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

UNIVERSITÉ BADJI MOKHTAR
- ANNABA-

جامعة باجي مختار- عنابة

Année / 2018

Faculté des sciences
Département de chimie

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

Option : *Chimie Analytique et Environnement*

THÈME

Etudes QSPR des propriétés contrôlant l'évolution de quelques HAP dans l'environnement

Présentée par: BOUARRA Nabil

Devant le jury :

Président	Mme. Linda DIB	Pr	Université Badji Mokhtar Annaba
Rapporteur	Mr. Djelloul MESSADI	Pr	Université Badji Mokhtar Annaba
Examinateurs	Mr. Abdelhak GHEID	Pr	Université Mohamed Chérif Messaadia Souk Ahras
	Mr. Mohamed AbdEsselem DEMS	MRA	Centre de Recherche en Biotechnologie (CRBTs) Constantine
	Mme. Douniazed HANNACHI	MCA	Université de Setif-1
	Mme. Hayet SAIFI	MCA	Université Badji Mokhtar Annaba

REMERCIEMENTS

La reconnaissance est la mémoire du cœur

« Hans Christian Andersen »

À la mémoire de mon père

A ma famille, (au sens large du terme),

A ceux qui me sont chers, qui m'ont toujours encouragé et supporté,

À ceux, qui ont cru en moi, je dédie ce travail.

REMERCIEMENTS

Le travail présenté dans ce mémoire a été réalisé au Laboratoire de sécurité environnementale et alimentaire(LASEA). Je remercie Monsieur le Professeur MESSADI Djelloul, directeur du LASEA, de m'avoir permis d'entreprendre cette thèse parallèlement à mes activités d'attaché de recherche au centre de recherche en analyses physico-chimiques (CRAPC)

Ma reconnaissance va en premier lieu au *Pr MESSADI Djelloul.* Travailler avec lui a été, est une expérience vraiment enrichissante dans la mesure où il parvient à diriger efficacement les recherches en développant un sens et un goût de l'autonomie et de la rigueur scientifique. Je pense avoir énormément appris à son contact, et je lui en suis réellement reconnaissant.

Je remercie ensuite les membres du jury, Madame DIB L, Madame HANNACHI D, Madame SAIFI H, Monsieur GHEID A, et Monsieur DEMS M A pour avoir pris le temps de lire en détail mon manuscrit de thèse et tout particulièrement le rapporteur (Pr. MESSADI Djelloul).

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

Résumé

Le travail présenté dans cette thèse a pour objectifs d'élaborer des modèles QSPR fiables, stables et prédictifs pour la prédiction des propriétés physico-chimiques (température d'ébullition, température de fusion, solubilité aqueuse, indice de rétention en chromatographie gazeuse) de divers ensembles d'hydrocarbures aromatiques polycycliques (HAP).

Le logiciel de modélisation moléculaire HyperChem (V.6.03) a été utilisé pour représenter les molécules, puis à l'aide de la méthode semi-empirique PM3 les géométries finales ont été obtenues. Différents descripteurs moléculaires sont calculés à l'aide du logiciel Dragon (V.5.5).

Des modèles QSPR ont été développés pour la prédiction des propriétés physico-chimiques importantes de divers ensembles de HAP. Des approches basées sur la régression linéaire multiple (RLM), et les réseaux de neurones artificiels (RNA), conduisent à des modèles de qualités différentes. Les algorithmes génétiques (GA), ont été associés pour sélectionner les descripteurs les plus importants,

L'approche hybride algorithme génétique/ régression multilinéaire a été utilisée pour modéliser la température d'ébullition (61 HAP), l'indice de rétention (209 HAP), la solubilité aqueuse (72HAP) et la température de fusion (77 HAP); la méthode des réseaux de neurones a été utilisée pour améliorer les résultats obtenus pour la température de fusion.

Les modèles établis, ont été validés selon les cinq principes avancés par *l'Organisation de Coopération et de Développement Economiques* (OCDE). Le domaine d'application des modèles est étudié à l'aide du diagramme de Williams pour détecter les composés aberrants en X et/ou en Y.

Des validations rigoureuses interne et externe ont été considérées pour juger la stabilité et la capacité prédictive de ces modèles afin de combler les lacunes dans les données physico-chimiques des HAP.

Mots-clés: Hydrocarbures aromatiques polycycliques- Propriétés physico-chimiques - Descripteurs moléculaires - Régression linéaire multiple - Réseaux de neurones artificiels.

Abstract

The work presented in this thesis aims to develop reliable, stable and predictive QSPR models for the prediction of physicochemical properties (boiling point, melting temperature, aqueous solubility, gas chromatographic retention index) of Polycyclic Aromatic Hydrocarbons compounds (PAHs).

The HyperChem molecular modeling software (V.6.03) was used to represent the molecules, then using the semi-empirical PM3 method the final geometries were obtained. Different molecular descriptors are calculated using the Dragon software (V.5.5).

QSPR models have been developed for the prediction of important physicochemical properties of various set of PAHs. Multiple linear regression (MLR) approach, and artificial neural networks (ANNs), lead to different quality models. Genetic algorithms (GA), have been associated to select the most important descriptors.

The hybrid genetic algorithm / multilinear regression approach was used to model the boiling temperature (61 HAP), the retention index (209 HAP), the aqueous solubility (72HAP) and the melting temperature (77HAP), the Neural network method was used to improve the results obtained for the melting temperature.

The established models have been validated according to the five principles put forward by the Organization for Economic Cooperation and Development (OECD). The field of application of the models is studied using the Williams diagram to detect aberrant compounds in X and / or in Y.

Rigorous internal and external validations were used to judge the stability and predictive capacity of these models in order to fill gaps in physico-chemical data on PAHs.

Keywords: Polycyclic aromatic hydrocarbons - physicochemical properties - molecular descriptors - multiple linear regression - artificial neural networks.

الملخص

يهدف العمل المقدم في هذه الرسالة إلى تطوير نماذج QSPR موثوقة ومستقرة وتنبؤية للتنبؤ بالخصائص الفيزيائية والكيميائية (نقطة الغليان، درجة حرارة الانصهار، القابلية للذوبان في الماء، مؤشر الاحتفاظ الكروماتوغرافي للغازية للهيدروكربونات .العطرية متعددة الحلقات (HAP).

تم استخدام برنامج النمذجة الجزيئية HyperChem (V.6.03) لتمثيل الجزيئات، ثم باستخدام طريقة PM3 شبه التجريبية تم الحصول على هندستها النهائية .وتم حساب الواصفات الجزيئية المختلفة باستخدام برنامج Dragon (V.5.5)

وقد وضعت نماذج QSPR للتنبؤ الخصائص الفيزيائية - الكيميائية هامة لمجموعات مختلفة من HAP إن الانحدار الخطي المتعدد (RLM)، والشبكات العصبية الاصطناعية (RNA)، تؤدي إلى نماذج ذو جودة مختلفة . الخوارزميات الجينية (AG)، وقد تم استعمالها لتحديد أهم الواصفات.

تم استخدام المقاربة الهجينة الخوارزم الوراثي / الانحدارالخطي المتعدد لنمذجة درجة حرارة الغليان (61 HAP)، ومؤشر الاحتفاظ (HAP 209)، والذوبان المائي (72HAP)ودرجة حرارة الانصهار (HAP 77) ، و تم استخدام طريقة الشبكة العصبية لتحسين النتائج التي تم الحصول عليها لدرجة حرارة الانصهار .

تم التحقق من صحة النماذج التي تم الحصول عليها وفقا للمبادئ الخمسة التي وضعتها منظمة التعاون والتنمية في الميدان الاقتصادي .تم دراسة مجال تطبيق النماذج باستخدام مخطط ويليامز للكشف عن المركبات الشاذة في X و / أو فيY.

واستخدمت عمليات التحقق الداخلية والخارجية الصارمة للحكم على استقرار هذه النماذج وقدرتها التنبؤية من أجل سد الثغرات في البيانات الفيزيائية - الكيميائية للهيدروكربونات العطرية متعددة الحلقات.

.

الكلمات الدالة

الهيدروكربونات العطرية متعددة الحلقات - الخصائص الفيزيائية والكيميائية - الو اصفلت الجزيئية - الانحدار الخطي المتعدد- الشبكات العصبية الاصطناعية.

SOMMAİRE

SOMMAİRE

SOMMAİRE

SOMMAİRE

LİSTE DES FİGURES

LİSTE DES TABLEAUX

SYMBOLES ET ABREVIATIONS

AG: Algorithme génétique

CAS : Chemical Abstracts Service

EQM: Ecart quadratique moyen.

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

EQMP Ecart quadratique moyen de prédiction.

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

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

ei_std : Résidu standardisé.

F : Statistique de Fisher.

FIT: Fonction de KUBINYI.

FIV: Facteur d'inflation de la variance.

GA: Algorithme génétique (Genetic Algorithm).

HF : Hartree -Fock

hii : Eléments diagonaux de la matrice chapeau.

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

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

MM+: Mécanique Moléculaire (+).

MM2 Mécanique moléculaire 2

n: Dimension de la population (échantillon).

n-p : Nombre de degrés de liberté.

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

PM3 : Parametrization Method 3.

PRESS : Somme des carrés des erreurs de prédiction.

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

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

QSAR: Relations Quantitatives Structure/ Activité.

QSPR : Relations Quantitatives Structure/ Propriété.

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

R^2 : Coefficient de détermination.

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

RLM (MLR): Régression linéaire multiple.

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

RNA: Réseaux de neurones artificiels.

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

S : Erreur standard.

SCE : Somme des carrés des écarts.

SCT : Somme des carrés totale.

t : t de Student.

t_i : Résidu studentisé externe.

y_i: Valeur observée.

$\hat{y}_i$: Valeur estimée.

$\hat{y}_{(i)}$: Valeur prédite.

İNTRODUCTİON GÉNÉRALE

İNTRODUCTİON GÉNÉRALE

Pendant longtemps, les hydrocarbures aromatiques polycycliques (HAP) ont fait l'objet d'une grande attention de la part de la communauté scientifique en raison de leur impact sur la santé publique et l'environnement. Certains de ces composés tels que les benzo[a] anthracène, chrysène, dibenzo[a,h]anthracene et benzo[a]pyrène sont mutagènes et cancérigènes (IARAC,1987 ; Jacob, 1996). Habituellement, les HAP sont introduits dans l'environnement suite aux activités anthropiques, qui ont augmenté considérablement ces 50 dernières années. Les HAP ont été détectés dans l'atmosphère, l'eau, les sols, les sédiments et les aliments (Ariese *et al*., 1993; Arfsten *et al* ., 1994; Faber et Heijman, 1996; Tolosa *et al*., 1996; Salau *et al*., 1997; Boehm *et al*., 1998; Franz et Einsenreich, 1998) . Il est bien établi que le devenir des HAP dans l'environnement est principalement contrôlé par leurs propriétés physicochimiques. Selon la solubilité dans l'eau, la volatilité, ou la lipophilie présentées, leurs distributions dans les systèmes aquatiques, l'atmosphère et les sols peuvent être significatives. D'autre part, ces composés qui sont aussi très volatils (une pression de vapeur relativement basse) résistent aux réactions chimiques. En conséquence, ils sont persistants dans l'environnement et ont tendance à s'accumuler dans les sols et les sédiments ; en outre, ils sont également fortement dispersés dans l'atmosphère.

Par conséquent, l'évaluation de leur impact sur l'environnement requiert des données fiables de leurs propriétés physicochimiques. Malheureusement, très souvent, ces informations ne sont pas disponibles dans la littérature, et fréquemment les données existantes comportent des inexactitudes. Ce qui est principalement lié aux procédures expérimentales difficiles (préparation, manipulation et analyse des solutions) mises en œuvre pour leur obtention.

Avec l'avènement du calcul peu coûteux et rapide, il y a une croissance remarquable de l'intérêt pour les relations quantitatives structure-propriété (QSPR), qui utilisent des méthodes d'analyses multidimensionnelles afin de modéliser des propriétés pertinentes en fonction des paramètres de la structure moléculaire (appelés descripteurs).

Un grand nombre de descripteurs ont été proposés dans la littérature. La nature des descripteurs couramment utilisés (structurels, topologiques, électroniques et géométriques) et le degré de codification des caractéristiques structurales moléculaires liées à certaines propriétés physiques spécifiques sont au cœur de n'importe quelle étude QSPR. On trouve dans la littérature un nombre élevé d'études QSPR, dont plusieurs sont consacrées aux HAP.Warne *et al*. (1990) ont utilisé des descripteurs moléculaires et des propriétés physico-

chimiques pour modéliser la solubilité et le coefficient de partage n-octanol/ eau. Gerstl (1990) et Karickhoff (1981) ont utilisé la solubilité dans l'eau pour estimer le coefficient de partage octanol-carbone organique du sol (Koc) pour les HAP. Les mêmes auteurs ont également corrélé le Koc et le coefficient de partage n-octanol / eau. Hong *et al*. (1996) ont prédit le coefficient de sorption dans le sol à l'aide du temps de rétention en chromatographie liquide haute performance à phase inverse (CLHP-PI). Govers *et al*. (1984) de même que Sabljić *et al*. (1995) ont utilisé l'indice topologique pour la prédiction du Koc. Une bonne revue sur le sujet est publiée par Gawlik *et al*. (1997). Quelques modèles pour l'estimation du Koc des HAP ont été établis en utilisant simultanément plus d'un descripteur. Certaines études relient la constante de la loi de Henry (K_H), ou le coefficient de partage air/eau (Kaw) aux descripteurs de la structure moléculaire. Bamford *et al*. (1999) ont mesuré la constante de la loi de Henry à différentes températures et relié log K_H avec le volume molaire. De Maagd *et al*. (1998) ont également mesuré K_H en plus de la solubilité aqueuse et Kow. Leurs résultats ont été reliés au volume molaire.

Le manuscrit de cette thèse est articulé sur trois grandes parties :

- Dans la première partie nous présentons :
  - ✓ Un aperçu général sur les hydrocarbures aromatiques polycycliques (HAP),
  - ✓ Un bref historique sur les QSAR,
  - ✓ Les différentes bases théoriques et les outils d'analyse des données statistiques nécessaires à la mise en œuvre des modèles QSAR.
- Dans la deuxième partie nous présentons :
  - ✓ Les différentes propriétés étudiées.
  - ✓ Les bases des données (origine des données).
  - ✓ La méthodologie utilisée lors de cette étude.
- Dans la troisième partie nous présentons et nous discutons les résultats obtenus pour:
  - ✓ *Tfus*, *Teb*, *IR*, *Sw*, et log *Koc* des 16 HAP prioritaires,
  - ✓ *Tfus*, *Teb*, *IR* et *Sw* pour de large gammes diverses de HAP.

  Nous terminerons par une conclusion générale.

Références bibliographiques

Arfsten D P, Schaeffer D J, Mulveny D C, The effects of near ultraviolet radiation on the toxic effects of polycyclic aromatic hydrocarbons in animals and plants: a review. Ecotoxicol. Environ. Safety. 1996. 33, 1.

Ariese F, Kok S J, Verkaik M, Gooijer C, Velthorst N H, Hofstraat J W, Synchronous fluorescence spectrometry of fish bile a rapid screening method for the biomonitoring of PAH exposure. Aquat. Toxicol. 1993. 26, 273.

Bamford H A, Poster D L, Baker J E, Temperature dependence of Henry's law constants of thirteen polycyclic aromatic hydrocarbons between 4°C and 31°C. Environ. Toxicol. Chem. 1999.18, 1905.

Boehm P D, Page D S, Gilfillan E S, Bence A E, Burns W A, Mankiewicz P J, Study of the fates and effects of the Exxon Valdez oil spill on benthic sediments in two bays in Prince William Sound, Alaska. 1. Study design, chemistry, and source fingerprinting. Environ. Sci. Technol. 1998. 32, 567.

De Maagd P G J, TenHulscher D T E M, VandenHeuvel H, Opperhuizen A, Sijm D T H M., Physicochemical properties of polycyclic aromatic hydrocarbons: Aqueous solubilities, n-octanol/water partition coefficients, and Henry's Law constants. Environ. Toxicol. Chem. 1998. 17, 253.

Faber J H, Heijmans J S M, Polycyclic aromatic hydrocarbons in soil detritivores. In: VanStraalen, N.M., Krivolutskii, D.A.(Eds.),Bioindicator Systems for Soil Pollution. Kluwer Academic Publishers, Dordrecht, MA. 1996, 31.

Franz T P, Eisenreich S, Snow scavenging of polychlorinated biphenyls and polycyclic aromatic hydrocarbons in Minnesota. J. Environ. Sci. Technol. 1998. 32, 1771.

Gawlik B M , Sotiriou N , Feicht E A, Schulte-Hostede S, Kettrup A., Alternatives for the determination of the soil adsorption coefficient, Koc, of non-ionic organic compounds a review. Chemosphere. 1997. 34, 2525.

Gerstl Z, Estimation of organic chemical sorption by soils. J. Contam. Hydrol. 1990. 6, 357.

Govers H, Ruepert C, Aiking H, Quantitative structure / activity relationships for polycyclic aromatic hydrocarbons: correlation between molecular connectivity, physico-chemical properties, bioconcentration and toxicity in daphinia pulex. Chemosphere. 1984.13, 227.

Hong H, Wang L S, Han S K, Zou G W, Prediction of adsorption coefficients (Koc) for aromatic compounds by HPLC retention factors (k,). Chemosphere. 1996. 32, 343.

IARC, Monographs on the Evaluation of the Carcinogenic Risk of Chemical to Humans, vol.32, Suppl. 7. International Agency for Research on Cancer, Lyon. 1987.

Jacob J, The significance of polycyclic aromatic hydrocarbons as environmental carcinogens. Pure. Appl. Chem. 1996. 68, 301.

Karickhoff S W, Semi empirical estimation of sorption of hydrophobic pollutants on natural sediments and soils. Chemosphere. 1981. 10, 833.

Sabljić A, Gutsen H, Verhaar H, Hermens J, QSAR modeling of soil sorption-improvements and systematics of logKoc vs log Kow correlations. Chemosphere. 1995. 31, 4489.

Salau J S I, Tauler R, Bayona J M, Tolosa I, Input characterization of sedimentary organic contaminants and molecular markers in the Northwestern Mediterranean Sea by exploratory data analysis. Environ. Sci. Technol. 1997. 31, 3482.

Tolosa I, Bayona J M, Albaiges J, Aliphatic and polycyclic aromatic hydrocarbons and sulfur/oxygen derivatives in northwestern Mediterranean sediments: spatial and temporal variability,fluxes, and budgets.Environ. Sci. Technol. 1996. 30, 2495.

Warne M St J, Connell D W, Hawker D W, Schuurmann G, Prediction of aqueous solubility and the octanol-water partition coefficient for lipophilic organic compounds using molecular descriptors and physicochemical properties. Chemosphere. 1990. 21, 877.

PARTIE I

SYNTHÈSE BİBLİOGRAPHİQUE

I. Les Hydrocarbures Aromatiques Polycycliques (HAP).

II. Évolution des QSAR/QSPR.

III. Bases théoriques.

IV. Relations (QSAR/QSPR) et méthodes statistiques.

I- Les Hydrocarbures Aromatiques Polycycliques (HAP) (Zedeck, 1980)

I.1.Définition des HAP :

Les HAP font partie de la famille des composés organiques hydrophobes (COH) qui comprend de nombreux autres types de substances (composés des résidus de pétrole et de fuel, composés chlorés dont les biphényles polychlorés, PCB, certains pesticides…). Les hydrocarbures aromatiques polycycliques sont très répandus dans l'environnement. Depuis quelques années, ils ont particulièrement attiré l'attention des scientifiques, et ce principalement à cause de leurs propriétés cancérigènes et mutagènes (Zedeck, 1980), reconnues chez certains animaux et fortement soupçonnées chez l'homme. Par conséquent, les HAP sont de plus en plus étudiés, à la fois chez les organismes vivants et dans les différents compartiments de l'environnement (atmosphère, sols, océans, rivières, etc…). Ces composés organiques peuvent notamment être utilisés comme traceurs du transport atmosphérique de contaminants d'origine anthropique. En effet, leur relative stabilité dans l'environnement et leur composition variable en fonction des sources et/ou de la distance par rapport à ces dernières font de ces molécules des marqueurs privilégiés (Halsall *et al.*, 1997). Contrairement aux composés organochlorés, les HAP sont produits en grandes quantités par des processus naturels, ce qui fait que les échantillons prélevés dans l'environnement peuvent contenir à la fois des HAP d'origines anthropique et naturelle. Toutefois, les émissions de HAP par les activités humaines ont crû dramatiquement durant le siècle dernier, à cause de l'augmentation de la combustion des énergies fossiles. Les HAP d'origine anthropique sont donc devenus largement prédominants dans les régions fortement industrialisées et urbanisées et constituent un problème majeur. Ces composés font ainsi partie des polluants à surveiller dans le cadre de la directive européenne nº 96/62/CE du 27 septembre 1996 (**https://aida.ineris.fr/consultation_document/1029**) et doivent être pris en compte dans les objectifs de qualité de l'air à atteindre.

I.2.Principales caractéristiques des HAP

I.2.1 Structure et propriétés physico-chimiques

Les HAP sont des composés organiques neutres, contenant uniquement des atomes de carbone et d'hydrogène et présentant au minimum deux cycles benzéniques fusionnés. A l'état pur et dans les conditions thermodynamiques standards (pression de 1 atm et température de 25°C), ils se présentent sous la forme de solides cristallins.

La figure 1 présente les 16 HAP qui ont été retenus comme polluants prioritaires par l'American Environmental Protection Agency (US-EPA) (**www.epa.gov**) du fait du risque qu'ils représentent pour l'environnement. Le tableau 1 (page suivante) résume les principales caractéristiques physico-chimiques de ces 16 HAP. De nombreuses études se concentrent sur ces 16 HAP, mais le nombre théorique de HAP susceptibles d'être rencontrés est supérieur à 1000 (Costes *et al*.,1997).

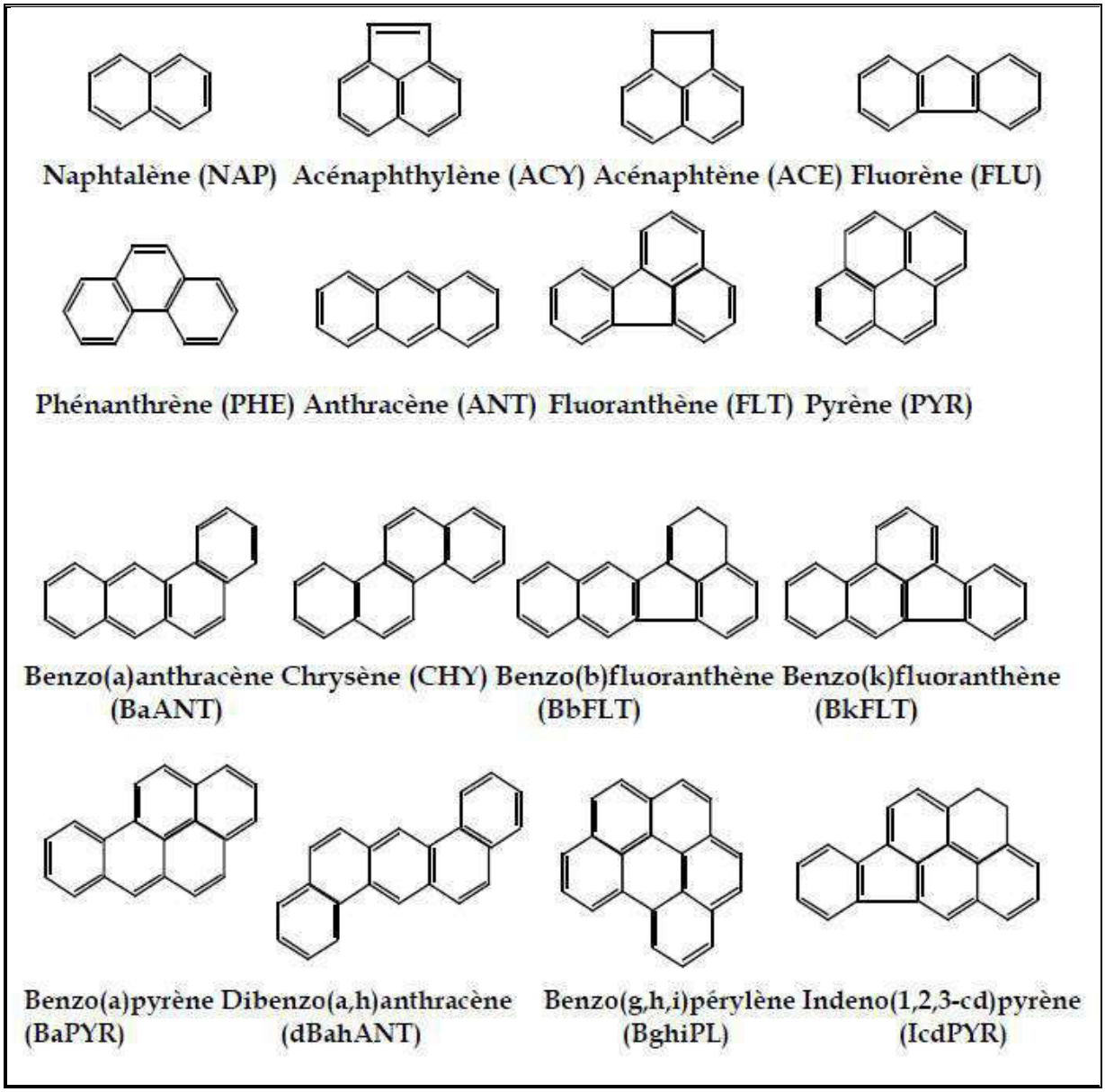


Figure. 1 Structure des 16 HAP retenus comme polluants majeurs par l'US-EPA

Les caractéristiques physico-chimiques des HAP sont extrêmement liées à leur structure particulière. La position des cycles de façon linéaire (anthracène) ou angulaire (phénanthrène) influence la stabilité des molécules, les HAP angulaires étant les plus stables

(Bouchez *et al*.,1996, Kanaly et Harayama, 2000). De la même façon la présence de cycles à 5 carbones parmi des cycles benzéniques augmenterait la stabilité des molécules vis-à-vis des attaques de micro- organismes (Wammer et Peter, 2005).

