by predictors **X** of the model, and the response **Y (** K_{xy}) must always be greater than the correlation measured with the set of predictors (K_{xx}) taken separately.

To calculated K_{xy} we have considered the response Y as an x variable and determined the corresponding correlation matrix. Generally [20], models that do not verify the relationship are rejected:

$$D(K) = K_{xy} - K_{xx} > 0,05 \qquad (2)$$

A model with a minimum number of explanatory variables is sought (rule of parsimony). Its variables can be related to the retention phenomena in the apolar stationary phase Apiezon MH used in the analyses and could be easly interpreted.

The model will be justified by means of different statistical parameters (R^2 , R^2_{aj}, Fisher parameter F, standard error S) and by considering the Leave Many Out (LMO) cross- validation, the randomization test of Y, as well as the bootstrap technique.

The adjusted R^2 (R^2_{aj}) calculated using the formula:

$$R^2_{aj} = 1 - \left[\frac{n-1}{n-p-1} \left(1 - R^2\right) \right] \qquad (3)$$

Is a better measure of the percentage of the total variation explained by the model than the coefficient of determination R^2

The Fisher parameter F is defined by the ratio of the average of the squares due to regression to the mean of the squares of the residuals, which to compare the variance explained by the model to the residual variance: a high value of R^2 is proof of the reliability of this model.

The cross validation consists in re-computing the model considering only (n-q) objects and using this new model to predict the dependent variable value of the q excluded compounds. The process is repeated for the n objects of the training set.

If $q = 1$ the technique is called LOO (Leave One Out), otherwise it is LMO (Leave Many Out). A prediction coefficient LOO or LMO designated by Q^2 or R^2_{CV} respectively, is calculated considering the dispersion of the estimation [21]:

ISSN: 0975-8585

$$Q^2 = 1 - \frac{\sum_{i=1}^{n} \left(y_i - \hat{y}_{i/i} \right)^2}{\sum_{i=1}^{n} \left(y_i - \bar{y} \right)^2} \qquad (4)$$

$\hat{y}_{i/i}$: corresponds to the response of the i[th] object using a model obtained without involving this object; y_i, $\overline{y}$: represent respectively, the value of the i[th] observation and the average value of the n observations; the summation covers all of the compounds in training.

In order to establish a nonrandom model, we have applied the randomization test of **Y**(Y-scrambling) [21]. The test consists in generating a vector of the studied propriety by a random permutation of the components of the real vector. Then, we calculate the result QSRR model vector according to the usual method. This process is repeated 100 times in this study. If a high score is reached, the original model is not acceptable.

In the Bootstrop validation technique, we simulate new samples of size (n), by random pooling with reduction. As such, the training set that maintains its initial size (n), is composed of generally, repeated objects, since the set of evaluation includes the removed objects [22,23].

The model is calculated both on the training set and on the predicted responses set combined. This construction procedure of the training and evaluation sets is repeated 3000 times in this study, and an average prediction capacity is calculated Q^2_{BOOT} [23].

The validation of the model has been completed using a test set. Equation (5) details the calculation of Q^2_{EXT} for the test set.

$$Q^2_{EXT} = 1 - \frac{\sum_{i=1}^{n_{EXT}} \left(y_i - \hat{y}_i \right)^2}{\sum_{i=1}^{n_{EXT}} \left(y_i - \bar{y}_{cal} \right)} \qquad (5)$$

y_i , $\hat{y}_i$: are respectively the observed and the predicted values, and $\bar{y}_{cal}$ is the average of the observed values of the training set. The sum considers all the samples of the test set.

According to Golbraikh et al, [9,24] a QSRR model can provide an acceptable prediction if it verifies the following conditions:

$$Q^2_{EXT} > 0,5 \qquad (6-a) \qquad r^2 > 0,6 \qquad (6-b)$$

;

$$\left(r^2 - r_0^2\right)/r^2 < 0,1 \quad \text{or} \quad \left(r^2 - r_0'^2\right)/r^2 < 0,1 \qquad (6-c)$$

$$0,85 \le k \le 1,15 \quad \text{or} \quad 0,85 \le k' \le 1,15 \qquad (6-d)$$

r is the correlation coefficient between the calculated and experimental values in the test set ;
r_0^2 (Calculated versus observed values) and $r_0'^2$ (observed versus calculated values) are the coefficients of determination; k , k' are slopes of the regression lines through the origin of calculated versus observed and observed versus respectively.