La présence de ces cycles aromatiques leur confère également la capacité à absorber un rayonnement ultra-violet (UV) et une partie du rayonnement visible (Lampi, 2005) ou encore à ré-émettre un rayonnement de fluorescence en réponse à une excitation par un rayonnement UV. Ces dernières propriétés sont utilisées pour la détection de ces molécules.

Tableau.1 Propriétés physico-chimiques des 16 HAP retenus comme polluants majeurs par l'United States Environmental Protection Agency (Schwarzenbach *et al*., 2003; ATSDR, 1995; Ferreira, 2001; Shiu et Mackay, 1997).

Composé	Symbole	Masse molaire ($g.mol^{-1}$)	K_H ($m^3.Pa.mol^{-1}$) à 25°C	P_{SAT} (Pa) à 25°C	K_{ow}	Solubilité ($mg.L^{-1}$)
Naphtalène	NAP	128.19	43.01	33	$2.34\ 10^{3}$	31
Acénaphtylène	ACY	154.21	12.17	1.35	$8.32\ 10^{3}$	3.8
Acénaphtène	ACE	152.20	8.4	4.14	$1\ 10^{4}$	16.1
Fluorène	FLU	166.20	7.87	$4.5\ 10^{-1}$	$1.51\ 10^{4}$	1.9
Phénanthrène	PHE	178.23	3.61	$5.7\ 10^{-2}$	$1.7\ 10^{3}$	4.57
Anthracène	ANT	178.23	3.96	$5.2\ 10^{-2}$	$3.5\ 10^{4}$	$4.5\ 10^{-2}$
Fluoranthène	FTH	202.26	1.037	$5.6\ 10^{-3}$	$1.7\ 10^{5}$	$2.6\ 10^{-1}$
Pyrène	PYR	202.26	$9.2\ 10^{-1}$	$4.1\ 10^{-3}$	$1.5\ 10^{5}$	$1.3\ 10^{-1}$
Benz(a)anthracène	BaA	228.29	$5.8\ 10^{-1}$	$2.3\ 10^{-4}$	$8.1\ 10^{5}$	$1.1\ 10^{-2}$
Chrysène	CHR	128.29	5.86	$4.8\ 10^{-5}$	$4.5\ 10^{5}$	$3.3\ 10^{-3}$
Benzo(b)fluoranthène	BbF	252.31	(-)	(-)	$6.3\ 10^{5}$	$1.5\ 10^{-3}$
Benzo(k)fluoranthène	BkF	252.31	$1.6\ 10^{-2}$	$4.1\ 10^{-6}$	$1\ 10^{6}$	$8\ 10^{-4}$
Benzo(a)pyrène	BaP	252.31	$4.6\ 10^{-2}$	$3.2\ 10^{-6}$	$1\ 10^{6}$	$3.8\ 10^{-3}$
Dibenz(ah)anthracène	DahA	278.35	$1.7\ 10^{-4}$	$8.1\ 10^{-8}$	$3.2\ 10^{6}$	$6\ 10^{-4}$
Benzo(ghi)pérylène	BghiP	268.35	$7.5\ 10^{-2}$	$1.1\ 10^{-12}$	(-)	$2.6\ 10^{-4}$
Indéno(1,2,3-cd)pyrène	IcdP	276.33	$3.07\ 10^{-5}$	(-)	(-)	(-)

De la forte influence des cycles aromatiques sur les propriétés physico-chimiques découlent des différences très marquées entre les HAP à deux ou trois cycles et ceux à quatre cycles et plus. On distingue ainsi généralement deux classes de HAP (Bouchez *et al*.,1996), ceux à faible poids moléculaire ou "légers" (deux et trois cycles, donc inférieurs à

180 g mol^{-1}) et ceux à poids moléculaire élevé ou "lourds" (quatre cycles et plus et supérieurs à 200 g mol^{-1}).

Le naphtalène qui est le plus léger a une solubilité notable de 32 mg L^{-1}, mais la solubilité décroît fortement avec le nombre de cycles jusqu'à 10^{-4} mg L^{-1}. Il en va de même pour leur volatilité puisque la constante de Henry K_H qui traduit la répartition du composé entre la phase liquide et la phase gazeuse va de 43 Pa m^3 mol^{-1}pour le naphtalène à près de dix-mille de fois moins ($3{,}07.10^{-3}$ Pa m^3 mol^{-1}) pour l'indeno-pyrène. L'hydrophobie des HAP est caractérisée par une constante de partage n-octanol-eau (K_{OW}) élevée. Cette constante traduit la répartition d'un composé entre une phase lipophile (octan-1-ol) et la phase hydrophile (l'eau). La structure en cycles aromatiques des HAP influence directement leur devenir dans l'environnement puisqu'elle est responsable de leur faible solubilité et forte hydrophobie et par conséquence de leur forte adsorption sur les phases solides du sol.

I.2.2 Toxicité des HAP

Actuellement, les effets toxicologiques de tous les HAP sont imparfaitement connus. Toutefois, les données expérimentales disponibles chez l'animal ont montré que certains HAP pouvaient induire spécifiquement de nombreux effets sur la santé, des effets systémiques (hépathiques, hématologiques et immunologiques), et/ou des effets sur la reproduction ainsi que des effets génotoxiques et cancérigènes (www.ineris.fr).

La toxicité des HAP peut s'exprimer selon trois mécanismes principaux :

- La toxicité narcotique qui affecte la fluidité et les fonctions des cellules membranaires sans la présence d'un récepteur spécifique. Cette toxicité semble être en partie liée à l'hydrophobie des HAP (Sverdrup *et al.,* 2002) qui leur permettrait d'interagir avec les membranes lipidiques (qui sont lipophiles et donc hydrophobes). Cette toxicité est dite "directe " (Boese *et al.,* 1999) par opposition entre autres à la toxicité photo-induite ou photo-toxicité.

- La photo-toxicité correspond à une toxicité indirecte des molécules de HAP sous l'effet d'un rayonnement UV. Une molécule de HAP excitée par un rayonnement UV peut transmettre son énergie à une molécule d'oxygène et former ainsi un radical oxygène. Ce radical peut alors s'attaquer aux molécules biologiques et en particulier aux membranes cellulaires par peroxydation des lipides (Mc Donald et Chapaman., 2002). Cette toxicité est fonction à la fois de la quantité de HAP présents en surface de l'organisme et de la dose de

radiations UV. Bien que les effets nocifs de cette photo-toxicité soient démontrés en laboratoire, il semble cependant que ces effets ne sont pas significatifs dans l'environnement, les espèces animales concernées ayant probablement développé des moyens de protection (Mc Donald et Chapaman., 2002).

- La génotoxicité correspond à l'apparition d'altérations dans la structure de l'ADN. Le principal risque que présentent ces composés sur la santé, est leur capacité à induire le développement de cancer dans les organismes exposés. L'induction du cancer chez les mammifères par les HAP passe par la participation d'un groupe d'enzymes capables de transformer les composés xénobiotiques en produits solubles dans l'eau. Ces enzymes sont des mono oxygénases qui appartiennent au groupe cytochrome P450 (voir figure 2). Les HAP sont transformés en HAP diolépoxydes qui sont alors particulièrement réactifs avec l'ADN, l'ARN et les protéines cellulaires, créant ainsi de nombreuses mutations irréversibles et induisant la formation de tumeurs. Ce système enzymatique est stimulé dans un organisme par exposition aux composés lipophiles persistants.

Les expositions répétées à ces composés induisent de grandes quantités d'enzymes. La capacité d'induction de ces enzymes dépend de chaque organisme. Les mammifères par exemple, ont une grande capacité d'adaptation à ces agents mutagènes et une exposition chronique aux HAP provoque à terme la production d'anticorps dégradant les composés lipophiles persistants. Par contre, les poissons ont une capacité limitée de dégradation (Sutherland *et al*., 1995).

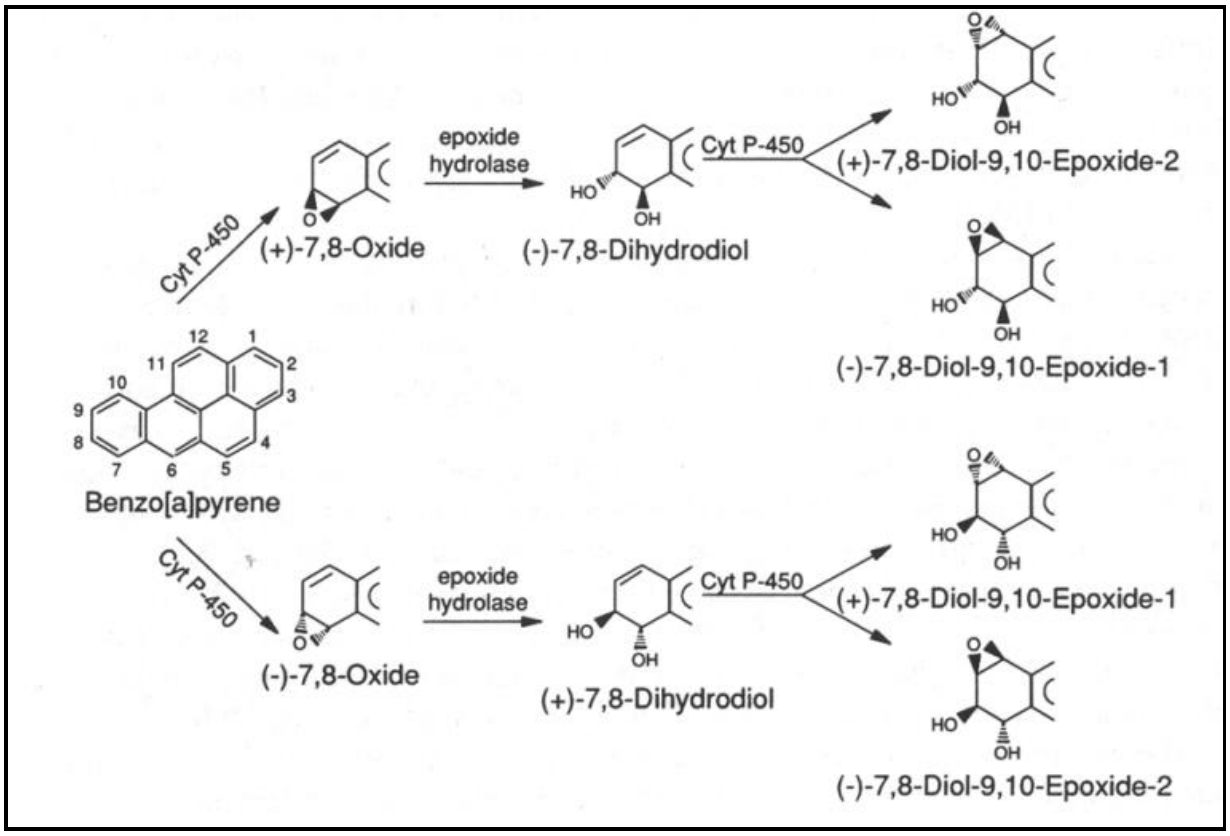


Figure. 2 Schéma d'activation métabolique du Benzo[a]pyrène chez les mammifères d'après (Sutherland *et al.*, 1995).

Une dernière toxicité indirecte est la toxicité potentielle des produits de dégradation des molécules de HAP. Que la dégradation soit physique (comme la photo-modification) ou biologique, elle est susceptible de transformer le HAP parent en un produit beaucoup plus toxique comme les HAP oxygénés (Lampi, 2005; Lundstedt, 2003).

Le benzo[a]pyrène est potentiellement le plus cancérigène. D'autres HAP sont également reconnus comme étant fortement génotoxiques et cancérigènes, comme le fluoranthène, le benzo[b]fluoranthène, le benzo[k]fluoranthène, le chrysène, le benzo[g,h,i]pérylène et l'indéno[1,2,3-cd]pyrène (Feix et Wiart, 1995).

Le naphtalène est peu toxique car, en général, l'ingestion d'une dose correspondant à une boule d'antimite (4 grammes) n'entraîne qu'une irritation des muqueuses et éventuellement quelques troubles neurobiologiques réversibles. Les HAP sont absorbés par l'homme (Boufetta *et al.*, 1997; Wornat *et al.*, 2001) par :

- Les voies respiratoires via l'inhalation de particules atmosphériques contaminées ou de fumées de cigarettes. Le taux d'absorption par les poumons dépend du type de HAP, de la taille et de la composition des particules sur lesquelles ils sont adsorbés.
- Le système digestif via l'ingestion de produits alimentaires contaminés notamment les produits grillés ou fumés. Le poisson fumé peut contenir jusqu'à 80µg de HAP par kilogramme.
- La peau.

Les personnes travaillant dans l'industrie du bois, dans des locaux confinés contenant des fourneaux ou utilisant du goudron ou de l'asphalte présentent un risque accru de cancer des poumons, de l'œsophage et de la peau (Boufetta *et al*., 1997; Wornat *et al*., 2001; Partanen et Boufetta, 1994). Les tests de toxicité sont souvent effectués sur des animaux ou des micro-organismes ce qui est un problème pour extrapoler et évaluer la toxicité des HAP envers les humains. Des tests d'écotoxicologie permettent de caractériser un niveau de toxicité des polluants présents dans les sols. Pour cela, il est possible de faire des tests sur les plantes ou sur les organismes du sol (bactéries, vers de terre) mais ils sont coûteux, longs et les microorganismes doivent être au préalable adaptés aux polluants.

Une autre méthode consiste à réaliser des bio-essais sur les lixiviats du sol. Différents tests sont utilisés en laboratoire, et l'un des plus sensibles et des plus rapides à mettre en oeuvre pour les HAP est le test Microtox (Renoux *et al*., 1999; Bipso *et al*.,1999) qui repose sur l'extinction de luminescence de la bactérie Vibrio fischeri. Ce test permet de déterminer la concentration de polluant nécessaire pour diminuer de moitié la luminescence initiale de la bactérie. La toxicité d'une molécule est différente lorsqu'elle est seule ou en mélange. Par exemple, Renoux *et al*. (Renoux *et al*., 1999) reportent une toxicité plus élevée du fluorène lorsqu'il est en mélange avec du phénanthrène et du p-crésol ; ceci indique des effets de synergie entre les molécules

I.3.Origines des HAP

Les différentes sources de HAP ont pour point commun de produire des mélanges de composés. Il est établi que la majeure partie d'entr'eux est d'origine pyrolytique (Mc Elroy *et al*., 1981). Ces HAP proviennent de la combustion incomplète de la matière organique à haute température (Hase et Hites, 1978). Cette origine peut être naturelle (feux de forêts; (Youngblood et Bammer, 1975; Freeman et Cattel, 1990), éruptions volcaniques(Greiner *et al*., 1977) ou plus généralement anthropique (activités industrielles), automobile, incinération

des déchets, chauffage domestique). La fumée de cigarette (Shmeltz et Hoffman, 1976) ainsi que certains procédés de préparation et de cuisson des aliments (fumage, grillade, chauffage de l'huile de cuisine) constituent également une source pyrolytique non négligeable d'exposition de l'homme aux HAP.

Les HAP peuvent également être d'origine pétrogénique, issus de la maturation lente (plusieurs millions d'années) de la matière organique lors de la diagénèse et de la catagénèse dans le milieu sédimentaire profond (plusieurs milliers de mètres de profondeur à des températures de 100 à 150°C et des pressions de 300 à 1500 bars). Ces mécanismes produisent des mélanges complexes appelés pétroles dont la part massique des composés aromatiques atteint 20 à 45%, les HAP (HAP alkylés et soufrés compris) représentant environ 65% des aromatiques(Tissot et Welte, 1978). Ces composés sont introduits dans le milieu marin soit lors de déversements pétroliers dus à l'activité humaine (naufrages, dégazages, etc…), soit par des fuites de réservoirs naturels à travers l'écorce terrestre (Neff *et al*., 1976). Enfin, les HAP peuvent dériver de la modification chimique lors de la diagénèse précoce de précurseurs naturels (origine diagénétique), tels que les pigments, les stéroïdes, les quinones, accumulés dans les dépôts sédimentaires (Neff *et al*., 1976; Aizen shtat, 1973); cette dernière source est toutefois minoritaire par rapport aux deux autres excepté en certains endroits très localisés (Budzinski *et al*., 1997).

I.4. Distribution Des HAP

- Eau

Par leur nature hydrophobe, les HAP contenus dans les milieux aquatiques sont principalement liés aux particules organiques et minérales en suspension dans l'eau (Herbes, 1977). Le dépôt de ces particules transfère la contamination en HAP vers les sédiments. Les HAP peuvent s'accumuler dans les organismes marins selon la contamination de l'eau, des sédiments ou de la chaîne alimentaire, mais aussi selon la physiologie de l'organisme (Meador *et al*., 1995). La taille de l'organisme, le taux d'ingestion, la perméabilité membranaire, le temps de résidence et l'osmorégulation affectent la contamination des organismes marins par les HAP (Juhasz et Naidu, 2000). La température, la teneur en oxygène, le pH et la salinité, quant à eux, influencent la biodisponibilité des HAP dans le milieu aquatique.

- Sédiment

Bien que la majorité des HAP soit émise dans l'atmosphère, le sol et les sédiments constituent le principal point de fuite environnementale de ces polluants (Wilcke,2000). Les HAP, molécules hydrophobes, se lient aux particules organiques des sédiments où ils s'accumulent (Meador *et al*., 1995). Les sources des HAP sédimentaires marins sont atmosphériques, liées à l'industrie pétrolière en haute mer, au transport fluvial ou maritime, aux contaminations issues du littoral (Shiaris et Jambard-Sweet, 1986; Yunker *et al*.,1993; Baumard *et al*., 1998; Johnson *et al*., 1985; Trapido, 1999). La concentration des HAP dans les sédiments est variable (du $\mu g.Kg^{-1}$ au $g.Kg^{-1}$). Par exemple, Shiaris M. P.*et al.*(1986) ont mesuré 718 $mg.Kg^{-1}$de HAP dans les sédiments se trouvant à proximité du port de Boston. Cependant, la contamination en HAP des sédiments peut s'avérer être plus importante, puisque Johnson A. C. *et al.* (1985) ont montré la présence de plus de 100 $mg.g^{-1}$ de HAP dans les sédiments d'estuaires urbanisés.

- Sol

C'est au cours du XIX^{ème} siècle que les concentrations en HAP dans les sols industrialisés ont commencé à augmenter avec un maximum vers 1950-1960 (Juhasz et Naidu, 2000), constituant une source importante de pollution de ce compartiment de l'écosystème. Trapido M. *et al.* (1999), Jones K. *et al.* (1989) ainsi que Motelay-Massei A. *et al* (2004) estiment qu'un sol sans contamination a des teneurs en HAP de l'ordre de 0,1 à 55 $mg.Kg^{-1}$. Wild S. *et al.* (1995) estiment que 90 % des HAP émis dans l'environnement sont stockés dans les sols, sans comptabiliser ceux des sites industriels. En effet, la contamination des sols à proximité des sites industriels est importante et elle diminue de façon exponentielle en s'éloignant de la source émettrice cela dépend du mode de transport et de la concentration en polluants (Crepineau *et al*., 2003). Les teneurs en HAP dans les sols dépendent du type d'industrie. Juhasz A. L. *et al.* (2000) montrent que le stockage du bois représente la plus importante source de contamination en HAP apportant aux sols 18700 $mg.Kg^{-1}$, contre plus de 7000 $mg.Kg^{-1}$pour les usines à gaz. Les sols agricoles peuvent également être contaminés par les HAP via l'épandage des boues de stations d'épuration (Wild *et al*., 1992).

I.4.1. Devenir des HAP dans l'environnement

Afin d'avoir une meilleure compréhension de l'impact sur l'environnement de polluants organiques tels que les HAP, tant sur la qualité des sols que sur son effet sur les

organismes telluriques, on va rappeler les devenirs possibles des HAP et la transformation voire la dégradation de ces polluants via des processus physico-chimiques ou biologiques.

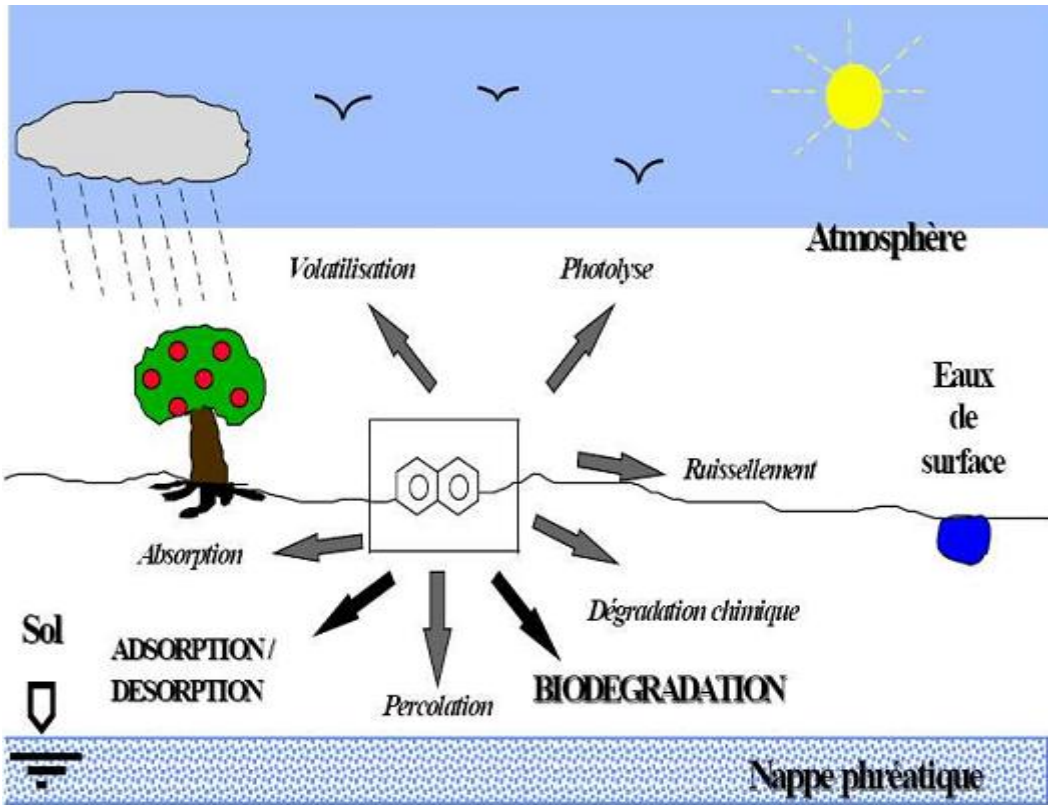


Figure. 3 Devenir des HAP dans les sols (d'après Mahjoub (1999)).

Le transport des HAP dans les sols fait principalement intervenir les phénomènes suivants : la solubilisation, la percolation, l'adsorption, la volatilisation.

Les phénomènes de transformation des HAP peuvent se produire essentiellement par hydrolyse, photolyse et biodégradation.

I.4.2. **Conséquences des propriétés physico-chimiques sur le devenir des HAP**

- ***Sorption***

La persistance des HAP dans le sol dépend de leur capacité à se sorber et à se désorber aux surfaces ou aux interfaces du sol (Huang *et al*., 2003) au niveau des argiles et de la matière organique (MO). Celle-ci représente le sorbant principal des HAP dans les sols du fait des nombreuses liaisons hydrophobes qui la constituent (Luthy *et al*., 1997; Wilcke, 2000, Hwang *et al*., 2003). La sorption des HAP dans le sol se déroule avec une première phase rapide qui tend vers un équilibre relatif (adsorption sur des sites de surface de

macromolécules organiques) et qui est le plus souvent réversible, suivie d'une période de sorption lente (diffusion progressive des HAP dans les sites internes) se déroulant sur plusieurs semaines voire des années (Rao, 1990; Hatzinger *et Al*exander, 1995; Pignatello, 1998; White *et al*., 1999).

Les études portant sur la sorption des HAP dans les sols considèrent qu'elle suit un modèle linéaire sauf dans le cas des sols riches en matière organique (MO). La non linéarité augmenterait ainsi avec la densité de la MO (Pignatello et Xing, 1996). Il semblerait que cet écart à la linéarité soit corrélé avec le rapport atomique O/C de la MO du sol. Le modèle de partage linéaire considère la MO des sols comme une phase de type gel et amorphe, et qu'il n'y a pas de limitation du nombre de sites offerts à la sorption lorsque la concentration en soluté augmente (Karichoff *et al*., 1979). La disponibilité des HAP dépend également de leur fixation sur des substances humiques : un poids moléculaire élevé réduit la disponibilité du polluant par stabilisation et formation de résidus liés alors qu'un faible poids moléculaire augmente la mobilité du HAP.

Les HAP se trouvent disponibles en faibles quantités dans la solution du sol (Weissenfelds *et al*., 1992) par leur adsorption sur les surfaces et la formation de phases non aqueuses. La présence de surfactants dans le milieu (issus de microorganismes ou de végétaux) facilite le transport des HAP de la phase solide à la phase aqueuse (Gao *et al*., 2006).

A long terme, les HAP séquestrés dans les particules du sol évoluent. Une partie des HAP s'avère alors récalcitrante à la biodégradation : les HAP sont fixés sur les fractions de granulométrie fine du sol (Amellal *et al*., 2001) inaccessibles aux microorganismes dégradants, ce qui limite leur désorption (Sun *et al*.,2003).

- ***Transport***

Même si la plupart des HAP se retrouvent dans les couches superficielles du sol, riches en MO, ils peuvent également se trouver dans les couches plus profondes après lessivage ou lixiviation [Wild *et al*., 1992, Wilcke, 2000]. Le transport des HAP vers les horizons profonds se produit par l'intermédiaire de la matière organique dissoute (MOD).

RÉFÉRENCES BİBLİOGRAPHİQUES

Aizen shtat Z, Perylene and its geochemical significance. Geochim. Cosmochim. Acta. 1973.37, 559.

Amellal N, Portal J-M, Berthelin J, Effect of soil structure on bioavailability of polycyclic aromatic hydrocarbons within aggregates of a contaminated soil. Appl. Geochem. 2001. 16, 1611.

ATSDR (Agency for Toxic Substances and Disease Registry), Toxicological profile for Polycyclic Aromatic Hydrocarbons. U.S. Department of Health and Human Services, Public Health Service, Atlanta, GA, USA. 1995, 454.

Baumard P, Budzinski H, Garrigues P, Polycyclic aromatic hydrocarbons in sediments and mussels of the western Mediterranean Sea. Environ. Toxicol. Chem. 1998. 15, 765.

Bispo A , Jourdain M J, Jauzein M, Toxicity and genotoxicity of industrial soils polluted by polycyclic aromatic hydrocarbons (PAHs), Organic Geochemistry. 1999. 30, 947.

Boese B L, Ozretich R J, Lamberson J O, Swartz R C, Cole Pelletier F A J, Jones J, Toxicity and phototoxicity of mixtures of highly lipophilic PAH compounds in marine sediment: Can the Sigma PAH model be extrapolated?. Arch. Environ. Con. Tox. 1999. 36, 270.

Boffeta P, Jourenkova N, Gustavsson P, Cancer risk from occupational and environmental exposure to polycyclic aromatic hydrocarbons. Cancer Cause. Control. 1997. 8, 444.

Bouchez M, Blanchet D, Haeseler F, Vandecasteele J P, Les hydrocarbures aromatiques polycycliques dans l'environnement. 1ère partie : Propriétés, origines, devenir. Rev. I. Fr. Petrol. 1996.51, 407.

Budzinski H, Jones I, Bellocq J, Pierard C, Garrigues P., Evaluation of sediment contamination by polycyclic aromatic hydrocarbons in the Gironde estuary. Mar. Chem. 1997.58, 85

Costes J M, Druelle V, Les hydrocarbures aromatiques polycycliques dans l'environnement : La réhabilitation des anciens sites industriels. Rev. Inst. Franç. Pétr. 1997 .52, 425.

Crepineau C, Rychen G, Feidt C, Le Roux Y, Lichtfouse E, Laurent F, Contamination of pastures by polycyclic aromatic hydrocarbons (PAHs) in the vicinity of a highway. J. Agr. Food. Chem. 2003. 51, 4841.

Feix I, Wiart J, Les micropolluants organiques dans les boues résiduaires des stations d'épuration urbaines. ADEME. 1995.

Ferreira M M C, Polycyclic aromatic hydrocarbons : a QSPR study. Chemosphere. 2001. 44, 125

Freeman D J, Cattell C R, Wood burning as a source of atmospheric polycyclic aromatic hydrocarbons. Environ. Sci. Technol. 1990.24, 1581

Gao Y, Ling W, Wong M H, Plant-accelerated dissipation of phenantrene and pyrene from water in the presence of a nonionic-surfactant. Chemosphere. 2006. 63, 1560.

Greiner A C, Spyckerelle C, Albrecht P, Ourisson G, Hydrocarbures aromatiques d'origine géologique. V. Dérivés mono- et di-aromatiques du hopane. J. Chem. Res., Miniprint. 1977. 3829.

Halsall C J, Coleman P J, Jones K C, Atmospheric deposition of polychlorinated dibenzo-p-dioxins / dibenzofurans (PCDD/Fs) and polycyclic aromatic hydrocarbons (PAHs) in two UK cities. Chemosphere. 1997.35, 1919.

Hase A, Hites R A, On the origin of polycyclic aromatic hydrocarbons in the aqueous environment. In Identification and analysis of organic pollutants in water (Keith LH ed), Ann Arbor Science, Ann Arbor, MI.1978.205.

Hatzinger P B, Alexander M, Effect of aging of chemicals in soil on their biodegradability and extractability. Environ. Sci. Tech. 1995. 29, 537.

Herbes S E, Partioning of polycyclic aromatic hydrocarbons between dissolved and particulates phases in natural waters. Water Ressources. 1977. 11, 493.

Huang W, Peng P, Yu Z, Fu J, Effect of organic matter heterogeneity on sorption and desorption of organic contaminants by soils and sediments. Appl. Geochem. 2003. 18, 955.

Hwang S, Ramirez N, Cutright T J, Ju L K, The role of soil properties in pyrene sorption and desorption. Water. Air. Soil. Poll. 2003. 143, 65.

Johnson A C, Larsen P F, Gadbois D F, Humason A W, Distribution of polycyclic aromatic hydrocarbons in the surficial sediments of Penobscot Bay (Maine, USA) in relation to possible sources and to other sites worldwide. Mar. Environ. Res. 1985.15, 1

Jones K C, Stratford J A, Waterhouse K S, Vogt N B, Organic contaminants in Welsh soil: polynuclear aromatic hydrocarbons. Environ. Sci. Technol. 1989. 23, 540.

Juhasz A L, Naidu R, Bioremediation of high molecular weight polycyclic aromatic hydrocarbons: a review of the microbial degradation of benzo[a]pyrene. Int. Biodeter. Biodegr. 2000. 45, 57.

Kanaly R A, Harayama S, Biodegradation of high-molecular-weight Polycyclic Aromatic Hydrocarbons by bacteria. J. Bacteriol. Mycol. 2000. 182, 2059.

Karickhoff S W, Brown D S, Scott T A, Sorption of hydrophobic pollutants on natural sediments. Water. Res. 1979.13, 241.

Lampi M A, Environmental photoinduced toxicity of Polycyclic Aromatic Hydrocarbons: Occurence and toxicity of photomodified PAHs and predictive modeling of photoinduced toxicity. PhD thesis, Waterloo, Ontario, Canada. 2005. 146

Lundstedt S, Analysis of PAHs and their transformation products in contaminated soil andremedial processes. PhD Thesis, Umea University, Sweden. 2003. 56.

Luthy R G, Aiken G R, Brusseau M L, Cunningham S D, Gschwend P M, Pignatello O J J, Reinhard M, Traina S J, Weber W J, Westall J C, Sequestration of hydrophobic organic contaminants by geosorbents. Environ. Sci. Tech. 1997. 31, 3341.

Mahjoub B, Comportement dans le sol de polluants aromatiques issus du goudron de houille. Thèse LAEPSI. Lyon : INSA de Lyon. 1999, 262.

McDonald B G, Chapman P M, PAH phototoxicity-an ecologically irrelevant phenomenon ?. Mar. Pollut. Bull. 2002. 44, 1321

McElroy A E, Farrington J W, Teal J M, Bioavailability of PAH in the aquatic environment. In Metabolism of polycyclic aromatic hydrocarbons in the aquatic environment (Varanasi U, ed), CRC Press, Boca Raton, FL.1989.1.

Meador J P, Stein J E, Reichert W L, Varanasi U, Bioaccumulation of polycyclic aromatic hydrocarbons by marine organisms. Rev. Environ. Contam.T. 1995. 143, 79

Motelay-Massei A, Ollivon D, Garban B, Teil M J, Blanchard M, Chevreuil M., Distribution and spatial trends of PAHs and BPCs in soils in the Seine River basin, France. Chemosphere. 2004. 55, 555.

Neff J M, Cox B A, Dixit D , Anderson J W, Accumulation and release of petroleum-derived aromatic hydrocarbons by four species of marine animals. Mar. Biol. 1976.38, 279

Partanen T, Boffetta P, Cancer risk in asphalt workers and roofers : review and meta-analysis of epidemiologie studies. Am. J. Ind. Med. 1994. 26, 721.

Pignatello J J, Soil organic matter as a nanoporous sorbent of organic pollutants. Adv. Colloid. Interfac. 1998.76/77, 445.

Pignatello J J, Xing B, Mechanisms of slow sorption of organic chemicals to natural particles. Critical review. Environ. Sci. Tech. 1996. 30, 1.

Rao P S C, Sorption of organic contaminants. Wat. Sci. Tech. 1990. 22, 1.

Renoux A Y, Millette D, Tyagy D, Samson R, Detoxification of fluorene, phenanthrene, carbozole and p-cresol in columns of aquifer sand as studied by the Microtox® assay, Wat. Res. 1999. 33, 2045.

Schmeltz I, Hoffmann D, Formation of polynuclear aromatic hydrocrabons from combustion of organic matter. In Carcinogenesis-A comprehensive survey. Vol. 1. Polynuclear Aromatic Hydrocarbons. Chemistry, Metabolism, and Carcinogenesis. Raven Press, New York. 1976. 225

Schwarzenbach R P, Gschwend P M, Imboden D M, Environmental Organic chemistry. John Wiley & Sons. 2003. 1313.

Shiaris M P, Jambard-Sweet D, Polycyclic aromatic hydrocarbons in surficial sediments of Boston Harbour, MA, USA. Mar. Pollut. Bull. 1986. 17, 469.

Shiu W Y, Mackay D, Henry' s Law Constants of selected aromatic hydrocarbons, alcohols and ketones. Journal of Chemical Engineering Data. 1997.42, 27.

Sun H, Taleda M, Ike M, Fujita M, Short and long term sorption/desorption of polycyclic aromatic hydrocarbons onto artificial solids: effects of particle and pore size and organic matters. Water. Res. 2003. 37, 2960.

Sutherland J B, Raffi F, Khan A A, Cerniglia C E, Mechanisms of Polycyclic Aromatic Hydrocarbon Degradation. In Microbial Transformation and degradation of toxic organic chemicals. Edited by Young, L.L and C.E. Cerniglia. Wiley-Liss. New York. 1995.

Sverdrup L E, Nielsen T, Krogh P H, Soil ecotoxicity of polycyclic aromatic hydrocarbons in relation to soil sorption, lipophibicity, and water solubility. Environmental Science and Technology. 2002. 36, 2429.

Tissot B P, Welte D H, Petroleum formation and occurrence. A new approach to oil and gas exploration. Springer-Verlag, Berlin Heidelberg, 1978.

Trapido M, Polycyclic aromatic hydrocarbons in Estonian soil: contamination and Profiles. Environ. Pollut. 1999. 105, 67

Wammer K H, Peters C A, Polycyclic Aromatic Hydrocarbon biodegradation rates: A structure based study. Environ. Sci. Technol. 2005. 39, 2571.

Weissenfelds W D, Klewer H J, Langhoff J, Adsorption of polycyclic aromatic hydrocarbons (PAHs) by soil particles: influence on biodegradation and biotoxicity. Appl. Microbiol. Biot. 1992. 36, 689.

White J C, Alexander M, Pignatello J J, Enhancing the bioavailability of organic compounds sequestred in soil and aquifer solids. Environ. Sci. Tech. 1999. 18, 182.

Wilcke W, Polycyclic aromatic hydrocarbons (PAHs) in soils - a review. J. Plant. Nutr. Soil. Sc. 2000. 163, 229.

Wilcke W, Polycyclic aromatic hydrocarbons (PAHs) in soils - a review. J. Plant. Nutr. Soil. Sc. 2000. 163, 229.

Wild S R, Berrow M L, McGrath S P, Jones K C, Polynuclear aromatic hydrocarbons in crops from long term field experiments amended with sewage sludge. Environ. Pollut. 1992. 76, 25.

Wild S R, Jones K, Polynuclear aromatic hydrocarbons in the United Kingdom environment: a prelimary source inventory and budget. Environ Pollu. 1995. 88, 91.

Wornat M J, Ledesma E B, Sandrowitz A K, Polycyclic Aromatic Hydrocarbons identified in soot extracts from domestics coal-burning stoves of Henan provinces, China. Appl. Environ. Microbiol. 2001. 62, 4174.

www.ineris.fr.

Youngblood W W, Blummer M, Polycyclic aromatic hydrocarbons in the environment: homologous series in soils and recent marine sediments. Geochim. Cosmochim. Acta. 1975.39 ,1303.

Yunker M B, MacDonald R W, Cretney W J, Fowler B R, MacLaughlin F A, Alkane, terpene and polycyclic aromatic hydrocarbon geochemistry of the Mackensie River and Mackensie Shelf: riverine contributions to Beaufort Sea coastal sediment. Geochimica et Cosmo. Chimica. Acta. 1993. 57, 3041.

Zedeck M S, Polycyclic aromatic hydrocarbons: a review. J. Environ. Pathol. Toxicol. 1980.3, 537.

ÉVOLUTİON DES QSAR/QSPR

I. Bref historique sur les QSAR (Gramatica, http://www. QSARworld.com)

La méthodologie QSAR repose sur l'hypothèse que la structure d'une molécule (c'est-à-dire ses propriétés géométriques, stériques et électroniques) doit contenir les caractéristiques responsables de ses propriétés physiques, chimiques et biologiques et sur la capacité de représenter un composé chimique par un, ou plusieurs, descripteur (s) numérique (s). Par les modèles QSAR, l'activité biologique (ou la propriété, la réactivité, etc...) d'un produit chimique nouveau ou non testé peut être déduite de la structure moléculaire de composés similaires dont les activités (propriétés, réactivités, etc…) ont déjà été évaluées. L'acronyme QS/XR [X= A (activité), P (propriétés), T (toxicologie)…] est utilisé lorsqu'une propriété est modélisée.

Cela fait près de 50 ans que la modélisation QSAR a d'abord été utilisée dans la pratique de l'agrochimie, de la conception de médicaments, de la toxicologie, de la chimie industrielle et environnementale. Son pouvoir croissant au cours de ces années peut aussi être attribué au développement rapide et étendu des méthodologies et des techniques de calcul qui ont permis de délimiter et d'affiner les nombreuses variables et les techniques utilisées dans cette approche de modélisation.

La modélisation QSAR est née dans le domaine de la toxicologie. En fait, les tentatives de quantification des relations entre la structure chimique et la puissance toxique aiguë font partie de la littérature toxicologique depuis plus de 100 ans. Dans la défense de sa thèse intitulée «Action de l'alcool amylique sur l'organisme» à la Faculté de médecine de l'Université de Strasbourg, France, le 9 janvier 1863, Cros a noté qu'une relation existait entre la toxicité des alcools aliphatiques primaires et leur solubilité dans l'eau. Cette relation a démontré l'axiome central de la modélisation structure-toxicité: la toxicité des substances est régie par leurs propriétés qui, à leur tour, sont déterminées par leur structure chimique. Par conséquent, il existe des interrelations entre la structure, les propriétés et la toxicité.

Il y a plus d'un siècle, Crum-Brown et Fraser (Crum-Brown et Fraser-Trans, 1868-1869) ont exprimé l'idée que l'action physiologique d'une substance dans un certain système biologique (Φ) était une fonction (f) de sa constitution chimique C:

$$\mathbf{\Phi = f\,(C)} \qquad (1)$$

Ainsi, une altération de la constitution chimique, ΔC, serait reflétée par une altération de l'activité biologique ΔΦ.

Au début du 20ème siècle, Meyer et Overton (Meyer, 1899; Overton, 1901) ont laissé entendre que l'action narcotique (dépresseur) d'un groupe de composés organiques était parallèle à leurs coefficients de partage huile d'olive / eau. Dans les années suivantes, sur le front physico-organique, le travail de Hammett a donné lieu à la culture "σ-ρ" (Hammet, 1935; Hammet, 1970) dans la délimitation des effets substitutifs sur les réactions organiques, tandis que Taft a conçu un moyen de séparer le polaire, le stérique et l'effet de résonance et l'introduction du premier paramètre stérique, ES (Taft, 1952).

En 1962 Hansch *et al*. ont publié leur étude sur les relations structure-activité des régulateurs de croissance des plantes et leur dépendance vis-à-vis des constantes de Hammett et de l'hydrophobie (Hansch *et al.,* 1962). En utilisant le système octanol / eau, une série complète de coefficients de partage a été mesurée, et donc une nouvelle échelle hydrophobe a été introduite. Le paramètre π, qui est l'hydrophobie relative d'un substituant, a été défini de manière analogue à la définition de sigma.

$$\pi = \log P_X - \log P_H \qquad (2)$$

P_X et P_H représentent les coefficients de partage d'un dérivé et de la molécule originelle, respectivement.

Les contributions de Hammett et Taft ont jeté les bases du développement du paradigme QSAR par Hansch et Fujita, qui ont combiné les constantes hydrophobes avec les constantes électroniques "σ-ρ" de Hammett pour produire l'équation linéaire de Hansch et ses nombreuses formes étendues.

Il existe un consensus parmi les toxicologues prédictifs actuels que Corwin Hansch est le fondateur du QSAR moderne. Dans son article classique (Hansch et Fujita, 1964), il a montré que, en général, l'activité biologique pour un groupe de produits chimiques «congénères» peut être décrite par un modèle complet:

$$Log 1/C_{50} = a\pi + b\varepsilon + cS + d \qquad (3)$$

dans lequel C, la concentration toxique à laquelle la toxicité se manifeste (par exemple, 50% de mortalité ou d'effet) est liée à un terme d'hydrophobie (qui est une constante de substituant indiquant la différence d'hydrophobie entre un composé parent et un analogue substitué, il a été remplacé par un terme moléculaire plus général le logarithme du coefficient de partage octanol / eau, log Kow). En raison de la relation curviligne, ou bilinéaire, entre log1 / C_{50} et

l'hydrophobie dans les doses uniques de test, le terme quadratique π^2 a été introduit plus tard dans le modèle.

La raison d'être de l'éq. (3) a été donnée par McFarland (1970). Il a supposé que l'activité relative d'une molécule biologiquement active, comme un produit toxique, dépend de:

(1) la probabilité (Pr1) que le toxique atteint son site d'action,

(2) la probabilité (Pr2) que le toxique interagira avec la cible sur ce site, et

(3) la concentration ou la dose externe.

La délimitation de ces modèles a entraîné un développement explosif dans l'analyse QSAR et les approches connexes (Hansch et Leo, 1995). Outre l'approche de Hansch, d'autres méthodologies ont également été développées pour aborder les questions d'activité structurelle. L'approche de Free-Wilson traite des études structure-activité dans une série congénère, comme décrit ci-après:

$$AB = \sum a_i x_i + u \qquad (4)$$

Lorsque AB est l'activité biologique, u est la contribution moyenne de la molécule mère, et a_i est la contribution de chaque caractéristique structurelle; x_i désigne la présence (x_i = 1) ou l'absence (x_i = 0) d'un fragment structurel particulier.

À l'heure actuelle, la science QSAR, fondée sur l'utilisation systématique de modèles mathématiques et sur le point de vue multivarié, est l'un des outils de base de la conception moderne des médicaments et des pesticides et joue un rôle croissant dans les sciences de l'environnement (Gramatica, http://www. QSARworld.com).

Le développement d'un modèle QSAR nécessite ces trois composantes: 1) un ensemble de données qui fournit des mesures expérimentales d'une activité biologique pour un groupe de produits chimiques; 2) la structure moléculaire et / ou les données de propriétés (c'est-à-dire les descripteurs, les variables ou les prédicteurs) pour ce groupe de produits chimiques; et 3) les méthodes statistiques, pour trouver la relation entre ces deux ensembles de données.

Le facteur limitant dans le développement des QSAR est la disponibilité de données expérimentales de haute qualité. Dans l'analyse QSAR, il est impératif que les données

d'entrée soient à la fois justes et précises pour développer un modèle significatif. En fait, il faut comprendre que tout modèle QSAR qui a été développé n'est valable que statistiquement comme les données qui ont conduit à son développement.

Les données utilisées dans les évaluations QSAR sont obtenues soit à partir de la littérature, soit générées spécifiquement pour les analyses de type QSAR. Ces données peuvent consister en séries congénères de produits chimiques ou assurer une diversité structurale même dans une classe chimique. Cette diversité a permis la généralisation de QSAR plus robustes, applicables de manière étendue. Un modèle structure-activité est défini et limité par la nature et la qualité des données utilisées dans le développement du modèle et devrait être exploité uniquement dans le domaine d'applicabilité de ce modèle.

Un modèle QSAR/ QSPR idéal devrait: (1) considérer un nombre suffisant de molécules pour une représentation statistique suffisante, (2) avoir une large gamme de Propriété/Activité quantifiée (c'est-à-dire plusieurs ordres de grandeur) pour les modèles de régression ou une répartition adéquate des molécules dans chaque classe (c'est-à-dire actifs et inactifs) pour les modèles de classification, (3) être applicable pour des prédictions fiables de nouveaux produits chimiques (domaine de validation et d'applicabilité) et (4) permettre l'obtention d'informations mécanistes sur l'Activité / Propriété modélisée. Les descripteurs chimiques comprennent des paramètres quantiques empiriques, ou non empiriques. Les descripteurs empiriques peuvent être mesurés ou estimés et intègrent des propriétés physicochimiques (par exemple logP). Les descripteurs non empiriques peuvent être basés sur des atomes individuels, des substituants ou toute la molécule, ce sont généralement des caractéristiques structurelles. Ils peuvent être basés sur la topologie ou la théorie des graphes et, en tant que tels, ils sont développés à partir de la connaissance de la structure 2D; ils peuvent également être calculés à partir des conformations structurelles 3D de la molécule.

Pareillement une variété de propriétés a été également utilisée dans la modélisation QSAR, notamment les propriétés physico-chimiques, quanto-chimiques et de liaison. Comme exemples de propriétés moléculaires nous citerons ; la distribution électronique, la disposition spatiale et le volume moléculaire. Les propriétés physico-chimiques comprennent des descripteurs pour les propriétés hydrophobes, électroniques et stériques d'une molécule ainsi que d'autres propriétés, y compris les constantes de solubilité et d'ionisation. Les propriétés quanto chimiques comprennent les valeurs de charge et d'énergie. Les propriétés de liaison

impliquent des macromolécules biologiques et sont importantes dans les réponses récepteur-médiation.