Applicability domain:

The applicability domain (AD) is a theoretical region of space defined by the descriptors of the model and the modeled response, for which a given QSRR model is expected to lead to reliable predictions. This region, which depends on the nature of the compounds of the training set, can be characterized in different

ways. In this work the structure of AD has been determined by the leverages approach, defined by the diagonal elements of the **H** matrix that allows, by simple multiplication to associate the vector $\mathbf{y}$ to the vector $\hat{\mathbf{y}}$ .The diagonal h_{ii} element is defined by:

$$h_{ii} = x_i \left(X^T X \right)^{-1} x_i^T \qquad (7)$$

$\boldsymbol{x}_i$ is the line-vector of the compound descriptors i, and **X** the matrix of the model deduced from the descriptors values of all the training set; the exponent T denoting the transposition vector (or matrix).

The h_{ii} element determines the influence of observation i on estimators obtained by the least squares method. A leverage point is an observation that significantly influences the estimators. In practice, an observation i is considered as a point of leverage if:

$$h_{ii} > h^* = 3\left(\sum_i h_{ii} \right) \Big/ n = 3(p+1)/n \qquad (8)$$

The Williams plot displaying the standardized residual of predictions against the leverage values h_{ii} was used with the aim of detecting both **X** outliers (leverage points) and **Y** outliers in (standard residuals higher in absolute values than 3 standard deviation units: $3s$).

RESULTS AND DISCUSSION

Model development and validation:

Table 2 shows that the retention index is linearly correlated to the descriptor $X1sol$, or better to its square $X1sol2$

Table 2: Comparison of the statistical parameters of different models.

n[a]	Descriptors	R^2	R^2_{aj}	Q^2_{LOO}	$Q^2_{L(5)O}$	Q^2_{BOOT}	F	S	DK
28	$X1sol$	98.00	97.7	97.64	97.86	97.64	1247.37	23.16	-
28	$X1sol2$	98.19	98.12	97.95	98.14	97.76	1409.32	21.82	-
28	$X0Av$. $X1sol2$	99.59	99.56	99.47	99.52	99.4	3068.88	10.53	0.123
32[b]	$NOCH_3$. $XV0$. VOL . $DIMO$	99.63	99.57	99.46	99.47	99.36	1814.17	10.12	0.126

[a] Training set compounds, [b] Jalali-Heravi et al. model.

We have adopted the model with 2 descriptors, $X\ 0AV$, $X1sol\ 2$. The corresponding equation, calculated using the centered reduced values is given by:

$$IR = 0.168\ X0Av + 0.872\ X1sol2 \qquad (9)$$

Where $X0Av$ denotes the valence connectivity index of zero order, and $X1sol$ the Solvation connectivity index of the first order [25-26].

The combination of these two descriptors provides an improvement of all statistical parameters as detailed and compared in the table 2. In particular, the standard error is divided by a factor greater than two

(21.82 to 10.58) and close to the value (10.12) obtained with the four descriptor model, published by Jalali-Heravi et al. Also note DK (= 0.123) is higher than the prescribed limit of 0.05.

The obtained statistical parameters provide substantial ground that the proposed model (equation 9) establishes a strong correlation between the 2 selected variables and the studied property, characterized by an excellent coefficient of determination R^2 =99.59%that explains about 99.60% of data variation. In addition, the very high value of the Fisher parameter F (= 3068.88), indicates the excellent capability of the model in the prediction of RI values, with an acceptable standard error (s = 10.53). Equation (9) presents a R^2_{aj} = 99.56 indicating excellent agreement between correlation and variation of the data.

The minor difference between Q^2_{LOO} and $Q^2_{L(5)O}$ informs about the robustness of the model. The cross-validation prediction coefficient provides indication of the reliability of the model when addressing the sensitivity against the elimination of any chosen five data. The value of $Q^2_{Boot}(=99.4)$ confirms both the internal predictability and stability of the proposed model. Figure 1 plots the graph of statistical coefficients Q^2 and R^2 which allows comparing the results for randomized models (circles) to the initial model (rhombus). It appears clearly that retention indices statistics obtained for the modified vectors are lower than those of the real QSRR model. This observation ensures that a real structure/retention relationship has been established.

Figure 1: Graphical representation of the randomization test.

The following statistical parameters obtained for the external tests set verify the well-accepted conditions (6-a to 6-d), which reinforces the predictive capabilities of the present model.

$$Q^2_{EXT} = 0.9869 > 0.5 \quad r^2 = 0.9765 > 0.6$$

$$\left(r^2 - r_0^2\right)/r^2 = \left(0.9765 - 1.000\right)/0.9765 = -0.0240 < 0.1$$

$$\text{or} \quad \left(r^2 - r_0^{'2}\right)/r^2 = \left(0.9765 - 1.000\right)/0.9765 = -0.0240 < 0.1$$

$$0.85 < k = 1.0002 < 1.15 \quad \text{or} \quad 0.85 < k^{'} = 0.9996 < 1.15$$

Application domain:

Figure 2 compares Williams plots derived either from our2-descriptors model, and the 4- descriptors model [2].In both cases, the leverage values of all training and test compounds, are lower than the corresponding critical values h^* (respectively 0.321 and 0.468) and, in both cases, none of the compounds is found influential.