Dans les approches QSAR modernes, il est devenu assez courant d'utiliser une large gamme de descripteurs moléculaires théoriques de différents types, capable de saisir tous les aspects structurels d'un produit chimique pour transformer la structure moléculaire en nombres.

Beaucoup de logiciels calculent de larges ensembles de descripteurs théoriques différents, des SMILES, des graphiques 2D aux coordonnées 3Dx,y,z, parmi lesquels nous mentionnerons: ADAPT (Stuper, 1976), OASIS (Mekenyan et Bonchev, 1986), CODESSA (Katritzky et Lobanov, 1994), MolConnZ (2003) et DRAGON (Todeschini et al., 2006). On estime que plus de 10 000 descripteurs moléculaires sont maintenant disponibles, et la plupart d'entr'eux ont été résumés et expliqués (Devillers et Balaban, 1999; Karleson, 2000, Todeschini et Consonni, 2000). Le grand avantage des descripteurs théoriques est qu'ils peuvent être calculés de manière homogène par un logiciel défini pour tous les produits chimiques, même ceux qui n'ont pas encore été synthétisés, le seul besoin étant une structure chimique présumée, ils sont donc reproductibles.

La technique mathématique la plus utilisée est l'analyse de régression linéaire multiple (MLR). C'est une approche simple qui aboutit à un résultat facile à comprendre et, pour cette raison, la plupart des QSAR sont construits à l'aide d'une analyse de régression. L'analyse de régression est un moyen puissant pour établir une corrélation entre les variables indépendantes ou explicatives (descripteurs moléculaires X) et une variable dépendante ou à expliquer Y, comme l'activité biologique:

$$Y = b + aX_1 + cX_2 + \cdots \qquad (5)$$

Selon Hansch, la sélection du descripteur est guidée par l'opinion du modélisateur pour avoir une connaissance a priori du mécanisme de l'activité / propriété étudiée et la présomption d'attribuer un sens mécaniste à tout descripteur moléculaire utilisé choisi parmi un groupe limité. Des variables de modélisation potentielles, normalement connues et utilisées à plusieurs reprises (par exemple: log K_{OW} est un paramètre universel mimant la perméation de la membrane cellulaire, ainsi utilisé dans beaucoup de modèles de toxicité, mais est également lié à divers coefficients de partage tels que la bioconcentration / bioaccumulation, coefficient de sorption du sol.

D'autre part, l'approche «statistique» ou chimiométrique, une approche parallèle à la précédente, appelée «mécaniste», repose sur la conviction fondamentale que le modélisateur QSAR ne devrait pas influencer, a priori et personnellement, la sélection du descripteur par hypothèses mécanistes, mais devrait appliquer des outils mathématiques impartiaux pour sélectionner, à partir d'une large gamme de descripteurs d'entrée, les descripteurs les plus corrélés à la réponse étudiée. Le nombre et la typologie des descripteurs d'entrée disponibles doivent être aussi larges et aussi différents que possible afin de garantir la représentation de n'importe quel aspect de la structure moléculaire. Les différents descripteurs sont des façons ou des perspectives différentes de voir une molécule, mais les modèles doivent être développés en tenant compte du principe de parcimonie, nommé Okham's Razor: "les entités ne doivent pas être multipliées au-delà du nécessaire" ou "éviter la complexité si pas nécessaire ". Ce principe est souvent paraphrasé comme "La solution la plus simple est la meilleure".

En ce qui concerne l'interprétation des descripteurs, il est important de prendre en compte que la réponse modélisée est souvent le résultat d'une série de mécanismes biologiques ou physico-chimiques complexes, il est donc très difficile d'accorder trop d'importance au sens mécaniste des descripteurs moléculaires utilisés dans un modèle QSAR.

En outre, il faut souligner que, dans des modèles multivariés tels que les modèles MLR, bien que l'interprétation d'un seul descripteur moléculaire puisse être certainement utile, c'est seulement la combinaison de l'ensemble de descripteurs sélectionnés qui est capable de modéliser la propriété / activité étudiée. Si l'objectif principal de la modélisation QSAR est de combler les lacunes dans les données disponibles, l'attention du modélisateur devrait être axée sur la qualité du modèle. En ce qui concerne ce point, Livingstone (2000) affirme: «Le besoin d'interprétation dépend de l'application, puisqu'il s'agit d'une méthode mathématique validée. Le modèle reliant une propriété cible aux caractéristiques chimiques peut-être, dans certains cas, tout ce qui est nécessaire, bien qu'il soit évidemment souhaitable d'essayer d'expliquer le «mécanisme» en termes chimiques, mais ce n'est souvent pas nécessaire, en soi ». Zefirov et Palyulin (2001) ont adopté la même position, en différenciant les QSAR prédictifs, où l'attention concerne essentiellement la meilleure qualité de prédiction, des QSAR descriptifs où une attention majeure est accordée à l'interprétation du descripteur.

En fait, le premier objectif de tout modélisateur devrait être la validation pour l'application prédictive du modèle QSAR, tant pour l'approche mécaniste que pour la

statistique. Le fameux "Paradoxe de Kubinyi" (Van Drie, 2003; Bultinck, 2004), souligné également par Tropsha *et al*. dans leurs célèbres articles: méfiez-vous de Q^2 (Beware of Q^2) (Golbraikh et Tropsha, 2002) et l'importance d'être sérieux (The Importance of being Earnest) (Tropsha et al, 2003) c'est que: les modèles « meilleurs ajustements » ne sont pas les meilleurs pour la prédiction! En fait, un modèle QSAR doit, avant tout, être un modèle réel, robuste et prédictif, pour être considéré comme un modèle fiable (Gramatica, 2007); seul un modèle stable et prédictif peut être utilement interprété pour sa signification mécaniste, même si cela n'est pas toujours facile ou réalisable.

Ces dernières années, la validation des modèles QSAR a été reconnue par des groupes d'experts spécifiques de l'OCDE comme un point crucial et urgent, ce qui a conduit au développement, pour des raisons réglementaires, des « principes de l'OCDE pour la validation des modèles (Q)SAR » (www.oecd.org). La nécessité de cette action importante est principalement due à la nouvelle politique des produits chimiques de la Commission Européenne (REACH: enregistrement, évaluation et autorisation des produits chimiques) (www.europa.eu.it), qui énonce explicitement la nécessité d'utiliser des modèles (Q) SAR pour réduire les tests expérimentaux (y compris les essais sur les animaux). De toute évidence, pour satisfaire aux exigences de la législation REACH, il est essentiel d'utiliser des modèles (Q) SAR qui produisent des estimations fiables, c'est-à-dire des modèles (Q) SAR validés. Ainsi, un modèle QSAR fiable doit être associé aux informations suivantes: 1) Une définition précise de la propriété prédite par le modèle, incluant le protocole et les conditions expérimentales; 2) Une équation mathématique (ou un algorithme) sans équivoque (reproductible), incluant la définition des différents paramètres employés ainsi que les méthodes de calculs éventuellement utilisées pour les obtenir; 3) Un domaine d'applicabilité défini, permettant de déterminer pour quelles molécules les prédictions sont fiables; 4) Des mesures appropriées des performances du modèle en termes de corrélation et de prédiction, incluant donc la mesure de son pouvoir prédictif pour un jeu de molécules de validation; 5) Si possible, une interprétation des mécanismes moléculaires mis en jeu au travers des descripteurs employés et de la structure du modèle.

Le besoin d'interprétation dépend de l'application, car un modèle mathématique validé reliant une propriété cible à des caractéristiques chimiques peut être tout ce qui est nécessaire, en particulier lorsque des données prédites sont nécessaires pour le criblage de grandes bases de données de produits chimiques, bien qu'il soit évidemment souhaitable de tenter certaines

explications du «mécanisme» en termes chimiques (Livingstone, 2000; Zefirov et Palyulin, 2001).

Références bibliographiques

Crum-Brown A, Fraser-Trans T R, On the connection between chemical constitution and physiological action. Part 1. R. Soc. Edinburgh. 1868–1869. 25, 151.

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

Golbraikh A, Tropsha A, Beware of q²!. J. Mol.Graph. Mod. 2002. 20, 269.

Gramatica P, A short story of QSAR evolution,http://www.QSARworld.com/Temp_fileup load/short story of QSAR.pdf

Gramatica P, Principles of QSAR models validation: internal and external, QSAR Comb.Sci. 2007. 26, 694.

Hammett L P, Physical Organic Chemistry, 2nd ed., McGraw-Hill, New York, 1970. p 420

Hammett L P, Some Relations between Reaction Rates and Equilibrium Constants. Chem. Rev. 1935. 17, 125.

Hansch C, Fujita T, A new substituent constant, π, derived from partition coefficients. J. Am. Chem. Soc. 1964. 86, 1616.

Hansch C, Leo A, Exploring QSAR, Fundamentals and Applications in Chemistry and Biology, ACS Professional Reference Book, American Chemical Society, Washington, DC, 1995. p 580.

Hansch C, Maloney P P, Fujita T, Muir R M, Correlation of biological activity of phenoxyacetic acids with Hammett substituent constants and partition coefficients. Nature. 1962. 194, 178.

http://europa.eu.int/comm/environment/chemicals/reach.htm.

http://www.oecd.org/document/23/0,2340,en_2649_201185_33957015_1_1_1_1,00. html.

Karelson M, Molecular descriptors in QSAR/QSPR. New York: Wiley-InterScience, 2000. 448

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

Livingstone D J, The Characterization of Chemical Structures Using Molecular Properties. A Survey. J. Chem. Inf. Comput. Sci. 2000. 40, 195.

McFarland J W, Parabolic relation between drug potency and hydrophobicity. J. Med. Chem. 1970. 13, 1092.

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

Meyer H, What is the property of anesthetics. Arch.Exp. Pathol. Pharmakol. 1899, 42, 109.

MolConnZ, Ver. 4.05, 2003, Hall Ass. Consult., Quincy, MA

Overton C E, Studien Uber die Narkose, Fischer, Jena, Germany. 1901.

Patrick B, Hans D W, Wilfried L, Jan P. Tollenare. Computational Medicinal Chemistry for Drug Discovery. Eds., Marcel Dekker, 2004. p 1169

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

Taft R W, polar and steric substituent constants for aliphatic and o-benzoate groups from rates of esterification and hydrolysis of esters. J. Am. Chem. Soc. 1952. 74, 3120.

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

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

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

Van Drie J H, Pharmacophore discovery-lessons learned. Curr. Pharm.Des. 2003. 9, 1649.

Zefirov N S, Palyulin V A, QSAR for boiling points of "small" sulfides. Are the "high-quality structure property activity regressions" the real high quality QSAR models?. J. Chem. Inf. Comput. Sci. 2001. 41, 1022.

BASES THÉORİQUES

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

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

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

I. Optimisation de la géométrie des molécules

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

II.1.1. Energie d'un micro système représenté par un déterminant de Slater

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

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

$$\psi(1,2,\dots,n)=\frac{1}{\sqrt{n!}}\begin{vmatrix}\psi_1(1) & \bar{\psi}_1(1)\dots\dots\dots\bar{\psi}_n(1)\\ \psi_1(2) & \bar{\psi}_1(2)\dots\dots\dots\bar{\psi}_n(2)\\ \vdots & \vdots \quad\quad\quad \vdots\\ \vdots & \vdots \quad\quad\quad \vdots\\ \psi_1(n) & \bar{\psi}_1(n)\dots\dots\dots\bar{\psi}_n(n)\end{vmatrix} \tag{6}$$

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

$$\psi_m(n)=\varphi_m(n)\alpha(n)$$
$$\bar{\psi}_m(n)=\varphi_m(n)\beta(n) \tag{7}$$

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

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

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

$$H(1,2,\dots,n)=\sum_{i=1}^{n}h_{(i)}^{c}+\sum_{i<j}\frac{e^2}{r_{ij}} \tag{8}$$

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

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

$$E=<\psi|H|\psi> \tag{9}$$

Lorsqu'on développe cette intégrale on arrive (Pople et Beveridge, 1967) au résultat :

$$E=\sum_{i=1}^{m}2h_{ii}^{c}+\sum_{i=1}^{m}\sum_{j=1}^{m}(2J_{ij}-K_{ij}) \tag{10}$$

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

$$h_{ii}^{c}=<\psi_i(\mu)\left|h_{(\mu)}^{c}\right|\psi_i(\mu)> \tag{11}$$

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

$$J_{ij}=\iint\psi_i^*(\mu)\psi_i(\mu)\frac{e^2}{r_{\mu\nu}}\psi_j^*(\nu)\psi_j(\nu)d\tau_\mu d\tau_\nu \tag{12}$$

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

$$K_{ij} = \iint \psi_i^*(\mu)\psi_i^*(\nu)\frac{e^2}{r_{\mu\nu}}\psi_j(\mu)\psi_j(\nu)d\tau_\mu d\tau_\nu \qquad (13)$$

est **l'intégrale biélectronique moléculaire d'échange** ; elle représente également une somme de répulsions entre charges élémentaires, l'électron occupant deux orbitales moléculaires ψ_i et ψ_j.

$r_{\mu\nu}$ représente la distance entre les deux électrons μ et ν.

Remarques :

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

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

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

$$E=E^c+E^{RE} \qquad (14)$$

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

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

$$E_T = E + \sum_{N<L}\frac{Z_K E_L e^2}{R_{KL}} \qquad (15)$$

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

La relation (10) est équivalente à :

$$E = \sum_{i=1}^{m}\{h_{ii}^c + (h_{ii}^c + \sum_{j=1}^{m}(2J_{ij} + K_{ij})\} \qquad (16)$$

Le terme :

$$e_i = h_{ii}^c + \sum_{j=1}^{m}(2J_{ij} + K_{ij}) \qquad (17)$$

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

E se réduit donc à :

$$E = \sum_{i=1}^{m}(h_{ii}^{c} + e_i) \quad (18)$$

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

$$E = \sum_{i=1}^{m} 2\, e_i \quad (19)$$

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

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

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

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

II.1.2. Détermination des Orbitales ou équations de Hartree-Fock

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

Quelles propriétés doivent posséder les ψ_i pour être acceptables au sens de la mécanique ondulatoire, et qu'elles puissent s'adapter au système particulier envisagé ?

Il faut que le déterminant de Slater soit une solution approchée de l'équation de Schrödinger totale :

$$H(1, 2, \ldots., n)\psi(1, 2, \ldots., n) = E\psi(1, 2, \ldots., n) \quad (20)$$

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

Comme la variation du déterminant de Slater s'exprime par la variation du jeu des $\{\psi_i\}$, il faudrait avoir, pour une variation première du jeu d'orbitales choisies, une variation première de l'énergie totale nulle, et pour cela il faut que les ψ_i soient solutions des équations de Hartree-Fock (Hatree, 1928 ; Fock,1930 ; Slater,1928):

$$\{\delta\psi_i\} \rightarrow \delta E^1 = 0 \qquad (21)$$

Ces deux conditions contiennent les équations de Hartree-Fock :

$$F_{(\mu)}\psi_i(\mu) = e_i\psi_i(\mu) \qquad (22)$$

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

Opérateur de Hartree-Fock :

$$F_{(\mu)} = \left[h^c_{(\mu)} + \sum_{i=1}^{m} 2J_l(\mu) - K_l(\mu)\right] \qquad (23)$$

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

II.1.3. Equations de Roothaan et Hall

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

Chaque orbitale moléculaire psi se présentera sous la forme :

$$\psi_i(\mu) = \sum_{p=1}^{N} C_{pi}\varphi_p(\mu) \qquad (24)$$

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

Les équations de Hartree-Fock prennent, en tenant compte de (24), une expression vectorielle assez simple :

$$\sum_{p=1}^{N} C_{pi}\left[F_{pq} - e_i S_{pq}\right] = 0 \quad , \quad q \in [1, N] \qquad (25)$$

Les coefficients :

$$S_{pq} = \int \varphi_p^* \varphi_q d\tau$$

$$F_{pq} = \int \varphi_p^* (F\varphi_q) d\tau \qquad (26)$$

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

L'équation (25) est un système linéaire homogène (*N* équations à *N* inconnues) qu'on peut écrire sous la forme matricielle :

$$[\mathbf{F}- e_i \mathbf{S}]\mathbf{C_i} = \mathbf{0} \qquad (27)$$

Où **F** est la matrice $[F_{pq}]$; **S** est la matrice $[S_{pq}]$; $\mathbf{C_i}$ est la matrice $[C_{pi}]$.

$$F_{pq} = h^c_{pq} + \sum_{l=1}^{N}\sum_{m=1}^{N} p_{lm}\left[< pq \left| lm > - \frac{1}{2} < pm \right| lq >\right] \qquad (28)$$

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

$$h^c_{pq} = \left\langle \varphi_p(\mu)\left| h^c_{(\mu)} \right| \varphi_q(\mu)\right\rangle \qquad (29)$$

-* $\langle pq|lm\rangle = \iint \varphi_p(\mu)\varphi_q(\mu)\frac{e^2}{r_{\mu\nu}}\varphi_l(\nu)\varphi_m(\nu)d\tau_\mu d\tau_\nu$ (30)

$$\langle pm|lq\rangle = \int \varphi_p(\mu)\varphi_m(\mu)\frac{e^2}{r_{\mu\nu}}\varphi_l(\nu)\varphi_q(\nu)d\tau_\mu d\tau_\nu \qquad (31)$$

-* $p_{lm} = \sum_{i=1}^{N} 2C_{li}C_{mi}$ =éléments de la matrice densité (32)

-* $\boldsymbol{P} = [p_{lm}]$ = matrice densité (33)

II.1.4. Quelques remarques sur les processus de résolution des équations de Hartree-Fock-Roothaan.

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

$$[\mathbf{F}- e_i \mathbf{S}]\mathbf{C_i} = \mathbf{0} \qquad (27)$$

Löwdin (1950) a proposé un procédé qui permet de se ramener dans tous les cas au calcul des valeurs propres et vecteurs propres d'une matrice moyennant une transformation de la base des orbitales atomiques (**orthogonalisation de Löwdin**).

Multiplions à gauche les deux membres de (27) par la matrice $\mathbf{S}^{-1/2}$, qui n'est jamais singulière puisque **S** ne l'est pas ; il vient successivement:

$$\boldsymbol{S}^{-1/2}\boldsymbol{F}\boldsymbol{C}_i = ei\,\boldsymbol{S}^{-1/2}\boldsymbol{S}\boldsymbol{C}_i$$

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

Soit en posant :

$$\boldsymbol{S}^{-1/2}\,\boldsymbol{F}\,\boldsymbol{I}\,\boldsymbol{S}^{-1/2} = \overline{\boldsymbol{F}} \quad \text{et} \quad \boldsymbol{S}^{-1/2}\,\boldsymbol{C_i} = \overline{\boldsymbol{C_i}} \tag{34}$$

$$\overline{\boldsymbol{F}}\,\overline{\boldsymbol{C}}_{\boldsymbol{i}} = e_i\,\overline{\boldsymbol{C}}_{\boldsymbol{i}} \quad \text{c'est-à- dire } [\overline{\boldsymbol{F}} - e_i\,\boldsymbol{I}]\overline{\boldsymbol{C}}_{\boldsymbol{i}} = \boldsymbol{0} \tag{35}$$

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

$$\overline{\boldsymbol{F}}\,\overline{\boldsymbol{C}} = e_i\,\overline{\boldsymbol{C}}_{\boldsymbol{i}} \tag{35}$$

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

II.1.5. Détermination des intégrales de la méthode de Hartree-Fock-Roothaan (HFR)

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

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

$$h_{pq}^{c} = < \varphi_p(\mu) \left| h_{\mu}^{c} \right| \varphi_q(\mu) > \tag{36}$$

Il existe deux types d'intégrales de ce genre : **monocentres** lorsque φ_p et φ_q appartiennent au même atome R, **bicentres**, lorsque φ_p et φ_q appartiennent à des atomes différents.

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

$$h_{pq}^{c} = \overbrace{-\frac{\hbar^2}{2m}\int \varphi_p(\mu)\Delta_{(\mu)}\varphi_q(\mu)d\tau_\mu}^{Intégrales\ cinétiques} \overbrace{- \sum_k Z_K \int \varphi_p(\mu)\,\frac{e^2}{r_{k\mu}}\,\varphi_q(\mu)d\tau_\mu}^{intégrales\ d'attractions\ nucléaires} \tag{37}$$

Les intégrales d'attractions nucléaires peuvent être monocentres, bicentres ou tricentres (très compliquées à calculer).

- **Intégrales bi-électroniques**

$$G_{pq} = \sum_l \sum_m p_{lm} \left[\langle pq|lm\rangle - \frac{1}{2}\,\langle pm|lq\rangle\right] \tag{38}$$

$\langle pq|lm\rangle$, $\langle pm|lq\rangle$ et p_{lm} sont respectivement définis par les relations (30), (31) et (32).

On a plusieurs types d'intégrales :

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

Le calcul des intégrales biélectroniques prend le plus grand temps, et il n'est pas possible, en prenant des orbitales de Slater (39) d'en donner des expressions analytiques.

$$\varphi_{n,l,m}(k,\vec{r}) = Nr^{n-1}e^{-kr}y_{l,m}(\theta,\varphi) \qquad (39)$$

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

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

II.2. Les méthodes semi-empiriques

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

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

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

II.2.1. Définition du semi-empirisme

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

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

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

II.2.2. Quelques théories semi-empiriques

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

La première théorie des orbitales moléculaires semi-empirique tri-dimensionnelle est l'approximation au recouvrement différentiel nul (CNDO pour : Complete Neglect of Differential Overlap**)**, introduite par Pople, Santry et Segal (Pople *et al.*, 1965), pour être appliquée à tous les électrons de valence de molécules quelconques organiques ou minérales.

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

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

et - Théorie de Hartree- Fock (HF).

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

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

Cependant, en 1965, les ressources computationnelles nécessaires pour l'approche SCF complète n'étaient pas encore disponibles. La pratique des théories des orbitales moléculaires nécessitaient donc encore des approximations. Le principal problème réside dans le calcul et le stockage des intégrales tétracentres notées $\langle \mu\nu|\lambda\sigma\rangle$, nécessaires pour le calcul des interactions électron- électron dans le cadre de l'approximation SCF. Les indices $\mu, \nu, \lambda \; et \; \sigma$ dénotent quatre centres d'orbitales atomiques de sorte que le nombre de telles orbitales à calculer croît proportionnellement à N^4, où N est le nombre d'orbitales atomiques. En fait, le nombre de telles intégrales n'est pas exactement égal à la puissance quatrième du nombre de fonctions de base parce que beaucoup d'entr'elles sont reliées par symétrie. Ce qui était une tâche très difficile en 1965 ; ainsi Pople, Santry et Segal ont introduit (Pople *et al.*,1965) l'approximation que seules les intégrales pour lesquelles $\mu = \nu$ et $\lambda = \sigma$ c'est-à-dire : $\langle \mu\mu|\nu\nu\rangle$ seront prises en compte et que, de plus, toutes les orbitales atomiques seront traitées de la même façon (comme si elles étaient des orbitales s), de sorte que l'équation (40) s'applique, où μ est centrée sur l'atome A et λ sur l'atome B et ainsi γ_{AB} ne dépend que des identités de A et B, et peut être traité comme paramètre.

$$\langle \mu\mu|\lambda\lambda\rangle = \gamma_{AB} \tag{40}$$

Une première approximation, due à Pariser et Parr (Pariser et Parr,1953) consiste à traiter le terme mono-centre γ_{AA} comme différence entre le potentiel d'ionisation PI_A et l'affinité électronique AE_A de A [Eq.(41)] :

$$\gamma_{AA} = PI_A - AE_A \tag{41}$$

Les termes di-centres sont alors données par l'éq.(42) :

$$\gamma_{AB} = \frac{\gamma_{AA} + \gamma_{BB}}{2 + r_{AB}(\gamma_{AA} + \gamma_{BB})} \tag{42}$$

Ce qui conduit à : $\gamma_{AB} = (\gamma_{AA} + \gamma_{BB})/2$ pour une distance interatomique, r_{AB}, nulle et $\gamma_{AB} \approx {}^{1}/_{r_{AB}}$ pour des distances interatomiques plus grandes. Ces expressions (Eqs. (40) – (42)) montrent la simplicité de la technique CNDO, qui a été utilisée pour calculer des propriétés électroniques comme les moments dipolaires ou les énergies d'excitation, généralement à partir des géométries expérimentales. Il y a eu beaucoup de modifications des eqs. (41 et (42), mais elles restent d'une simplicité comparable. Pareillement, des expressions simplifiées ont aussi été utilisées pour les intégrales mono-électroniques.

Cependant, la méthode CNDO montra des insuffisances systématiques directement imputées aux simplifications ébauchées précédemment, aussi fut-elle remplacée par la méthode **INDO (Intermediale Neglect of Differential Overlap)**, introduite en 1967 par Pople, Beveridge et Dobosh (Pople *et al.*, 1967). L'approximation qui conduit à l'éq. (40) s'étant avérée très sévère, elle fut remplacée par des valeurs individuelles pour les différents types d'interactions entre deux orbitales atomiques. Ces valeurs individuelles, souvent désignées par G_{ss}, G_{sp}, G_{pp} et G^2_{pp} dans la littérature, peuvent être ajustées pour donner un accord avec l'expérience meilleur que celui obtenu avec la méthode CNDO. Cependant, en INDO les termes di-centres sont maintenus du même type que ceux apparaissant dans les éqs. (41) et (42). Cette approximation conduit à des affaiblissements systématiques, comme par exemple dans le traitement des interactions entre doublets isolés.

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

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

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

Les méthodes MINDO sont maintenant largement obsolètes.

La méthode avantageuse pour la plupart des techniques modernes d'orbitales moléculaires semi-empiriques est la MNDO, qui a été publiée par Dewar et Thiel en 1977

(Dewar et Thiel, 1977; Thiel, 1998). La MNDO est une méthode NDDO dans laquelle Dewar Et Thiel ont introduit un formalise basé sur les multipoles pour le calcul des intégrales bi-électroniques. Elle a été paramétrée pour reproduire les chaleurs de formation expérimentales, les géométries, les moments dipolaires et les potentiels d'ionisation. Elle s'avéra très supérieure aux méthodes MINDO pour la plupart des grandeurs calculées. Cependant la MNDO présente une faiblesse qui limite sévèrement son utilité ; elle ne reproduit pas la liaison hydrogène. Cette faiblesse a été surmontée de façon pragmatique par Bustein et Isaev (Bustein et Isaev,1984) qui modifièrent simplement le potentiel de répulsion cœur-cœur par addition de fonctions gaussiennes en vue d'obtenir des liaisons hydrogène. Ce « fixe » a été adopté par le groupe Dewar pour leur méthode suivante AM1 (Dewar *et al.*,1985; Holder, 1998) qui est par ailleurs identique à la MNDO. AM1, en retour, s'avéra présenter une faiblesse dans le traitement des composés nitrosés et hypervalents. Ces faiblesses ont été abordées par Stewart dans une nouvelle paramétrisation nommée PM3 (Stewart, 1989-1998) qui est par ailleurs identiques à AM1. Cependant, MNDO, MNDO/H, AM1 et PM3 sont pour l'essentiel identiques du point de vue quanto-mécanique. Leurs différences se limitent à la « correction » classique des potentiels entre atomes et pour laquelle les paramètres sont traités comme variables dans la procédure de paramétrisation.

II.2.3. Limites et avantages des méthodes semi-empiriques (Jensen, 2007)

La négligence de toutes les intégrales bi-électroniques tri et tétracentres réduit la matrice de Fock d'un ordre formel M^4 à M^2. Toutefois, le temps requis pour la diagonalisation de la matrice ***F*** croît comme le cube de la dimension de la matrice. La diagonalisation d'une matrice devient importante lorsque la dimension dépasse ~10 000 x 10 000. De nombreuses itérations sont nécessaires pour la résolution des équations SCF, et habituellement la géométrie est également optimisée, nécessitant de nombreux calculs pour différentes géométries. Ce qui situe la limite actuelle des méthodes semi-empiriques à environ 1000 atomes. Il est à noter que la méthode classique de résolution des équations HF par diagonalisation de la matrice de Fock s'impose rapidement comme l'étape limitante réelle dans les méthodes semi-empiriques. Des développements ultérieurs se sont ainsi focalisés sur la formulation de méthodes alternatives pour l'obtention d'orbitales SCF sans passer par la diagonalisation (Stewart, 1996; Daniels et al.,1997). De telles méthodes utilisent des ajustement (combinaisons) linéaires avec le nombre d'atomes, ce qui permet d'effectuer des calculs pour des systèmes comprenant plusieurs milliers d'atomes.

La paramétrisation de MNDO/AM1/PM3 est réalisée en ajustant les constantes impliquées dans les différentes méthodes de façon à ce que les résultats des calculs HF ajustent les données expérimentales aussi près que possible. Ce qui est faux dans un sens. On sait que la méthode HF ne peut conduire au résultat correct, même à la limite d'un ensemble de base infini et sans approximations. Les résultats HF ne reproduisent pas la corrélation électronique, mais les données expérimentales impliquent naturellement de tels effets. Ceci peut être considéré comme un avantage, les effets de corrélation électronique sont implicitement pris en compte dans la paramétrisation, et il n'est pas besoin d'exécuter des calculs compliqués pour surmonter les déficiences de la procédure HF. Cependant, il y a réellement problème quand la fonction d'onde HF ne peut décrire le système correctement, même qualitativement, comme par exemple avec les bi-radicaux et les états excités.

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

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

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

La forme la plus économique est celle pour laquelle seuls les micro-états dans lesquels un électron est promu de l'état fondamental à une orbitale virtuelle (excitations simples) sont utilisés. Ce qu'on désigne, dans une forme abrégée, par CIS. En ajoutant toutes les excitations doubles (pour lesquelles deux électrons sont promus) on est conduit à CISD, et ainsi de suite (Figure 4).

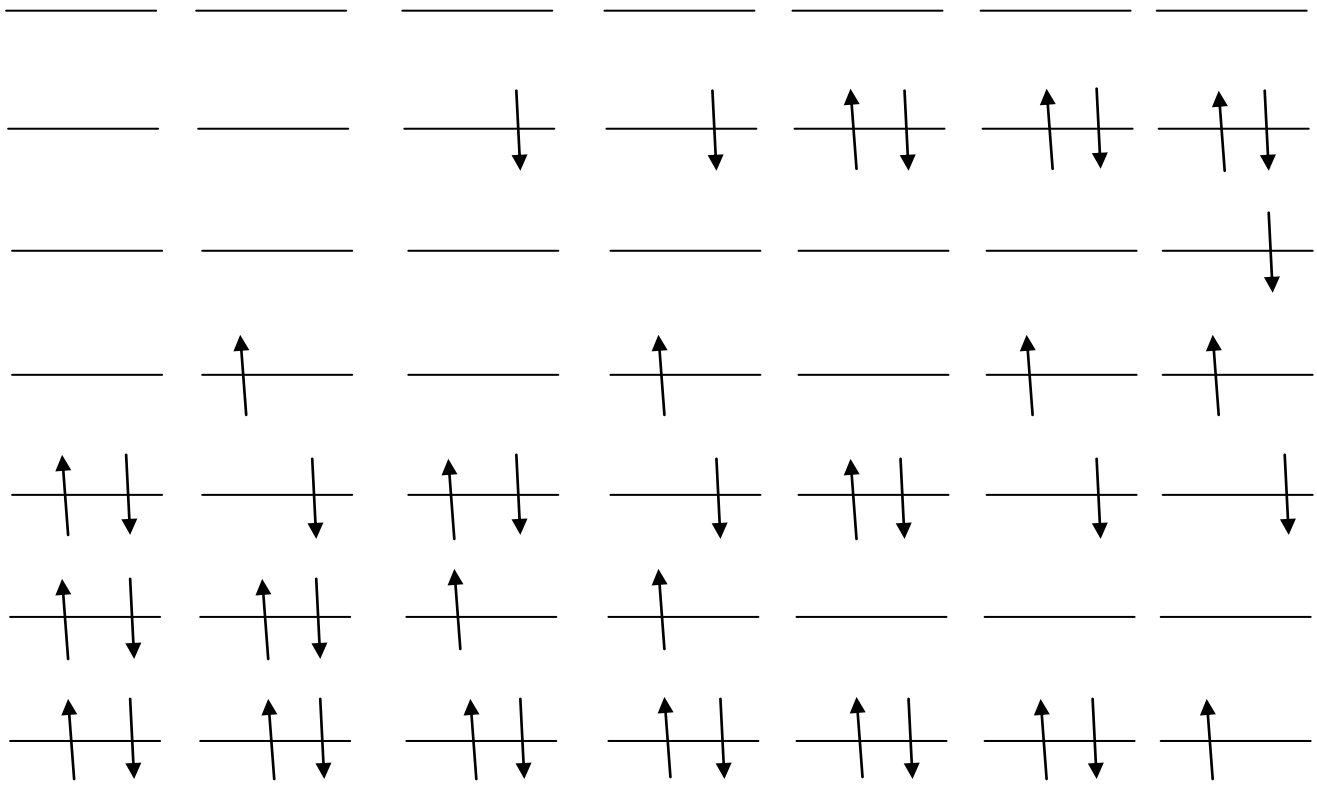


Figure. 4 Déterminants de Slater excités générés à partir d'une référence HF

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

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

$$\psi_{CI} = a_0\phi_{SCF} + \sum_{Simples\ \ (S)} a_s\phi_s + \sum_{Doubles\ \ (D)} a_D\phi_D + \cdots = \sum_{i=0} a_i\phi_i \qquad (43)$$

La méthode des multiplicateurs indéterminés de Lagrange (Ramachandran *et el.*, 2008) est ensuite appliquée pour minimiser l'énergie :

$$E = (\langle\psi|\hat{H}|\psi\rangle/\langle\psi|\psi\rangle) \quad (44)$$

Les méthodes semi-empiriques partagent les avantages/désavantages des méthodes de champ de force (cf : III), elles sont davantage performantes avec les systèmes pour lesquels on dispose de données expérimentales en quantités, mais il leur est impossible de faire des prédictions pour des types de composés totalement inconnus. La dépendance des données expérimentales n'est pas aussi sévère que pour la méthode du champ de force, à cause de la forme complexe de la fonctionnelle du modèle. Les méthodes NDDO nécessitent uniquement des paramètres atomiques, et nullement des paramètres di-, tri- et tétra-atomiques comme dans les méthodes de champ de force. Une fois un atome donné paramétré, tous les types de composés possibles contenant cet élément peuvent être traités. Le plus petit nombre de paramètres et la forme plus complexe de la fonctionnelle ont l'inconvénient, par rapport aux méthodes de champ de force, qu'il est très difficile de « réparer » un problème spécifique par reparamétrisation.

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

Les méthodes semi-empiriques fournissent une méthode de calcul de la fonction d'onde électronique, qui peut être utilisée pour la prévision d'une variété de propriétés. Il n'y a rien qui entrave le calcul, par exemple, de la polarisabilité d'une molécule, bien qu'il soit connu des calculs *ab-initio* que l'obtention de bons résultats nécessite un grand ensemble de base polarisé incluant des fonctions diffuses. Les méthodes semi-empiriques comme AM1 ou PM3 n'ont qu'une base minimale (absence de polarisation et de fonctions diffuses), la corrélation électronique n'est qu'implicitement incluse par les paramètres et aucune donnée de polarisabilité n'a été utilisée pour dériver ces paramètres. Il est douteux que de tels calculs puissent conduire à des résultats comparables à ceux fournis par l'expérience, et ils nécessitent, pour le moins, un calibrage soigné (Jensen, 2007) Encore une fois, il convient de souligner que la capacité d'effectuer un calcul ne garantit pas la fiabilité des résultats obtenus.

II.3. Analyse des distributions de charges

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

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

Il existe une famille d'analyses de population, parmi lesquelles nous citerons celles de Coulson et Longuet-Higgins (Coulson et Longuet-Higging, 1947), exprimée en termes de charges (ou « densités de charge ») et d'ordres de liaison, celle de Mulliken (Mulliken,1962), que nous rappellerons brièvement, et qui fait intervenir les populations atomiques et de recouvrement.

II.3.1. Analyse de population de Mulliken

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

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

$$P_\nu = \sum_k^{OM.occup\ ées} N_k C_{k\nu}^* C_{k\nu} \qquad (45)$$

où N_k est la population de l'O.M. ψ_k ; P_ν est la population électronique localisée dans l'O.A. $\varphi_{\mu\nu}$, que l'on appelle la population nette de l'O.A. φ_ν, dans la molécule.

$$R_{\mu\nu} = 2\sum_k^{OM.occup\ ées} C_{k\mu}^* C_{k\nu} S_{\mu\nu} \qquad (46)$$

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

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

$$\sum_\mu R_{\mu\nu} = \sum_\mu \sum_\nu P_{\mu\nu} S_{\mu\nu} = N \quad \text{[Décomposition sur les OA]} \qquad (47)$$

$$\int \psi^* \psi \, d\tau = N \quad \text{[Décomposition sur les OM]} \qquad (48)$$

Posons :

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

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

$$Q_M = \sum_{\mu(M)} q_\mu\ (M) \tag{49}$$

q_μ=densité électronique de l'orbitale μ ;

Q_M= densité électronique de l'atome M.

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

$$\delta_M = Z_M - Q_M \tag{50}$$

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

II.3.2. Calcul du moment dipolaire

Le moment dipolaire d'une molécule peut être décomposé, de façon unique, en trois composantes : une composante atomique ou d'hybridation, une composante de recouvrement, et une composante de transfert de charge (qui permet de définir les charges atomiques nettes), chacune étant définie de façon univoque dans le cadre du schéma OM-CLOA.

Dans ce schéma, l'expression en u.a du moment dipolaire d'une molécule, dans la convention des chimistes, est (Kutzelingg *et al.*, 1971)

$$\vec{\mu} = \sum_P \sum_Q \sum_{r\in P} \sum_{s\in Q} P_{rs}^{PQ} \int \varphi_r^* \vec{r}\, \varphi_s d_s d_r - \vec{\mu}_{nucl} \tag{51}$$

avec :

$$P_{rs}^{PQ} = \sum_i n_i C_{ir} C_{is} \tag{52}$$

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

On peut alors faire les substitutions suivantes :

$\vec{r}=\vec{r}_p+\vec{\xi}$, dans les termes tels que ***P* = *Q*** (53)

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

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

$$\vec{\xi}_{rr'}^{P} = \int \varphi_r^* \vec{\xi}\, \varphi_{r'} d\tau \;;\; \vec{\chi}_{rs}^{PQ} = \frac{\int \varphi_r^* \vec{\chi} \varphi_s d\tau}{S_{rs}^{PQ}} \quad (54)$$

Le moment dipolaire (51) devient (Mulliken, 1962) :

$$\vec{\mu} = \sum_p \delta_p \vec{r}_p + \vec{\mu}_{hybrid} + \vec{\mu}_{recouvr} \quad (55)$$

Avec :

$$\vec{\mu}_{hybrid} = \sum_p \sum_{r,r' \in P} P_{rr'}^{PP} \vec{\xi}_{rr'}^{P} \quad (56)$$

Et :

$$\vec{\mu}_{recouvr} = \sum_p \sum_{r,r' \in P} \sum_Q \sum_{s \in Q} P_{rs}^{PQ} S_{rs}^{PQ} \vec{\chi}_{rs}^{PQ} \quad (57)$$

II.3.3. Application

Nous avons réuni dans la figure 5 quelques applications (Pullman, 1969) des indices électroniques de la méthode des orbitales moléculaires.

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

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

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

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

$$P_f = \frac{q_{max} - q_{min}}{F(r_{max})} \qquad (58)$$

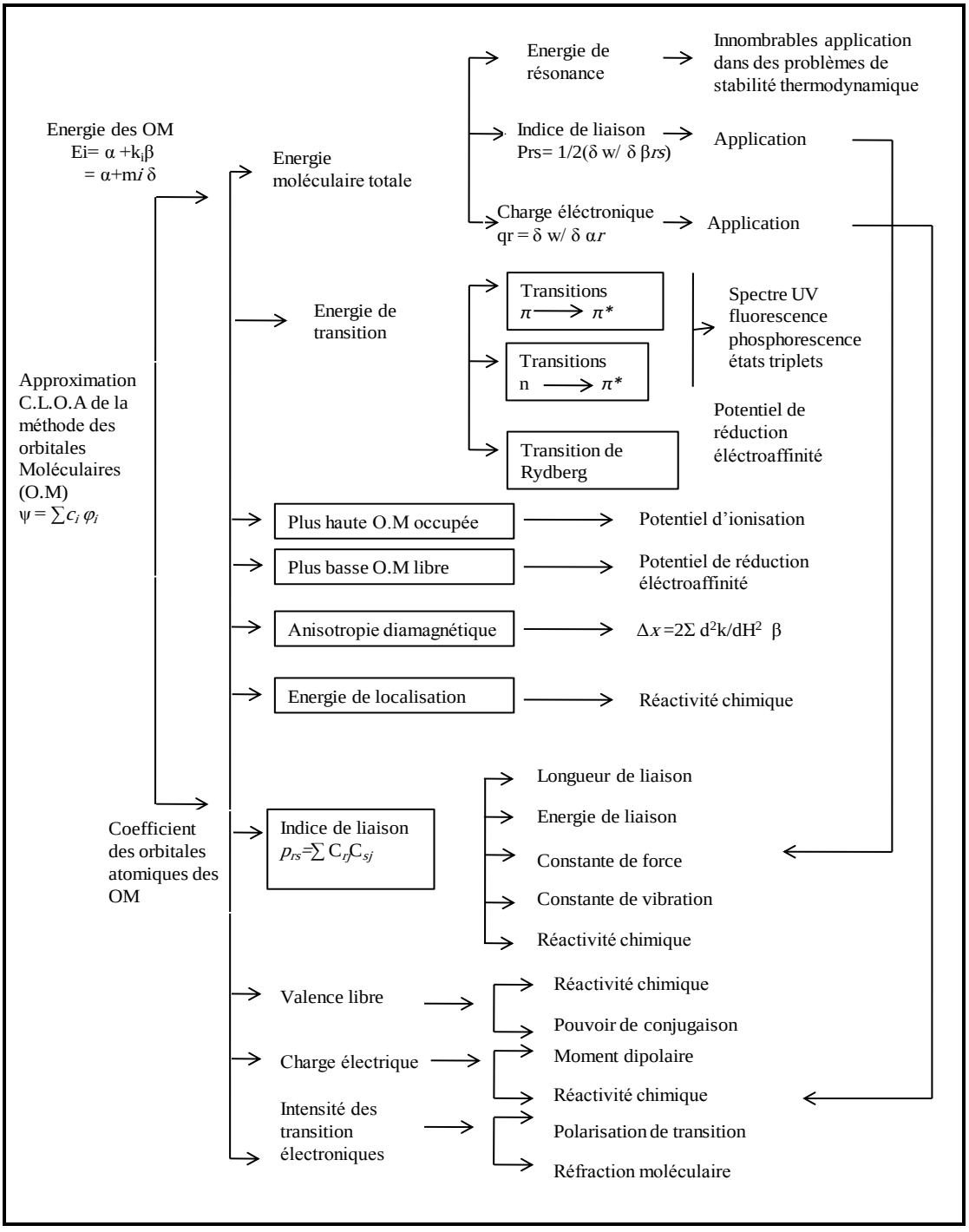


Figure. 5 Les indices électroniques de la méthode des orbitales moléculaires et leurs applications, d'après (Pullman, 1969).

III. LA MÉCANİQUE MOLÉCULAİRE

Si une molécule est trop grosse pour subir un traitement semi-empirique, il est toujours possible de modéliser son comportement en évitant complètement la mécanique quantique. Les méthodes désignées par mécanique moléculaire, établissent une expression algébrique simple de l'énergie d'un composé, sans avoir à calculer une fonction d'onde ou une densité électronique totale (Boyd et Lipkowitz, 2000). L'expression de l'énergie consiste en des équations classiques simples, comme l'équation de l'oscillateur harmonique, dans le but de décrire l'énergie associée à l'étirement de liaison, de flexion, de rotation, et aux forces intermoléculaires, telles que les interactions de Van Der Waals et de liaison hydrogène. Toutes les constantes apparaissant dans ces équations doivent être obtenues à partir de données expérimentales ou d'un calcul *ab initio*.

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

La mécanique moléculaire permet la modélisation de très grosses molécules, comme les protéines et des segments de DNA, la faisant le premier outil de la biochimie computationnelle. Le défaut de cette méthode est qu'il y a beaucoup de propriétés chimiques qui n'y sont pas définies, comme par exemple les états électroniques excités. De plus, pour travailler avec des systèmes très grands et très compliquées, les logiciels doivent être très puissants et faciles dans l'utilisation des interfaces graphiques.

III.1. Pas de calculs de champ de force sans définition préalable des types d'atomes.

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

dipôles de liaisons. Au contraire de la mécanique quantique, la mécanique moléculaire nécessite plus d'informations que le numéro atomique seul. En fait, chaque atome doit être décrit de manière plus détaillée.

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

III.2. Forme fonctionnelle des champs de force courants

Un champ de force ne consiste pas uniquement en une expression mathématique qui décrit l'énergie d'une molécule en fonction des coordonnées atomiques. La deuxième partie indispensable est le jeu de paramètres lui-même. Deux champs de force différents peuvent présenter la même forme fonctionnelle, mais utilisent un paramétrage complètement différent. D'un autre côté, différentes formes fonctionnelles peuvent conduire à des résultats presque identiques, en fonction des paramètres mis en jeu. Cette comparaison montre que les champs de force sont empiriques : il n'y a pas de forme « correcte ».

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

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

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

III. 3. Quelques exemples

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

- **MM2, MM3, et MM4 :** (http://europa.chem.uga.edu/allinger/mm2mm3.html).
Introduit par Allinger *et al*. (Alliger ,1977; Burkert et Alliger 1982,1986; Alliger *et al*.,1989, Alliger *et al*., 1996), largement utilisé pour le traitement de petites molécules.

- **AMBER :** (Assisted Method Building and Energy Refinement) (http://amber.scripps.edu)
Introduit par Cornell *et al.* (Mc kerell *et al.*, 1998) très largement utilisé dans le traitement des protéines et des acides nucléiques.

- **CHARMM :** (Chemistry at Harvad molecular Modeling) (http://yuri.harvard.edu)
Développé par Mackerall, Karplus *et al.*,(Brook *et al.*, 1983; Mackerel *et al.*, 1995,1998) qui est largement utilisé pour la simulation de petites molécules et jusqu'aux complexes solvatés de grandes macromolécules biologiques.

CHARMm est une version commerciale disponible de CHARMM qui est également applicable aux petits composés organiques (Momany et Rone, 1992).

-MMFF : (Merck Molecular Force Field)
Développé par Halgren (Halgren, 1996; Halgren et Nechbar, 1996), il est similaire à MM3 dans la forme, mais en diffère par son application focalisée sur les processus de condensation de phases (en dynamique moléculaire). Il reproduit l'exactitude MM3 pour les petites molécules, et est applicable aux protéines et autres systèmes d'importance biologique.

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

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

III.4. Représentation simple d'un champ de force

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

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

$$\upsilon(r^N) = \sum_{liaisons} \frac{k_i}{2}(I_i - I_{i,0})^2 + \sum_{angles} \frac{k_i}{2}(\theta_i - \theta_{i,0})^2 \sum_{torsions} \frac{V_n}{2}(1 - \cos(n\omega - \gamma)) + \sum_{i=1}^{N} \sum_{j=i+1}^{N} \left(4\varepsilon_{ij}\left[\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6}\right] + \frac{q_i q_j}{4\pi\varepsilon_0 r_{ij}}\right) \quad (59)$$

$\upsilon(\mathrm{r}^{\mathrm{N}})$ représente l'énergie potentielle qui est fonction des positions (**r**) des N particules (habituellement les atomes).

IV. LA DYNAMİQUE MOLÉCULAİRE

La dynamique moléculaire a débuté avec l'arrivée, en 1957, des premiers ordinateurs (Alder et Wainwright, 1957). Mais les premières simulations réelles ont été faites par Rahman (Rahman, 1964), 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 (Rahman et Stillinger, 1971) en 1971.

IV.1. Principe de la dynamique moléculaire

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

$$\vec{F}_i = m_i \vec{a}_i = m_i \frac{d^2 \vec{r}_i(t)}{dt^2} \quad (60)$$

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

m_i : masse de l'atome i.

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

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

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

$$E_c = \sum_{i=1}^{N} \frac{|\vec{P}_i|}{2n_i} \quad (61)$$

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

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

$$E_c = \frac{3K_b T}{2}(3N - N_c) \quad (62)$$

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

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

$$\vec{F}_i = \frac{d\vec{E}(r_i \ldots r_n)}{dr_i(t)} \quad (63)$$

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

r_i : coordonnées cartésiennes de l'atome i.