ISSN: 0975-8585

Furthermore, for 2-descriptors model (figure 2-A) all training and test compounds exhibit standard residuals values lower, in absolute value to 3 units of standard deviation (3s), which confirms that the relevance of data set and the removal of insignificant data.
However, two outlier data points are found with the 4-descriptors model (figure 2-B), one of the training set (compound 30: o-xylene) and the other from the test set (compound 38: m-Bromofluorobenzene).

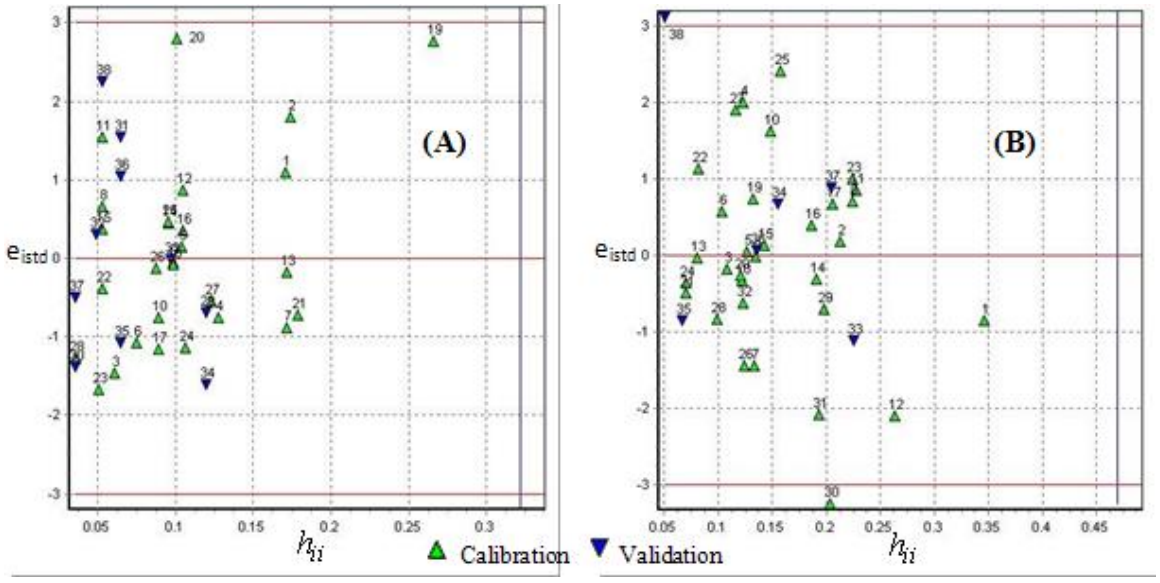


Figure 2: Williams plot for the 2-descriptors model (A) and the 4-descriptors model (B).

Interpretation of the model:

The descriptor X1sol2 which is highly correlated with RI significantly governs the model response as shown by the values of the coefficients of the 2- descriptors model, and the associated Student t values equal to 48.33 for X1sol2 and 9.31 for X0Av respectively.

Solvation connectivity indices are defined [27] for a H-depleted molecular graph, where fluorine atoms as well as hydrogen's are not taken into account, their dimension being very close to that of the hydrogen atom.

The Solvation connectivity index of the first order is derived from the equation:

$$X\,1sol = \frac{1}{4}\sum_{b=1}^{B}\frac{\left(L_i.L_j\right)_b}{\left(\delta_i\delta_j\right)_b^{0,5}} \qquad (10)$$

Where b runs to the number of bonds B, L_i and L_j are the principal quantum numbers of 2 vertices (atoms) incidents to the considered bond; δ_i and δ_j represent the degrees (valences) of the corresponding vertices. The solvation connectivity indices make it possible to model solvation entropy and describe the interactions of dispersion in solution which play a decisive role in the retention phenomenon.

The average valence connectivity index of order zero ($X\,0Av$) is obtained by dividing the valence connectivity index of order zero ($X\,0V$) by the number of edges B (bonds) of the H-depleted molecular graph. $X\,0V$ is defined [28,29] by:

$$X\,0Av = \sum_{i=1}^{N}\left(\delta_i^v\right)^{-0,5} \qquad (11)$$

N is the number of graph vertices, the number of atoms in the molecule other than hydrogen.

δ_i^v is calculated for the atom i, from the expression:

$$\delta_i^V = \frac{Z_i^V - H_i}{Z_i - Z_i^V - 1} \qquad (12)$$

Z_i, Z_i^δ respresent, respectively, the atomic number and the number valence electron of atom i.

Whereas H_i designates the number of hydrogen atoms bounded to the considered atom.