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

$$\vec{a}_i = \frac{d\vec{V}_i}{dt} \quad (64)$$

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

$$\vec{V}_i = \frac{d\vec{r}_i}{dt} \quad (65)$$

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

$$r_i(t + \Delta t) = r_i(t) + v_i \Delta t + \frac{1}{2} a_i \Delta t^2 \quad (66)$$

IV.2. Application de la dynamique moléculaire

Une application importante de la dynamique moléculaire est l'analyse des modes normaux de vibration le long de la trajectoire. Une autre application est l'optimisation et le raffinement des structures 3D d'après les données de la cristallographie et/ou de la RMN. La mise en œuvre de cette méthode requiert néanmoins des moyens de calcul particulièrement puissants et elle est coûteuse en temps et en argent. Elle se généralise cependant pour les études de peptides et de petites protéines (Mc Cammon et Harvery, 1978).

Références

Alder B J, Wainwright T E, Phase Transition for a Hard Sphere System, J. Chem. Phys. 1957.27, 1208.

Allinger N L, Conformational analysis. 130. MM2. A hydrocarbon force field utilizing V1 and V2 torsional terms. J. Am. Chem. Soc. 1977, 99, 8127.

Allinger N L, Yuk Y H, Lii J H, Molecular Mechanics. The MM3 Force Field for Hydrocarbon 3. 1 J. Am. Chem. Soc. 1989. 111, 8551.

Allinger N L. Chem K, Katzenellbogen J A, Wilson S R, Anstead G M, Hyperconjugative Effects on Carbon-Carbon Bond Lengths in Molecular Mechanics (MM4), J. Comput. Chem. 1996. 17, 747.

Boyd D B, Lipkowitz K B., eds. Reviews in Computational Chemistry, History of the Gordon Conferences on Computational Chemistry, Wiley- VCH, New York, 2000.399- 439.

Brooks B R, Bruccoleri R E, Olafson B D, States David J, Swaminathan S, Karplus M, CHARMM: A Program for Macromolecular Energy, Minimization, and Dynamics Calculations, J. Comput. Chem.1983.4, 187.

Burkert U, Allinger N L, Molecular Mechanics, ACS Monograph No. 177, American Chemical Society, Washington, DC, 1982, 1986.

Burstein K Y, Isaev A N, MNDO calculations on hydrogen bonds. Modified function for core-core repulsion, Theor. Chim. Acta. 1984. 64, 397.

Coulson C A, Longuet- Higgins H C, The Electronic Structure of Conjugated Systems. I. R Proc. Roy. Soc. (London) A .1947.191, 39 .

Daniels A D, Millam J M, Scuseria G E, Semiempirical methods with conjugate gradient density matrix search to replace diagonalization for molecular systems containing thousands of atoms, J. Chem. Phys. 1997.107 , 425.

Dewar M J S, Thiel W, Ground states of molecules. 38. The MNDO method. Approximations and parameters, J. Am. Chem. Soc., 1977, 99, 4899.

Dewar M J S, Zoebisch E G, Healy E F, Stewart J J P, Development and use of quantum mechanical molecular models. 76. AM1: a new general purpose quantum mechanical molecular model, J. Am. Chem. Soc. 1985. 107, 3902.

Fock V, Approximation method for solving the quantum mechanical multibody problem, Z. Physik. 1930.61, 126.

Halgren T A, Merck molecular force field. I. Basis, form, scope, parameterization, and performance of MMFF94, J. Comput. Chem. 1996. 17, 490.

Halgren T A, Nachbar R B, Merck Molecular Force Field. IV.Conformational Energies and Geometries for MMFF94, J. Comput. Chem, 1996.17, 587.

Hatree D R, The wave mechanics of an atom with a non-Coulomb central field. Part I. Theory and methods. Part I. Theory and Method. Proc. Cambridge. Phill. Soc. 1928.24, 328 .

Holder A J, Encyclopedia of Computational Chemistry, Vol. 1, P. v. R. Schleyer, N. L. Allinger, T. Clark, J. Gasteiger, P. A. Kollman, H. F. Schaefer III, P. R. Schreiner (Eds.), Wiley, Chichester. 1998, 8.

Jensen F, Introduction to Computational Chemistry, John Wiley & Sons Ltd, Chichester, , England. 2007, 94.

Kutzelnigg W, Delre G, Berthier G, s and p Electrons in Theoretical Organic Chemistry, Springer Verlag, Berlin, 1971.

Löwdin P O, On the Non-Orthogonality Problem Connected with the Use of Atomic Wave Functions in the Theory of Molecules and Crystals. J. Chem. Phys,1950. 18, 365.

Mac Kerell A D, Bashford D, Bellott M, Dumbrack R L, Evaseck J D, Field M J, Fischer S, Gao J, Gao H, He S, Joseph-Mac Carthy D, Kuchnir L, Kuczera K, Lau F T K, Mattos C, Michmick S, Nego T, Nguyen D T, Prodhom B, Reiher W E, Roux B, Schlemkrich M, Smith J C, Stote R, Straub J, Watanabe M, Wiorkiewicz-Kuczera J, Yin D, Karplus M, J. Phys. Chem. B. 1998. 102, 3586.

MacKerell A D, Bashford D, Bellott, Dunbrack R L, Evanseck J D, Field M J, Fischer S, Gao J, Guo H, Ha S, Joseph-McCarthy D, Kuchnir L, Kuczera K, Lau F T K, Mattos C, Michnick S, Ngo T, Nguyen D T, Prodhom B, Reiher W E, Roux B, Schlenkrich M, Smith J C, Stote R, Straub J, Watanabe M, Wirkiewicz-Kuczera J, Yin D, Karplus M, All-Atom Empirical Potential for Molecular Modeling and Dynamics Studies of Proteins. J. Phys. Chem. B. 1998. 102, 3586.

MacKerell A D, Wiókiewicz-Kuczera J, Karplus M, An All-Atom Empirical Energy Function for the Simulation of Nucleic Acids, J. Am. Chem. Soc. 1995.117, 11946.

McCammon J A, Harvey S C, Dynamics of Proteins and Nucleic Acids, Cambridge Univ. Press.1987, 234.

Momany F A, Rone R, Validation of the General Purpose QUANTAa3.2/CHARMm® Force Field J. Comput. Chem. 1992, 13, 888.

Mulliken R S. Criteria for the Construction of Good Self Consistent Field Molecular Orbital Wave Functions, and the Significance of LCAOMO Population Analysis , J. Chem. Phys. 1962.36, 3428.

Pariser R, Parr R G, A Semi-Empirical Theory of the Electronic Spectra and Electronic Structure of Complex Unsaturated Molecules I, J. Chem. Phys.1953. 21, 466.

Pariser R, Parr R G, A Semi-Empirical Theory of the Electronic Spectra and Electronic Structure of Complex Unsaturated Molecules. II J. Chem. Phys. 1953.21, 767.

Pople J A, Beveridge D L, Approximate Molecular Theory, Mc Graw- Hill. New York, 1970.

Pople J A, Beveridge D L, Dobosh P A, Approximate Self Consistent Molecular Orbital Theory. V. Intermediate Neglect of Differential Overlap J. Chem. Phys.1967. 47, 2026.

Pople J A, Electron interaction in unsaturated hydrocarbons, Trans. Faraday Soc.1953. 49, 1375.

Pople J A, Santry D P., Segal G A, Approximate Self Consistent Molecular Orbital Theory. I. Invariant Procedures. J. Chem. Phys.1965. 43, 5129 .

Pullman B, La Biochimie Electronique, Collection Que sais-je ? PUF, n°1075, Deuxième édition, Paris, 1969.

Rahman A F, Stillinger H., Molecular Dynamics Study of Liquid Water, J. Chem. Phys. 1971. 5, 3336.

Rahman A, Correlations in motion of atoms in liquid argon, Phys. Rev. 1964.136, 405.

Ramachandran K I, Deepa G, Namboori K, Comptutional Chemistry and Molecular Modeling, Principles and Applications, Springer- Verlag Berlin Heidebberg, 2008.

Slater J C. Phys. Rev. The self consistent field and the structure of atoms,1928. 32, 339 ; atomic shielding constants. 1930.35, 1210.

Stewart J J P, J. Comput. Chem. 10, 1989, 209- 220 ; 221- 264; PM3 : J. J. P. Stewart, Encyclopedia of Computational Chemistry; Vol. 3, Schleyer P V R, Allinger N L, Clark T, Gasteiger J, Kollman P A, Schaefer III H F, Schreiner P R. (Eds.), Wiley, Chichester, 1998, pp. 2080.

Stewart J J P. Application of Localized Molecular Orbitals to the Solution of Semiempirical Self-Consistent Field Equations, Int. J. Quantum. Chem. 1996. 58 , 133.

Thiel W, Encyclopedia of Computational Chemistry, Vol. 3, P. v. R. Schleyer, N. L. Allinger, T. Clarck, J. Gasteiger, P. A. Kollman, H. F. Schaefer III, P. R. Schreiner (Eds.), Wiley, Chichester, 1998, p. 1599.

PRİNCİPE ET MÉTHODES DES MODÈLES QSAR/QSPR

Le développement de nouvelles techniques de modélisation a permis la mise en place de nombreuses méthodes QSPR (QSPR : Relations Quantitatives Structure / Propriété) et QSAR (QSAR : Relations Quantitatives Structure /Activité) ; 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 de nouveaux composés ou des composés pour lesquels les données expérimentales ne sont pas disponibles. Dans cette partie, nous présentons une étude bibliographique sur les différentes méthodologies QSAR/QSPR ainsi que les différentes étapes de développement et de validation.

Les méthodes QSAR/QSPR (Roy *et al.*, 2015) sont basées sur l'hypothèse que l'activité ou la propriété d'un composé chimique est liée à sa structure, plus précisément cette approche indique que l'activité (ou la propriété) et la structure d'un composé chimique sont liées à certain algorithme mathématique. De plus, lorsque les paramètres moléculaires sont exprimés par des chiffres, on peut proposer une relation mathématique, ou relation quantitative structure activité/propriété. 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.

Les relations quantitatives structure-activité/propriété (QSAR/QSPR) sont de plus en plus utilisées, du fait de la croissance des moyens de calculs. Très récemment, la mise en place du nouveau règlement européen REACH (Registration, Evaluation, Authorisation and Restriction of Chemicals) (Margossian, 2007), qui recommande leur utilisation pour limiter le recours à l'expérience, donne un nouvel essor au développement de tels modèles prédictifs.

1. Principe des méthodes QSAR/QSPR

Comme leur nom l'indique le principe des méthodes QSAR/QSPR est de mettre en place une relation mathématique reliant de manière quantitative des propriétés moléculaires aussi bien électroniques que géométriques, appelées descripteurs, avec une observable macroscopique (activité biologique, toxicité, propriété physico-chimique, etc...), pour une série de composés chimiques similaires à l'aide de méthodes d'analyses de données. La forme générale d'un tel modèle est la suivante :

$$Propriété/Activité = f(Descripteurs) \quad (67)$$

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

- ✓ Régressions multi-linéaires (MLR) (Ghasemi *et al.*,2007),
- ✓ Régressions aux moindres carrés partiels (PLS) (Geladi et Kowalski, 1986),
- ✓ Arbres de décision (Myles *et al.*, 2004),
- ✓ Réseaux de neurones (Dupart *et al.*, 1998; Tetko *et al.*,1996; Gasteiger et Zupan, 1993),
- ✓ Algorithmes génétiques (Leardi, 2001),
- ✓ Machines à vecteur support (Gasteiger et Zupan, 1993).

Une fois cette relation établie et validée, elle peut alors être employée pour la prédiction de la propriété /activité de nouvelles molécules, pour lesquelles les valeurs expérimentales ne sont pas disponibles, voire pour des molécules non encore synthétisées. De tels modèles peuvent être également utilisés pour mieux comprendre les phénomènes moléculaires mis en jeu dans la propriété /activité d'intérêt et les modes d'action.

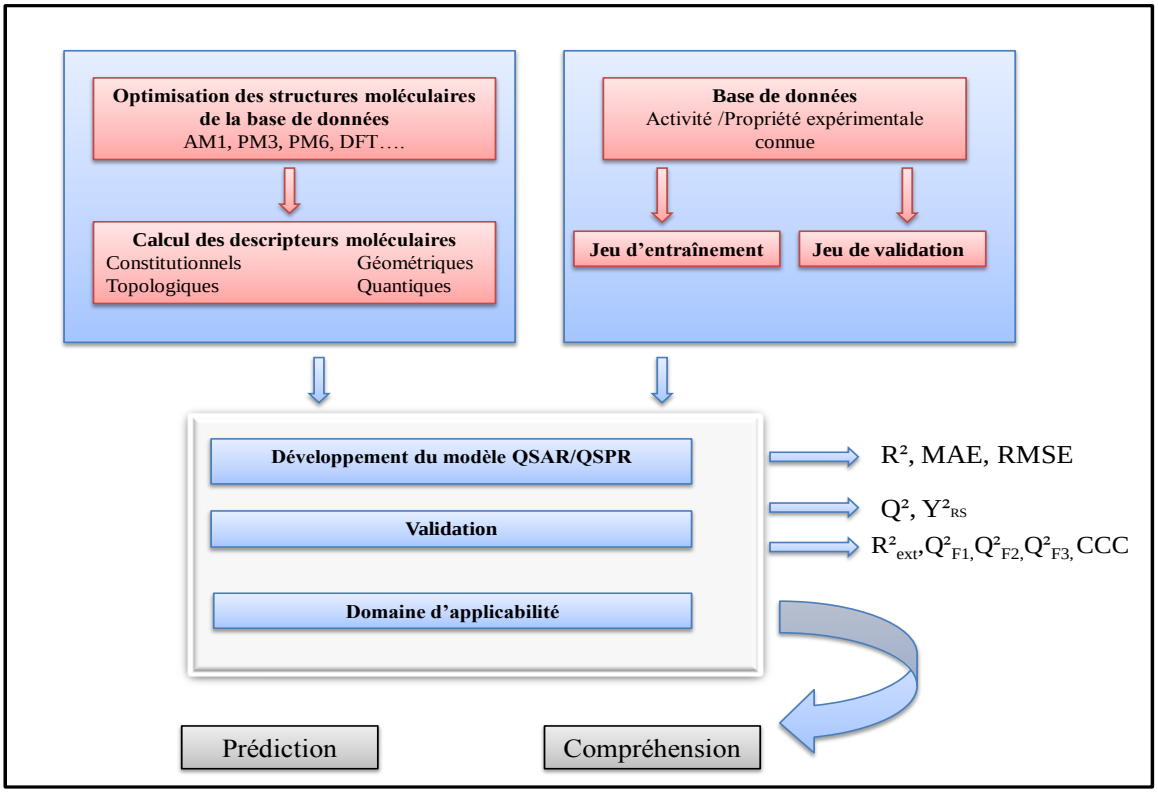


Figure. 6 Principes de la méthode QSAR/QSPR.

En pratique, le développement d'un modèle QSAR/QSPR suit les étapes suivantes :

1. Collecte de données expérimentales fiables et en nombre le plus important possible.
2. Développement d'une série de descripteurs qui caractérisent les structures moléculaires électroniques et géométriques des composés de la base de données en vue de les relier à la propriété expérimentale étudiée.
3. Diviser cette base de données, aléatoirement, selon l'ordre de réponses, ou par choix orienté selon l'algorithme Kennard et Stone (CADEX) (Kennard et Stone, 1969) (le choix utilisé dans cette thèse), en un ensemble d'apprentissage (training set) qui contient généralement les 2/3 de la base de données et un ensemble de test (test set) constitué par le 1/3 restant.
4. Etablir des modèles mathématiques en utilisant l'ensemble d'apprentissage.
5. Caractériser les modèles élaborés par leurs indices de validation internes et vérifier leur robustesse par un test de randomisation de la variable dépendante Y (réponse).
6. Valider les modèles élaborés en utilisant l'ensemble de test et calculer leurs paramètres statistiques de validation externe.
7. Elaborer le domaine d'applicabilité du modèle retenu.

2. Les bases de données (Goulon-Sigwalt-Abram, 2008)

Un modèle QSPR, de par sa construction, est très dépendant des données expérimentales de référence; le choix de la base de données est donc un point critique de son développement. Dans la plupart des cas, les données expérimentales sont issues de la littérature.

Pour être de qualité, une base de données doit être composée de données expérimentales aussi fiables que possible, puisque les barres d'erreurs sur celles-ci se propageront dans le modèle final, étant donné que les paramètres de ce dernier sont ajustés par rapport à ces données. Il est donc important de choisir des données présentant des incertitudes faibles afin de limiter les barres des erreurs expérimentales. En effet, un modèle ne pourra être plus robuste statistiquement que les données théoriques à partir desquelles il a été développé.

De plus, les données doivent être obtenues suivant un protocole expérimental unique. En effet, les conditions expérimentales ont, en général, une forte influence sur les valeurs

obtenues. La définition de la propriété en termes de conditions expérimentales est d'ailleurs un point important de la démarche.

D'autre part, les modèles QSPR étant plus performants en interpolation qu'en extrapolation (Fortuné, 2006), pour que le modèle soit lui-même applicable sur de larges gammes de valeurs, il est important de disposer d'une large gamme de valeurs expérimentales, représentative de l'éventail des valeurs possibles. Les outils d'analyse de données étant généralement adaptés au traitement de distributions normales de données, il est par conséquent préférable que la distribution des propriétés expérimentales employées soit la plus proche possible de la normalité, pour atteindre des modèles au meilleur pouvoir prédictif possible.

Les modèles QSPR sont généralement plus efficaces pour un jeu de molécules similaires. Pareillement, il est important de considérer une base de données dont les molécules se comportent de manière similaire vis-à-vis de la propriété ciblée. De même, afin de réduire la probabilité de rencontrer des molécules très différentes de la majorité des autres, qui peuvent peser significativement sur le modèle et accroître l'incertitude sur les prédictions (outliers) dans le jeu de données, plus la diversité structurale d'une base de données est importante et plus la taille de la base de données doit être importante.

Tous ces facteurs poussent à examiner les données collectées avant tout développement d'un modèle QSPR. Par conséquent, les molécules susceptibles d'être des outliers ou présentant potentiellement des erreurs expérimentales importantes doivent être considérées avec précaution, voire les retirer du jeu de données préalablement au développement des modèles.

3. Les Descripteurs Moléculaires

Afin d'exploiter au maximum les informations contenues dans les structures moléculaires, celles-ci sont traduites en une série de grandeurs (en général scalaires) qui quantifient leurs caractéristiques physico-chimiques et structurelles. Ces grandeurs sont appelées descripteurs.

Les descripteurs moléculaires 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 (Nieto-Draghi,1984; Sana et Leroy,1995)

4. Les types de descripteurs moléculaires

On dénombre aujourd'hui plus de 10 000 descripteurs moléculaires, qui quantifient des caractéristiques physico-chimiques ou structurelles de molécules. 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. Il existe cependant quelques descripteurs mesurés : il s'agit généralement de données expérimentales plus faciles à mesurer que la propriété ou l'activité à prédire (coefficient de partage eau-octanol (Hansh *et al.*, 1995) polarisabilité, ou potentiel d'ionisation). Les descripteurs moléculaires sont fréquemment classés par rapport à la dimensionnalité de la représentation moléculaire sur laquelle ils sont calculés: on parlera alors de descripteurs 0D, 1D, 2D, ou 3D (Dudek *et al.*,2006).

A. Les descripteurs 0D

Tous les descripteurs moléculaires pour lesquels aucune information sur la structure moléculaire et les connectivités atomiques n'est nécessaire appartiennent à la classe des descripteurs 0D. Les nombres d'atomes et de liaisons, ainsi que la somme ou la moyenne des propriétés atomiques sont typiques de cette classe de descripteurs. Ces descripteurs peuvent toujours être facilement calculés, interprétés naturellement, ne nécessitent pas d'optimisation de la structure moléculaire et sont indépendants de tout problème conformationelle. Ils montrent généralement une très forte dégénérescence, c'est-à-dire qu'ils ont des valeurs égales pour plusieurs molécules, telles que les isomères. Leur contenu informationnel est faible, mais ils peuvent néanmoins jouer un rôle important dans la modélisation de plusieurs propriétés physico-chimiques ou prendre part à des modèles plus complexes (Bonachera, 2011).

B. Les descripteurs 1D

Tous les descripteurs moléculaires qui peuvent être calculés à partir des informations de la sous-structure de la molécule appartiennent aux descripteurs 1D. Le comptage des groupes fonctionnels et des fragments de sous-structure, ainsi que des descripteurs centrés sur les atomes, sont les descripteurs 1D les plus connus. Ces descripteurs sont souvent présentés comme des empreintes digitales, c'est-à-dire un vecteur binaire où 1 indique la présence d'une sous-structure et 0 son absence. Un avantage pertinent dans la description des molécules par

les empreintes digitales est la possibilité d'effectuer des calculs rapides pour la similarité des molécules / problèmes de diversité. Comme les descripteurs 0D, ces descripteurs peuvent généralement être facilement calculés, sont naturellement interprétés, ne nécessitent pas une optimisation de la structure moléculaire et sont indépendants de tout problème conformationel. Ils montrent généralement une dégénérescence moyenne élevée et sont souvent très utiles dans la modélisation à la fois des propriétés physico-chimiques et biologiques (Nieto-Draghi, 1984; Todeschini *et al.*,2009).

C. Les descripteurs 2D

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

- **Les indices topologiques,** qui considèrent la structure du composé comme un graphe, les atomes étant les sommets et les liaisons les arêtes. De nombreux indices quantifiant la connectivité moléculaire ont été développés en se basant sur cette approche, comme par exemple l'indice de Wiener [20], qui compte le nombre total de liaisons dans les chemins les plus courts entre toutes les paires d'atomes (en excluant les hydrogènes). D'autres indices basés sur les chemins ont été développés (Schultz, 1989; Balaban, 1982; Randić, 1975).
- **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…

D. Les descripteurs 3D

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

- **Les descripteurs géométriques.** Les plus importants sont le volume moléculaire, la surface accessible au solvant, le moment principal d'inertie.
- **Les descripteurs électroniques.** Permettent de quantifier différents types d'interactions inter- et intramoléculaires, de grande influence sur l'activité biologique des molécules. Le calcul de la plupart de ces descripteurs nécessite la recherche de la géométrie pour laquelle l'énergie stérique est minimale, et fait souvent appel à la chimie quantique. Par exemple, les énergies de la plus haute orbitale moléculaire occupée (HOMO) et de la plus basse vacante (LUMO) (orbitales frontières) sont des descripteurs fréquemment sélectionnés. Le moment dipolaire, le potentiel d'ionisation, et différentes énergies relatives à la molécule sont d'autres paramètres importants.
- **Descripteurs spectroscopiques.** Les molécules peuvent être caractérisées par des mesures spectroscopiques, par exemple par leurs fonctions d'onde vibrationnelles. En effet, les vibrations d'une molécule dépendent de la masse des atomes et des forces d'interaction entre ceux-ci; ces vibrations fournissent donc des informations sur la structure de la molécule et sur sa conformation. 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) (Trevor *et al.*, 1998) est ainsi obtenu à partir des fréquences de vibration de chaque molécule. Les descripteurs de type MoRSE (shuur *et al.*, 1996) (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.

5. Méthodes D'analyse Des Données :

Pour élaborer un modèle QSPR/QSAR nous avons besoin d'une méthode d'analyse de données, cette méthode permet de quantifier la relation qui existe entre la Propriété/Activité et la structure (descripteurs). Il existe plusieurs méthodes pour construire un modèle et analyser les données statistiques de ce dernier, certaines sont linéaires telles que la régression linéaire multiple (MLR), la régression aux moindres carrés partiels (PLS), d'autres sont non linéaires comme les arbres de décisions, les réseaux de neurones…, ces méthodes sont disponibles dans des logiciels tels que, Excel stat, Minitab, Statistica, Molegro Data Modeller, SPSS, R, etc…

Les différentes méthodes qui seront présentées dans la partie suivante sont celles exploitées au cours de ce travail, pour développer des modèles avec des paramètres les plus pertinents, valider ces modèles (en interne et/ou en externe) et déterminer leurs domaines d'applicabilité.

6. Sélection Des Descripteurs

Lorsqu'une grande quantité de descripteurs est introduite, certains d'entre eux peuvent contenir des informations redondantes, entraînant un problème de colinéarité. Les paramètres employés doivent être, autant que possible, porteurs de sens et facilement interprétables d'un point de vue chimique. Les descripteurs sélectionnés seront donc d'autant plus pertinents qu'ils offrent une idée du mécanisme du phénomène étudié (Mannan, 2005). Au final, les modèles QSPR devraient être simples, transparents et compréhensibles d'un point de vue phénoménologique (Mannan, 2005). Dans une modélisation QSAR/QSPR, ce principe signifie que le modèle doit avoir le moins de paramètres possibles (principe de parcimonie) tout en traduisant au mieux l'information véhiculée par la propriété (Crawley, 2005).

Aussi, le processus de simplification doit-il être une partie intégrante des tests d'hypothèse. De manière générale, pour qu'une variable soit retenue, il faut que son retrait entraîne une décroissance significative de la performance du modèle. Bien entendu, en simplifiant le modèle, il faut être attentif à ne pas perdre des parts d'informations essentielles.

Finalement, le sens chimique des descripteurs utilisés doit, bien entendu, être pris en considération puisque, plus les descripteurs sont reliés chimiquement au phénomène, plus la probabilité de faire intervenir des descripteurs par le biais du hasard est réduite (Witten et Frank, 2005).

7. Validation et Interprétation d'un Modèle QSPR/QSAR

Le développement des ressources informatiques rapides et économiques permet de calculer un grand nombre de descripteurs en utilisant différents outils logiciels. En conséquence, on ne peut pas nier le risque de corrélations fortuites avec le nombre croissant de variables incluses dans le modèle QSAR/QSPR par rapport au nombre limité de composés généralement utilisés pour le développement du modèle (Topliss et Costello, 1972). D'autre part, en employant divers outils d'optimisation, il est possible d'obtenir des modèles qui peuvent bien satisfaire les données expérimentales, mais il reste toujours une chance de

surestimation (overfitting). L'ajustement des données n'exprime pas une bonne prédiction du modèle car le premier est un paramètre pour la qualité statistique du modèle. C'est la raison principale pour laquelle les outils de validation doivent être appliqués sur le modèle QSAR développé pour vérifier sa prédictivité pour de nouvelles molécules non testées. Un organigramme pour la méthode de développement d'un modèle QSAR/QSPR fiable avec les différentes méthodes de validation avec les mesures couramment utilisées est reproduit dans la figure. 7.

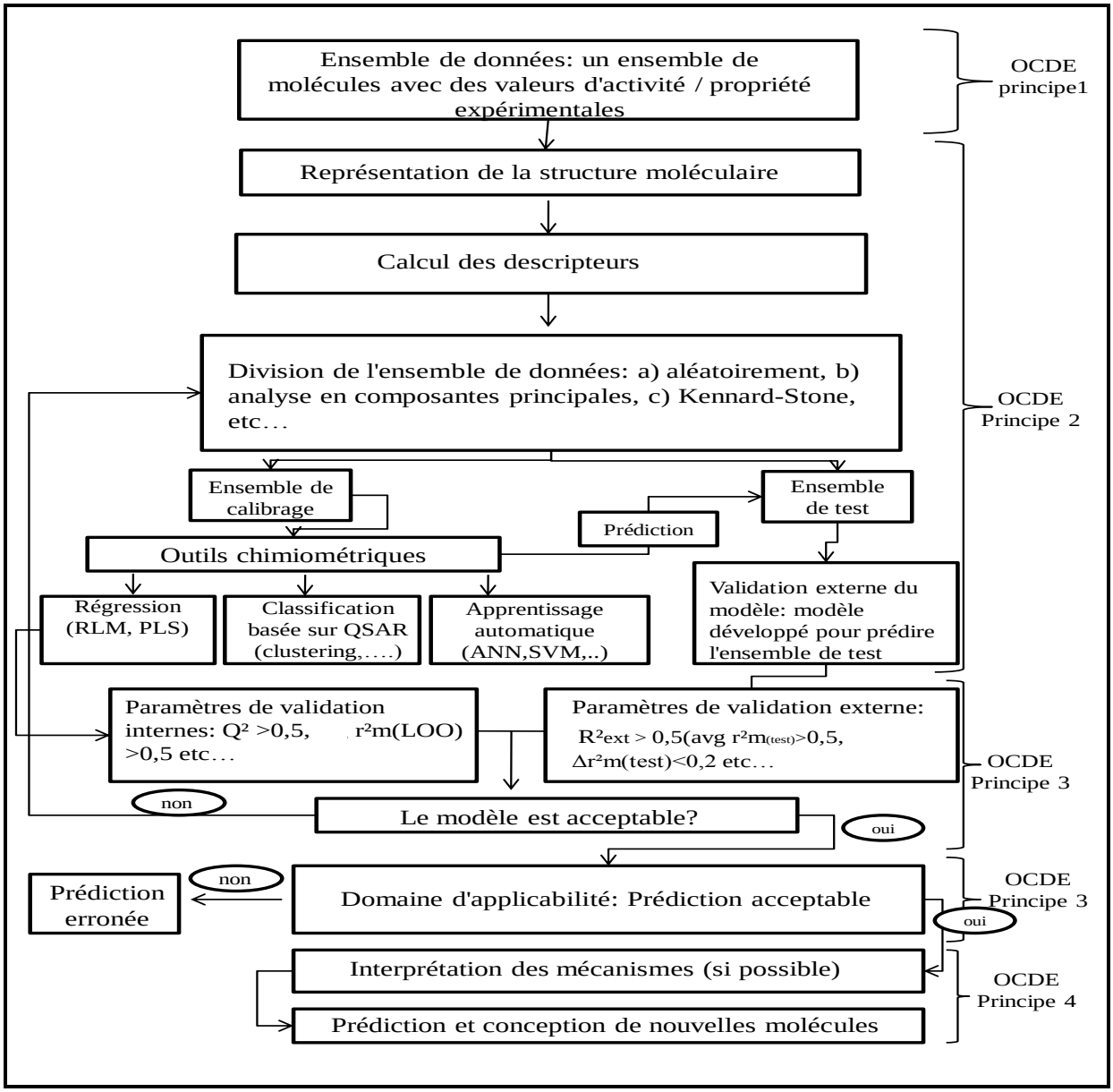


Figure.7 Étapes fondamentales pour la génération d'un modèle QSAR et méthodes de validation utilisées (Roy *et al.*, 2015).

7.1. Validation

7.1.1. Coefficients et tests statistiques standards

Afin de déterminer la qualité d'un modèle, différents indicateurs statistiques sont employés. Le plus répandu d'entre eux est le coefficient de corrélation R^2 qui évalue la part de la variance expliquée par le modèle; il est défini par la relation :

$$R^2 = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2} \quad (68)$$

Où $\bar{y}$ est la valeur moyenne des valeurs prédites, y_i et $\hat{y}_i$ sont respectivement les valeurs observées et estimée.

Plus la valeur de R^2 est proche de 1 (cas idéal) et plus les valeurs prédites et observées sont corrélées. L'erreur absolue moyenne (*MAE*, pour *mean absolute error*), est un autre indicateur utilisé défini par la relation (69);

$$MAE = \frac{\sum|\hat{y}_i - y_i|}{P} \quad (69)$$

La déviation standard ***s*** défini par la relation (70);

$$s = \sqrt{\frac{\sum(\hat{y}_i - y_i)^2}{P-n-1}} \quad (70)$$

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

$$F = \frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \hat{y}_i)^2} \frac{P-n-1}{n} \quad (71)$$

La pertinence des descripteurs dans le modèle, est également évaluée par le test-*t* de Student. Il s'agit de tester l'hypothèse considérant le descripteur comme non significatif. Pour une régression multi-linéaire, cela revient à supposer le coefficient ai qui lui est associé comme nul. Cette hypothèse est rejetée (avec un intervalle de confiance ***α***) si le ratio ti entre ai et son erreur type $\mathbf{s(a_i)}$ atteint la valeur du fractile d'ordre ***(1- α/2)*** de la loi de Student à ***(p-n-2)*** degrés de liberté.

$$|t_i| = \left|\frac{a_i}{s(a_i)}\right| > t_{1-\frac{\alpha}{2}}^{P-n-2} \quad (72)$$

7.1.2. Types de validation

A. Les Principes de l'OCDE

Les principes de l'OCDE sont les meilleurs éléments possibles des points essentiels à résoudre tout en développant des modèles QSAR fiables et reproductibles (Jaworska *et al.*, 2003). Les principes ont été formulés par des experts du QSAR lors d'une réunion tenue à Setúbal (Portugal) en mars 2002 comme lignes directrices pour la validation des modèles QSAR, notamment à des fins réglementaires. Les cinq directives adoptées par l'OCDE pour la validité du modèle QSAR ont été déjà citées (***partie histoire de QSAR***):

- ***Principe 1***— Définition précise de la propriété prédite,
- ***Principe 2***— Equation mathématique (ou algorithme) sans équivoque (reproductible),
- ***Principe 3***— Domaine d'applicabilité défini,
- ***Principe 4***— Mesures appropriées des performances,
- ***Principe 5***— Interprétation des mécanismes sous-jacents (si possible).

Le défi actuel dans le processus de développement d'un modèle QSAR/QSPR n'est plus au développement d'un modèle statistiquement solide pour prédire l'activité au sein de l'ensemble de calibrage, mais dans l'élaboration d'un modèle capable de prédire avec précision l'activité ou la propriété de nouveaux produits chimiques.

B. Validation interne

La validation interne d'un modèle QSAR/QSPR est effectuée en fonction des molécules utilisées dans le développement du modèle. Cela implique une prédiction d'activité des molécules étudiées puis une estimation des paramètres pour détecter la précision des prédictions. Pour juger de la qualité et de la qualité d'ajustement du modèle, la validation interne est une technique idéale. Mais, l'inconvénient majeur de cette approche est le manque de prévisibilité du modèle lorsqu'il est appliqué à un nouvel ensemble de données (Wold, 1978).

C. Validation externe

On ne peut pas juger la capacité prédictive du modèle développé à partir de la validation interne pour tout nouvel ensemble de composés. La validation externe doit être

effectuée dans le cadre d'une prévision des composés issus d'un ensemble n'ayant pas été utilisé dans l'élaboration du modèle. La validation externe assure la prédictivité et l'applicabilité du modèle QSPR développé pour la prédiction des molécules non testées (Roy, 2007).

D. Choix de l'ensemble de calibrage et de validation

Avant de commencer le développement des modèles, nous cherchons à diviser les données en deux sous-ensembles : un pour le calibrage et un pour la validation externe du modèle. Le sous-ensemble de calibrage doit être représentatif des données initiales et le sous-ensemble de la validation doit être choisis pour évaluer la qualité du modèle. Il existe de nombreux algorithmes de sélection de ses deux sous-ensembles qui se différencient principalement par leurs techniques de base. Nous citerons les plus utilisées par les praticiens du QSAR (Roy, 2007):

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

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

- ***Basé sur la réponse X***: Les propriétés et la similarité structurelle des molécules sont considérées pour le groupement de composés similaires. Ensuite, une fraction pré-décidée des composés est affectée au calibrage ou au jeu de tests manuellement ou de façon régulière.

Parmi les outils les plus couramment utilisés pour la division rationnelle des ensembles de données ont peut citer:

• K-Means clustering (Hartigan et Wong, 1979),

• La sélection par carte auto-organisée de Kohonen (Kohonen, 1984),

• La conception moléculaire statistique (Linusson *et al.*, 2000),

• Choix de Kennard-Stone (CADEX) (Kennard et Stone, 1969),

• Les sphères d'exclusion (Hudson *et al.*, 1996) et

• Sélection du jeu de test orientée vers l'extrapolation (Szàntai-kis *et al.*, 2003).

E. Domaine d'application

Même les modèles les plus exhaustifs, dignes de confiance et validés, ne peuvent prédire des propriétés de manière fiable pour l'intégralité des composés chimiques existants. Le domaine d'application (DA) permet de définir la zone dans laquelle un composé pourra être prédit avec confiance. Le DA correspond donc à la région de l'espace chimique incluant les composés du jeu d'apprentissage et les composés similaires, proches dans ce même espace (Netveza *et al.,* 2005). En effet, un modèle QSAR/QSPR n'est pas destiné à être employé en dehors de son domaine d'application, c'est-à-dire en dehors de l'espace chimique couvert par son jeu d'entraînement. La détermination des DA est donc d'une grande importance. Cette partie de l'analyse est d'ailleurs explicitement demandée dans les démarches de validation mises en place au niveau de l'OCDE [OCDE, 2009; Tunkel *et al.,* 2005).

Il existe plusieurs méthodes pour la détermination du domaine d'application d'un modèle QSPR/QSAR parmi ces méthodes on trouve la méthode du **''levier''**; cette méthode est basée sur la variation des résidus standardisés de la variable dépendante en fonction des leviers (la distance entre les valeurs des descripteurs et leur barycentre). Si un composé a un levier qui dépasse le seuil ***h*=3p/n*** (ou p est le nombre de descripteurs plus 1 et ***n*** le nombre d'observations (***training set***)), ce composé est considéré comme un composé influent sur modèle élaboré.

1.3. Métriques de validation pour les modèles de régression (QSAR/QSPR)

1.3.1. Métriques pour la validation interne

Les paramètres internes les plus couramment utilisés (Roy and Mitra, 2011) sont rappelés et discutés ci-après:

A. Validation croisée LOO (leave- one-out)

Pour déterminer la validation croisée LOO, l'ensemble de calibrage est principalement modifié en éliminant un composé de l'ensemble. Le modèle QSAR est ensuite reconstruit en fonction des molécules restantes de l'ensemble de calibrage en utilisant les descripteurs choisis, et l'activité du composé supprimé est calculée en fonction de l'équation QSPR résultante. Ce cycle est répété jusqu'à ce que toutes les molécules de l'ensemble de calibrage aient été supprimées une fois, et les données de la propriété prévue pour tous les composés de

calibrage sont utilisées pour le calcul de divers paramètres de validation internes. Enfin, la prédiction du modèle est jugée à l'aide de la somme des carrés des erreurs de prédiction (PRESS) et du Q^2 pour le modèle tandis que la valeur de l'écart quadratique moyen de prédiction (SDEP) est calculée à partir du PRESS.

$$PRESS = \sum\left(Y_{obs} - Y_{pred}\right)^2 \quad (73)$$

$$SDEP = \sqrt{\frac{PRESS}{n}} \quad (74)$$

$$Q^2 = 1 - \frac{\sum\left(Y_{obs\,(train\ )} - Y_{pred\ (train\ )}\right)^2}{\sum\left(Y_{obs\,(train\ )} - \bar{Y}_{training\ }\right)^2} = 1 - \frac{PRESS}{\sum\left(Y_{obs\,(train\ )} - \bar{Y}_{training\ }\right)^2} \quad (75)$$

Dans les équations (73) et (75), Y_{obs} et Y_{pred} correspondent aux valeurs de la propriété observée et prédite par LOO, ***n*** est le nombre d'observations, $Y_{obs\ (\text{train})}$ est la propriété observée, $Y_{pred\ (\text{train})}$ est la propriété prédite des molécules du jeu d'entraînement basée sur la technique LOO. La valeur seuil de Q^2 est de 0,5.

B. Validation croisée LMO (Leave-Many-Out)

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

C. La métrique r_m^2 pour la validation interne

Une valeur acceptable du Q^2 n'indique pas obligatoirement que les données d'activité/ propriété prévues se situent dans une proximité étroite avec celles observées bien qu'il puisse exister une bonne corrélation globale entre les valeurs. Ainsi, pour éviter ce problème et pour mieux indiquer la prévisibilité du modèle on peut exploiter les métriques r_m^2 introduites par Roy *et al*. (2012) elles sont définies par les équations suivantes:

$$\overline{r_m^2} = \frac{\left(r_m^2 + r_m'^2\right)}{2} \quad (76)$$

$$\Delta r_m^2 = \left|r_m^2 - r_m'^2\right| \quad (77)$$

Où $r_m^2 = r^2 \times \left(1 - \sqrt{\left(r^2 - r_0^2\right)}\right)$ et $r_m'^2 = r^2 \times \left(1 - \sqrt{\left(r^2 - r_0'^2\right)}\right)$ (78)

Les paramètres r^2 et r_0^2 sont les carrés des coefficients de corrélation entre les valeurs observées et les valeurs prédites par (LOO) des composés avec et sans interception, respectivement. Le paramètre $r_0'^2$ porte la même signification mais utilise les axes croisés.

Le $\overline{r_m^2}$ est la valeur moyenne de r_m^2 et $r_m'^2$, et Δr_m^2 est la différence absolue entre r_m^2 et $r_m'^2$. En cas de validation interne de l'ensemble de formation, les paramètres $\overline{r_m^2}_{(LOO)}$ et Δr_m^2 (LOO) peuvent être utilisés, la valeur de Δr_m^2 (LOO) devrait être inférieure à 0,2 à condition que la valeur de $\overline{r_m^2}_{(LOO)}$ soit supérieure à 0,5. Roy *et al* (2013) ont proposé que le calcul des métriques de r_m^2 soit basé sur les valeurs des données de réponse observées et les données de réponse prévues.

D. Test de randomisation

Afin de s'assurer qu'un modèle QSPR est fiable, les tests de Y-randomisation (Tropsha *et al.,* 2003) sont une des techniques les plus employées. En effet, il n'est pas rare d'obtenir des corrélations fortuites (ou « chance correlation »), c'est-à-dire un modèle affichant de bons résultats statistiques (R^2, Q^2) pour l'apprentissage, mais impliquant des descripteurs qui dans la réalité ne sont pas reliés à la propriété modélisée. Ces modèles aléatoires peuvent être détectés par la procédure Y-randomisation. Elle consiste à mélanger aléatoirement les propriétés expérimentales pour le jeu d'apprentissage et, en utilisant les mêmes descripteurs (figure 8), à entraîner à nouveau l'algorithme d'apprentissage pour tenter d'obtenir un modèle. Normalement, les modèles obtenus doivent avoir des performances très faibles. La distribution des modèles obtenus permet de fixer un seuil heuristique de signification des modèles.

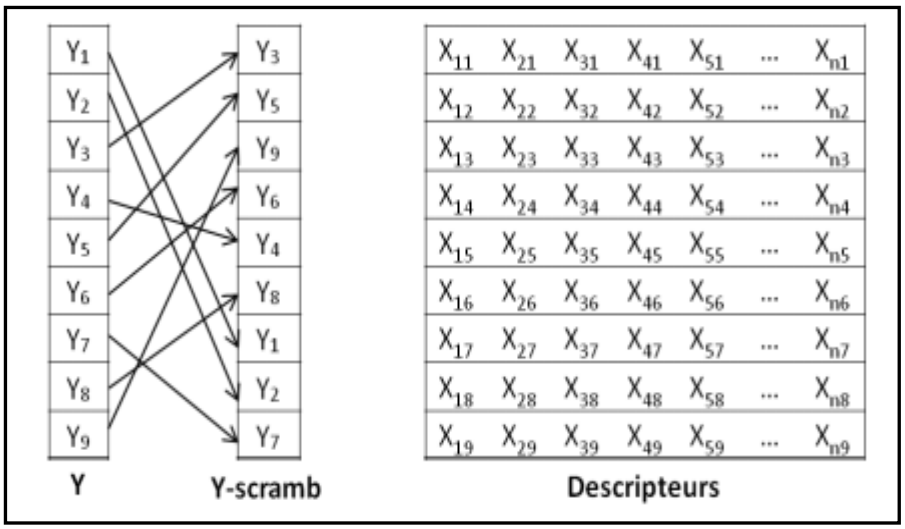


Figure.8 Illustration de la méthode « Y-scrambling »(randomisation de Y)

1.3.2. Paramètres de la validation externe

A. R² prédictif (R^2_{ext})

Le paramètre R^2_{ext} reflète le degré de corrélation entre les données de la propriété observées et prévues de l'ensemble de test.

$$R^2_{ext} = 1 - \frac{\sum_{i=1}^{n_{ext}} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n_{ext}} (y_i - \bar{y})^2} \quad (79)$$

Ici, Y_i et $\hat{Y}_i$ sont les données de la propriété observée et prédite pour les composés de test, tandis que $\bar{y}$ est la moyenne des valeurs expérimentales de l'ensemble de données entier. Ainsi, les modèles avec des valeurs de R^2_{ext} supérieures à la valeur stipulée de 0,5 sont considérés comme bien prédictifs.

B. Critères de Golbraikh et Tropsha

Golbraikh et Tropsha (2002) ont proposé un ensemble de paramètres pour déterminer la prédiction externe du modèle QSAR. Selon ces auteurs, les modèles sont considérés comme satisfaisants, si l'ensemble des conditions suivantes sont simultanément réalisées:

$$Q^2_{training} > 0.5 \quad (80\ a)$$

$$R^2_{test} > 0.6 \quad (80\ b)$$

$$\frac{r^2-r_0^2}{r^2} < 0.1 \; et \; 0.85 \leq k \leq 1.15 \text{ ou}$$

$$\frac{r^2-r_0'^2}{r^2} < 0.1 \; et \; 0.85 \leq k' \leq 1.15 \quad (80\text{ c})$$

$$\left|\boldsymbol{r_0^2 - r_0'^2}\right| < 0.3 \quad (80\text{ d})$$

C. La métrique r²m (test) pour la validation externe

Afin de vérifier la proximité entre les données observées et prédites, le paramètre r_m^2 (test), similaire à $r_{m(\mathrm{LOO})}^2$ utilisé dans la validation interne, a été développé par Roy *et al*. (2012). La valeur de $r_{m(\mathrm{test})}^2$ est calculée en utilisant les coefficients de corrélation carrés entre la propriété observée et la prédiction des composés testés. Pour une prévision acceptable, la valeur de Δr_m^2 (test) devrait de préférence être inférieure à 0,2 à condition que la valeur du test r²m soit supérieure à 0,5.

D. Erreur Quadratique Moyenne de Prédiction ($RMSE_P$)

La capacité prédictive externe d'un modèle QSAR/QSPR peut aussi être déterminée par l'erreur quadratique moyenne dans la prédiction ($RMSE_P$) donnée par l'équation.

$$RMSEp = \sqrt{\frac{\sum(y_{obs\ (test)} - y_{pred\ (test)})^2}{n_{ext}}} \quad (81)$$

n_{ext} désignant le nombre de composés de test (prédiction).

E. Q²F1, Q²F2, Q²F3 et CCC

Une mesure de validation externe (Golbraikh et al., 2002, 2003 ; Tropsha *et al.*, 2003) Q_{F1}^2 proposée est définie comme suit :

$$Q_{F1}^2 = 1 - \sum_{i=1}^{n_{ext}} (y_i - \hat{y}_i)^2 / \sum_{i=1}^{n_{ext}} (y_i - \bar{y}_{Tr})^2 \quad (82)$$

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

Après la formulation Q_{F1}^2 et la méthode proposée par Golbraikh et Tropsha, Schüürmann *et al* (2008), ont proposé un autre critère noté Q_{F2}^2, calculé selon :

$$Q_{F2}^2 = 1 - [\sum_{i=1}^{n_{ext}} (y_i - \hat{y}_i)^2 / \sum_{i=1}^{n_{ext}} (y_i - \bar{y}_{ext})^2] \quad (83)$$

Q^2_{F2} est différent de Q^2_{F1} car la valeur moyenne au dénominateur est calculée à l'aide de l'ensemble de données de prédiction et non de calibrage.

La méthode ne tient pas compte de la "distance" par rapport à la moyenne des valeurs de calibrage, ce qui la rend indépendante de celle ci. Les opinions divergent (Schüürmann *et al.*,2008; Consoni *et al.*,2009, 2010) quant à savoir s'il s'agit ou non d'un avantage, mais, dans le cas où l'ensemble de calibrage original utilisé pour construire le modèle n'est pas disponible pour l'utilisateur, Q^2_{F2} est un avantage. Un autre point, mis en évidence par Schüürmann (2008), concerne le fait que $Q^2_{F2} \leq Q^{2}_{F1}$, donc Q^{2}_{F1}fournit généralement des jugements plus optimistes et, par conséquent, le risque que les modèles qui ne sont pas en mesure de faire de bonnes prédictions soient acceptés est plus élevé. En outre, Q²F1, même s'il est rare, pourrait être même supérieur à R², ce qui conduit à la conclusion contrastée selon laquelle le modèle est capable de mieux prédire de nouvelles données que celles disponibles.

Par la suite, Consonni *et al.*, (2009, 2010) ont proposé une nouvelle mesure de validation externe Q^2_{F3}, en comparant et en mettant en évidence les différences avec Q^2_{F2}. Un commentaire préliminaire concerne la relation $Q^2_{F2} \leq Q^2_{F1}$, est discuté par Schüürmann (2008), qui, bien que vraie, n'implique pas nécessairement que Q^2_{F2}soit un moyen correct d'estimer la capacité d'un modèle à prédire. En outre, le fait qu'il n'y a pas de référence à l'ensemble de calibrage du modèle (c'est-à-dire $\bar{y}_{TR}$), est un inconvénient, d'après Consonni *et al.*, ce qui est contraire à l'opinion de Schüürmann. De plus, dans les analyses de Consonni, Q^2_{F1} et Q^2_{F2} se sont révélés biaisés en fonction de la distribution des données.