Besides the fact that it introduces relative corrections to the differences between halogen types contained in a given molecule, the descriptor $X\ 0Av$ is related to the size and degree of ramification of the molecules that may have a significant role in the distribution process of the solute between the two chromatographic phases (mobile/stationary)

CONCLUSION

A bi-parametric model was developed for the retention of 38 benzene derivatives separated by gas chromatography on apolar Apiezon MH column.

Solvation connectivity index describes the interactions of dispersion in solution. Whose role is crucial in the phenomenon of retention, while the average valence connectivity index of order zero plays a significant role in the distribution process of the solute between the two chromatographic phases (mobile/stationary).

The selection of these explanatory variables was carried out by genetic algorithm based software MobyDigs among 203 descriptors calculated using the Dragon software.

This optimal model was validated by different statistical approaches using the training set and the external validation set, defined rationally by adopting the Kennard and Stone technique.

The obtained statistical parameters show that the model with two descriptors established a strong correlation between the two selected variables and the studied property, which indicates the excellence of the model in the prediction of retention indices of benzene derivatives.

REFERENCES

[1] Bertinetto C, Duce C, Solaro R, Tiné MR.MATH Commun Math Comput Chem 2013; 70:1005-1021
[2] Jalali-Heravi M, Garkani-Nejad Z.J Chromatogr1993; 648: 389-393.
[3] Kaliszan R.Chem. Rev2007;107: 3212-3246.
[4] Leardi R, Boggia R, Terrile M. JChemom1992; 6: 267-281.
[5] Clark DE .Evolutionary Algorithms in Molecular Design. Wiley-VCH, Weinheim, 2000, 288p.
[6] Dewar M J S, Thiel W. J Am Chem Soc1977; 99: 4899-4907.
[7] Dewar M J S, Rzepa HS.J Am Chem Soc1978;100: 778-784.
[8] R. Todeschini, V. Consonni, M. Pavan ; 2005. DRAGON software for the calculation of molecular descriptors. Release 5.3 for Windows, Milano.
[9] Golbraikh A, Tropsha A.J Comput Aided Mol Des2002; 16: 357-369.
[10] Golbraikh A, Shen M, Xiao Z, Xiao YD, Lee KH, Tropsha A.J Comput Aided Mol Des 2003; 17: 241-253.
[11] Eriksson L, Jaworska J, Worth AP, Cronin MTD, Mc Dowell RM, Gramatica P.Environ HealthPerspect2003; 111: 1361-1375.
[12] Tropsha A, Gramatica P, Grombar VR. QSAR Comb Sci2003, 22: 69-76.
[13] Stewart JJP. JComput Chem1989, 10: 109-221.
[14] Dewar MJS, Zoebisch EG, Healy EF, Stewart JJP.J Am Chem Soc1985; 107: 3902-3909.
[15] Kennard RW, Stone LA.Technometrics1969;11: 137-148.
[16] Hyperchem TM, 2000. Release 6.03 for windows, molecular modeling System.

[17] Todeschini R., Ballabio D., Consonni V., Mauri A., Pavan M., 2009. MOBYDIGS software for multilinear regression analysis and variable subset selection by genetic algorithm. Release for windows, Milano.
[18] Todeschini R, Consonni V, Maiocchi A.Chemom Intell Lab Syst 1998;46: 13-29.
[19] Todeschini R. Anal Chim Acta 1997;348: 419-430.
[20] Todeschini R, Consonni V, Mauri A, Pavan M.Anal Chim Acta 2004; 515: 199-208.
[21] Wold S, Eriksson L. Statistical validation of QSAR results. In: H. Van de Waterbeemd ed. Chemometrics methods in molecular design. VCH, New York, 1995, Vol. 2, pp. 309-318.
[22] Efron B, Tibshirani RJ.An introduction to the bootstrap, Chapman and Hall,1993, 456p.
[23] Wehrens R, Putter H, Buydens LMC.Chemom Intell Lab Syst2000;54: 35-52.
[24] Golbraikh A, Tropsha A. J Mol Graph Model 2002; 20: 269-276
[25] Todeschini R, Consonni V. Handbook of molecular descriptors. Edited by Mannhold R, Kubinyi H, Timmerman H, Wiley-VCH Verlag GmbH, Weinheim, 2000, 688p.
[26] Todeschini R, Consonni V.Molecular descriptors for chemoinformatics. Second, Revised and Enlarged Edition.Vol. I: Alphabetical listing, Series Editors: Mannhold R, Kubinyi H, Folkers G, Wiley-VCH Verlag GmbH CO. KGaA, 2009, 967 p.
[27] Zefirov NS, Polyulin VA.J Chem Inf Comput Sci2001; 41: 1022-1027.
[28] Kier LB, Hall LH.J Pharm Sci1981; 70: 583-589.
[29] Kier LB, Hall LH.J Pharm Sci1983;72: 1170-1173.