Un autre critère de validation Q^2_{F3}, alternatif à Q^2_{F1} et Q^2_{F2}, a été proposé ((Schüürmann *et al.*, 2008):

$$Q^2_{F3} = 1 - [(\sum_{i=1}^{n_{ext}} (y_i - \hat{y}_i)^2 / n_{ext})/(\sum_{i=1}^{n_{tr}} (y_i - \bar{y}_{tr})^2 / n_{tr})] \quad (84)$$

Comme on peut le voir, la formule diffère de Q^2_{F1} et Q^2_{F2} car le dénominateur est calculé sur l'ensemble d'entraînement, et le numérateur et le dénominateur sont divisés par le nombre d'éléments correspondants.

De l'analyse précédente, il est évident que deux types de critères de validation ont été proposés: ceux basés sur les variations de la forme Q^2 (Schüürmann *et al.*,2008; Consoni *et al.*,2009, 2010; Shi *et al.*, 2001)et ceux basés sur la différence entre les données expérimentales et les données prédites (par le modèle) de l'ensemble de prédiction

(Roy,2007; Mitra *et al.*,2010; Golbraikh et Tropsha,2002; Roy *et al.*, 2008,2009; Ojha *et al.*, 2011).

Le coefficient de corrélation de concordance (CCC) proposé par Lin (Lin, 1989, 1992) légèrement réarrangé par rapport à l'original pour une lisibilité plus facile, parce qu' il est bien ajusté pour mesurer l'accord entre les données expérimentales et prédites, ce qui devrait être le véritable objectif de toute prédiction des modèles QSAR/QSPR:

$$CCC = 2\sum_{i=1}^{n_{ext}}(y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}}) / \sum_{i=1}^{n_{ext}}(y_i - \bar{y}_{ext})^2 + \sum_{i=1}^{n_{ext}}(\hat{y}_i - \bar{\hat{y}})^2 + n_{ext}(\bar{y}_{ext} - \bar{\hat{y}})^2 \quad (85)$$

Ce coefficient mesure à la fois la précision (à quelle distance des observations) et la justesse (c'est-à-dire à quel point la ligne de la régression dévie de la droite x = y dite « concordance line »). Il s'agit d'une validation externe car aucune information du jeu d'entraînement n'est nécessaire. Par conséquent, pour toute divergence de la droite de régression de la ligne de concordance donne en conséquence une valeur de CCC inférieure à 1. le point clé est que ce résultat est obtenu même si le coefficient de corrélation de Pearson est égale à 1, c'est-à-dire que les données correspondent parfaitement à n'importe quelle relation linéaire; cependant, dans ce dernier cas, le modèle peut être précis mais pas exacte.

2. Interprétation des modèles

L'interprétation des modèles est un point important (Staton, 2003; Katritzky *et al.*, 2001) recommandé dans la démarche de validation des modèles pour la prédiction dans un cadre règlementaire (Netveza *et al.*, 2005). Outre l'aspect d'interprétation pure des modèles, qui peut permettre une meilleure compréhension des phénomènes chimiques mis en jeu, l'utilisation de paramètres interprétables dans les modèles prédictifs présente l'intérêt de limiter les risques d'avoir choisi ces derniers par chance.

Les descripteurs ne sont pas forcément faciles à interpréter chimiquement, par exemple les indices topologiques, notamment, sont des constructions mathématiques caractérisant la taille et la forme des systèmes moléculaires mais sans caractérisation explicite. Si leur efficacité en termes de prédiction n'est plus à démontrer, une interprétation physico-chimique est en général très difficile.

Un autre obstacle à l'interprétation peut provenir du type de modèle choisi. Un grand nombre de descripteurs, par exemple, rend une équation difficilement interprétable du fait

d'un trop grand nombre d'informations. De même, certains modèles non linéaires rendent l'interprétation de descripteurs, pourtant significatifs chimiquement, totalement impossible.

En termes de démarche, l'interprétation peut être prise en compte dès l'étape de sélection des données. En effet, au cours du processus, il peut être nécessaire de choisir entre deux descripteurs très proches statistiquement. Une technique automatisée peut, par exemple, mener au choix du moins significatif du point de vue chimique pour peu que sa corrélation avec la propriété expérimentale soit très légèrement supérieure. Inclure une phase de choix manuel des descripteurs peut alors permettre d'intégrer plus aisément des considérations physico-chimiques.

Conclusion

Dans cette partie, le processus de construction d'un modèle QSAR/QSPR a été détaillé. La technique de modélisation QSAR / QSPR implique l'utilisation d'un nombre significatif d'outils statistiques. Les modèles QSAR/QSPR développés peuvent fournir des relations linéaires et non linéaires entre la réponse et les attributs chimiques grâce à des analyses basées sur la régression et la classification. Étant donné que les relations mathématiques quantitatives sont établies, la validation des modèles utilisant un algorithme statistique approprié devient essentielle pour confirmer la stabilité et la bonne capacité prédictive des modèles. Le jugement pour le choix de la méthode dépend d'une multitude de facteurs, y compris la réponse à modéliser, la nature des données du jeu d'entraînement, le type et le nombre de descripteurs utilisés et même l'objectif de l'analyse.

Références bibliographiques

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

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

Consonni V, Ballabio D, Todeschini R, Comments on the Definition of theQ² Parameter for QSAR Validation. J. Chem. Inf.Model. 2009. 49, 1669.

Consonni V, Ballabio D, Todeschini R, Evaluation of model predictive ability by external validation techniques. J. Chemom. 2010. 24, 194.

Crawley M J, Statistics : an introduction using R, Wiley, Chichester, UK, 2005.

Dudek A Z, Arodz T, Gàlvez J. Computational methods in developing quantitative structure-activity relationships (QSAR): A review. Comb. Chem. High. T. Scr, 2006. 9, 213.

Duprat A F, Huynh T, Dreyfus G, Toward a principled methodology for neural network design and performance evaluation in QSAR. Application to the prediction of LogP. J. Chem. Inf. Comput. Sci. 1998. 38, 586.

Fortuné A, Techniques de Modélisation Moléculaire appliquées à l'étude et à l'optimisation de Molécules Immunogènes et de Modulateurs de la Chimiorésistance, Thèse de Doctorat, Université Joseph Fourier – Grenoble I, France, 2006

Gasteiger J, Zupan J, Angew. Neural networks in chemistry. Chem. Int. Ed. Engl. 1993.32, 503.

Geladi P, Kowalski B R, Partial least-squares regression: a tutorial. Anal. Chim. Acta. 1986. 185, 1.

Ghasemi J, Saaidpour S, Brown S D, QSPR study for estimation of acidity constants of some aromatic acids derivatives using multiple linear regression (MLR) analysis. J. Mol. Struct. (Theochem). 2007. 805, 27.

Golbraikh A, Tropsha A, Beware of q2! J Mol Graph Model. 2002. 20,269.

Goulon-Sigwalt-Abram A, Une nouvelle méthode d'apprentissage de données structurées : applications à l'aide à la découverte de médicaments [thèse en ligne]. PhD thesis, Paris : Université Pierre et Marie Curie (Paris 6), 2008. 1, 22- 25, 30-32

Hansch C, Leo A, Hoekmann D, Exploring QSAR: hydrophobic, electronic and stericconstants. American Chemical Society, Washington, DC. 1995. 348.

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

Hawkins D M, Basak S C, Mills D, Assessing model fit, by cross-validation. J. Chem. Inf. Comput Sci. 2003. 43,579.

Heritage T W, Ferguson A M, Turner D B, Willett P, EVA : A novel theoretical descriptor for QSAR studies. Perspect Drug Discov. 1998. 9. 381.

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

Jaworska J S, Comber M, Auer C, Van Leeuwen C J, Summary of a workshop on regulatory acceptance of (Q)SARs for human health and environmental endpoints. Environ Health. Perspect. 2003. 111, 1358.

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

Kennard R W, Stone LA, Computer aided design of experiments. Technometrics. 1969. 11, 137.

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

Leardi R, Genetic algorithms in chemometrics and chemistry: a review. J. Chemometr. 2001. 15, 559.

Lin L I, A Concordance Correlation Coeffcient to Evaluate Reproducibility. Biometrics 1989, 45, 255.

Lin L I, Assay Validation Using the Concordance Correlation Coefficient. Biometrics 1992. 48, 599.

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

Mannan S, Lee's Loss Prevention in Process Industries: Hazard Identification, Assessment and Control, Elsevier Butterworth-Heinemann, Burlington, 2005.

Margossian N, Le règlement REACH-La règlementation européenne sur les produits chimiques, Dunod / L'usine Nouvelle, Paris, 2007.

Mitra I, Roy P P, Kar S, Ojha P, Roy K, On further application of r^2m as a metric for validation of QSAR models. J. Chemometrics. 2010. 24, 22.

Myles A J, Feudale R N, Liu Y, Woody N A, Brown S D, An introduction to decision tree modeling. J. Chemom. 2004. 18, 275.

Netzeva T I, Worth A P, Aldenberg T, Benigni R, Cronin M T D, Gramatica P, Jaworska J S, Kahn S, Klopman G, Marchant C A, Myatt G, Nikolova-Jeliazkova N, Patlewicz G Y, Perkins R, Roberts D W, Schultz T W, Stanton D T, Van de Sandt J J M, Tong W, Veith G, Yang, CECVAM WORKSHOP REPORT, Current status of methods for defining the applicability domain of (quantitative) structure-activity relationships, ATLA. Altern. Lab. Anim. 2005. 33, 155.

Nieto-Draghi C, A general guidebook for the theoretical prediction of physico-chemical properties of chemicals for regulatory purpose. Chemical review.2015.115, 13093

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

Ojha P K, Mitra I, Das R N, Roy K, Further exploring r^2m metrics for validation of QSPR models. Chemom. Intell. Lab. Syst. 2011. 107, 194.

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

Roy K, Chakraborty P, Mitra I, Ojha P K, Kar S, Das R N, Some case studies on application of "r²m" metrics for judging quality of QSAR predictions: emphasis on scaling of response data. J. Comput. Chem. 2013. 34, 1071.

Roy K, Kar S, Das R N, Statistical Methods in QSAR/QSPR. In: A Primer on QSAR/QSPR Modeling. SpringerBriefs in Molecular Science. Springer, Cham. 2015, 37.

Roy K, Mitra I, Kar S, Ojha P K, Das R N, Kabir H, Comparative studies on some metrics for external validation of QSPR models. J. Chem. Inf. Model. 2012. 52, 396.

Roy K, Mitra I, On various metrics used for validation of predictive QSAR models with applications in virtual screening and focused library design. Comb.Chem.High.Throughput. Screen. 2011. 14,450.

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

Roy P P, Roy K, On Some Aspects of Variable Selection for Partial Least Squares Regression Models. QSAR Comb. Sci. 2008. 27, 302.

Roy P P, Somnath P, Mitra I, Roy K, On two novel parameters for validation of predictive QSAR models. Molecules. 2009. 14, 1660.

Sana M, Leroy G, Graph theory, electronic structures and reaction mechanisms. J.Mol. Struct. (THEOCHEM). 1984. 109, 251

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

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

Schüürmann G, Ebert R, Chen J, Wang B, Keuhne, R. External Validation and Prediction Employing the Predictive Squared Correlation Coefficients Test Set ActivityMean vs Training Set Activity Mean. J. Chem. Inf. Model. 2008. 48, 2140.

Shi L M, Fang H, Tong W, Wu J, Perkins R, Blair R M, Branham W S, Dial S L, Moland C L, Sheehan D M, QSAR Models Using a Large Diverse Set of Estrogens. J. Chem. Inf. Comput. Sci. 2001. 41, 186.

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

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

Tetko I V, Villa A E P, Livingstone D J, Neural network studies. 2. Variable selection. J. Chem. Inf. Comput. Sci. 1996. 36, 794.

Todeschini R, Consonni V, Gramatica P, Chemometrics in QSAR. Edition: 4, Publisher: Elsevier. 2009.129.

Topliss J G, Costello R J, Chance correlation in structure-activity studies using multiple regression analysis. J. Med. Chem. 1972. 15,1066.

Tropsha A, Gramatica P, Gombar V K, The Importance of Being Earnest: Validation is the Absolute Essential for Successful Application and Interpretation of QSPR Models. QSAR. Comb. Sci. 2003. 22 69.

Tunkel J, Mayo K, Austin C, Hickerson A, Howard P, Practical considerations on the use of predictive models for regulatory purposes. Environ. Sci. Techn. 2005. 39, 2188.

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

Wold S, Cross-validation estimation of the number of components in factor and principal components models. Technometrics.1978. 20, 397.

PARTIE II

PROPRIÉTÉS ÉTUDIÉES ET MÉTHODOLOGIE

I. Propriétés étudiées

II. Méthodologie

I. LES PROPRIÉTÉS ÉTUDIÉES

I.1. La température d'ébullition (Teb)

La température d'ébullition est la température à laquelle les phases liquide et gazeuse d'une substance pure sont en équilibre à une pression donnée, c'est la température à laquelle la substance change d'état, du liquide au gaz à une pression donnée. Le point d'ébullition normal est le point d'ébullition à la pression atmosphérique normale ($1{,}013.10^5$ kPa). En termes d'interactions intermoléculaires, le point d'ébullition représente la température à laquelle les molécules possèdent l'énergie thermique suffisante pour surmonter les attractions intermoléculaires liant les molécules dans le liquide.

La température d'ébullition d'un composé pur augmente avec la taille, la ramification de la molécule, et avec la présence des liaisons hydrogènes et des interactions dipôle-dipôle.

La température d'ébullition est importante pour la caractérisation et l'identification du composé. Elle fournit également une indication de la volatilité d'un composé. D'autres propriétés physiques, telles que la température critique (Fisher, 1989), le point d'éclair (Satyanarayana et Kakati, 1991), et l'enthalpie de vaporisation (Rechsteiner, 1982), peuvent être prédits ou estimés à partir des points d'ébullition. Le besoin de données fiables pour l'optimisation des processus industriels, le développement des modèles QSPR fiables pour l'estimation des points d'ébullition normaux pour les composés qui ne sont pas encore synthétisés est devenu important.

De nombreuses méthodes ont été développées pour l'estimation des points d'ébullition normaux des composés, et de nombreuses corrélations QSPR ont été rapportées. Des tentatives préliminaires ont été faites pour corréler les points d'ébullition des hydrocarbures homologues avec le nombre d'atomes de carbone ou le poids moléculaire (Walker, 1894). Des méthodes ultérieures ont employé des paramètres physiques tels que le parachor et la réfractivité molaire (Meissner, 1949). Des méthodes pour l'estimation des points d'ébullition ont été résumées par Rechsteiner (Rechsteiner, 1982) et Horvath (Horvath, 1992). Des efforts ont été faits pour estimer les points d'ébullition par contribution de groupe additive (CGA) (Rechsteiner, 1982; Reid et Prausnitz, 1987) basée sur l'hypothèse que les forces de cohésion dans les liquides sont de courte portée (Benson et Buss, 1958) et procède de la division d'une molécule en groupes structuraux prédéfinis, dont chacun ajoute un incrément constant à la valeur de la propriété (Copeman *et al.*, 1988). Les méthodes de contribution de groupe

fournissent une bonne prédiction des points d'ébullition (Joback et Reid, 1987; Stein et Brown, 1994), avec une erreur absolue moyenne de 15,5 K, pour les petites molécules non polaires. Cependant, les méthodes ACG sont limitées aux molécules contenant des groupes présents dans l'ensemble de molécules d'étalonnage, et certains schémas de contribution de groupe ne sont pas suffisamment complets pour couvrir plusieurs substitutions de groupes fonctionnels.

Mis à part les simples corrélations des points d'ébullition avec le nombre d'atomes de carbone ou le poids moléculaire pour des séries homologues de composés, Wiener a été le premier à corréler les points d'ébullition avec des descripteurs topologiques (Wiener, 1947). Wiener a introduit deux paramètres structurels, appelé l'indice de Wiener (W), défini comme la somme des distances entre deux atomes de carbone dans la molécule (Wiener, 1947), et l'indice de polarité de Wiener (P), défini comme le nombre de paires non ordonnées de sommets dont la distance entre deux sommets est égale à 3. Sur la base de ces indices, il a prédit les points d'ébullition des paraffines avec une erreur moyenne de 1°C (Wiener, 1947). D'autres indices topologiques, y compris les indices de connectivité moléculaire de Randić (Randić1975), et de Kier & Hall (1976), ont permis de corréler les points d'ébullition des alcanes et des amines (Kier et Hall, 1976). Pendant plus de quatre décennies, la corrélation des points d'ébullition des hydrocarbures avec la structure chimique a suscité un intérêt considérable. Cependant, pour une meilleure prévisibilité d'une propriété sous la forme d'un modèle général, la recherche de meilleurs descripteurs a été un point focal de la recherche QSPR.

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

Références bibliographiques

Benson S W, Buss J H. Additivity rules for the estimation of molecular properties. Thermodynamic properties. J. Chem. Phys. 1958. 29, 546.

Copeman T W, Mathias P M, Klotz H C. In Physical Property Prediction in Organic Chemistry; Jochum, C., Hicks, M. G., Sunkel, J., Eds.; Springer-Verlag: New York, 1988; pp 351.

Fisher C H. Boiling-point gives critical-temperature. Chem. Eng. 1989. 96, 157.

Horvath A L. Molecular Design: Chemical Structure Generation from the Properties of Pure Organic Compounds; Elseiver: Amster-dam, 1992; Chapter 2.

Joback K G, Reid R C. Estimation of pure-component properties from group-contributions. Chem. Eng. Commun. 1987. 57, 233.

Kier L B, Hall L H. Molecular Connectivity in Chemistry and Drug Research; Academic: New York, 1976.

Meissner H P. Critical constants from parachor and molar refraction. Chem. Eng. Progr. 1949. 45, 149.

Randić M. Characterization of molecular branching. J. Am. Chem. Soc. 1975, 97, 6609.

Rechsteiner C E. In Handbook of Chemical Property Estimation Methods; Lyman, W. J., Reehl, W. F., Rosenblatt, D. H., Eds.; McGraw-Hill: New York, 1982; Chapter 12.

Reid R C, Prausnitz, J. M.; Poling, B. E. The Properties of Gases and Liquids, 4th ed.; McGraw-Hill: New York, 1987.

Satyanarayana K, Kakati M C. Note: Correlation of flash points. Fire. Mater. 1991. 15, 97.

Stein S E, Brown R L. Estimation of normal boiling points from group contributions. J. Chem. Inf. Comput. Sci. 1994. 34, 581.

Walker J. The boiling points of homologous compounds. Part I. Simple and mixed ethers .J. Chem. Soc. 1894. 65, 193.

Wiener H. Influence of interatomic forces on paraffin properties, J. Am. Chem. Soc. 1947. 69, 17.

I.2. La température de fusion (T_{fus})

La température de fusion est une propriété physique fondamentale spécifiant la température de transition de l'état (phase) solide à l'état liquide. Elle a été utilisée comme critère de pureté d'un composé et pour prédire d'autres propriétés physiques telles que la solubilité aqueuse (Yalkowsky et Banerjee, 1992; Ran et Yalkowsky, 2001) et la viscosité des liquides (Benoit-Guyod, 1984). Le point de fusion a également été utilisé avec succès comme descripteur dans les corrélations avec la solubilité dans l'eau des chlorophénols (Devillers et Bull, 1986; Barratt, 1995) et la corrosivité des acides organiques, des bases et des phénols (Dearden, 1991). Puisque le point de fusion affecte la solubilité d'un composé, les techniques d'estimation du point de fusion des composés organiques aideraient de manière significative les chimistes à concevoir de nouveaux médicaments avec une gamme spécifiée de point de fusion et de solubilité. La température de fusion affecte la solubilité, et la solubilité contrôle la toxicité, donc si un composé est faiblement soluble, sa concentration dans l'environnement aqueux peut être trop faible pour qu'il exerce un effet toxique (Dearden, 1999).

En général les températures de fusion des molécules organiques dépendent de l'agencement des atomes dans le réseau cristallin ainsi que de la force des interactions de groupes par paires (Dearden, 2003; Kataigorodsky, 1973). Le point de fusion est déterminé par la force du réseau cristallin qui, à son tour, est contrôlé principalement par trois facteurs: les forces intermoléculaires, la symétrie moléculaire et les degrés de liberté dans une molécule (Dearden, 1999). De plus, le mouvement moléculaire des cristaux affecte le point de fusion, qui dépend de la taille et de la forme des molécules, de leur orientation dans le cristal (Mackay *et al.,* 1982). De nombreux composés cristallisent sous plus d'une forme, avec une température de fusion différente, et présentent donc le phénomène de polymorphisme. Les transitions de phase sont compliquées par le polymorphisme; les molécules qui existent sous différentes formes cristallines ont des propriétés distinctes, y compris la capacité calorifique et la température de fusion. En outre, les mesures des points de fusion sont affectées par la pureté du composé et par l'erreur expérimentale. Malgré la grande quantité disponible de données sur la température de fusion, les corrélations et les prédictions des températures de fusion de divers ensembles de composés sont encore très difficiles. Diverses méthodes QSPR, telles que les relations quantitatives propriété/propriété (QRPP) (Simamora, 1993), et les contributions de groupes (Krzyzaniak *et al.,* 1995; Katritzky *et al*., 2001) ont été utilisées pour prédire le point de fusion.

Cependant, les méthodes de contribution de groupe ont généralement des difficultés à fournir des estimations fiables des points de fusion, car elles dépendent fortement des caractéristiques structurelles non additives, telles que les interactions intermoléculaires et la symétrie moléculaire (Katritzky *et al*., 2001). Une revue complète sur la relation entre les points de fusion et les structures chimiques est présentée dans (Needham *et al*., 1988).

Références bibliographiques

Barratt M D, Quantitative structure activity relationships for skin corrosivity of organic acids, bases and phenols. Toxicol. Lett. 1995. 75, 169.

Benoit-Guyod J L, Andre C, Taillandier G, Rochat J, Boucherle A, Toxicity and QSAR of chlorophenols on Lebistes reticulatus. Ecotoxicol. Environ. Saf. 1984. 8, 227.

Dearden J C, In Advances in Quantitative Structure-Property Relationships; Charton, M., Charton, B. I., Eds.; JAI Press Inc.: Stamford, 1999; Vol. 2

Dearden J C, Quantitative structure-property relationships for prediction of boiling point, vapor pressure, and melting point. Environ. Toxicol. Chem. 2003. 22, 1696.

Dearden J C, The QSAR prediction of melting point, a property of environmental relevance. Sci. Total Environ. 1991. 109, 59.

Devillers J, Chambon P, Acute toxicity and QSAR of chlorophenols onDaphnia magna. Bull. Environ. Contam. Toxicol. 1986. 37, 599.

Katritzky A R, Jain R, Lomaka A, Petrukhin R, Maran U, Karelson M, Perspective on the relationship between melting points and chemical structure. Cryst. Growth Des. 2001. 1, 261.

Kitaigorodsky A I, In Molecular Crystals and Molecules; E. M. Loebel, Ed.; Academic Press: New York, 1973.

Krzyzaniak J F, Myrdal P B, Simamora P, Yalkowsky S H, Boiling point and melting point prediction for aliphatic, non-hydrogen-bonding compounds. Ind. Eng. Chem. Res. 1995. 34, 2530.

Mackay D, Shiu W T, Bobra A, Billington J, Chan E, Yeun A, C Ng, Szeto F, Volatilization of Organic Pollutants from Water. U. S. Environmental Agency Report PB 82-230939; U. S. Environmental Agency: Athens, GA, 1982.

Needham D E, Wei I C, Seybold P G, Molecular modeling of the physical properties of alkanes. J. Am. Chem. Soc. 1988. 110, 4186.

Ran Y, Yalkowsky S H, Prediction of drug solubility by the general solubility equation (GSE) J. Chem. Inf. Comput. Sci. 2001. 41, 354.

Simamora P, Miller A H, Yalkowsky S H, Melting point and normal boiling point correlations: applications to rigid aromatic compounds. J. Chem. Inf. Comput. Sci. 1993. 33, 437.

Yalkowsky S H, Banerjee S, Aqueous Solubility. Methods of Estimation for Organic Compounds; Marcel Dekker: New York, 1992.

I.3. L'indice de rétention chromatographique (IR)

La chromatographie en phase gazeuse (CPG) est l'une des techniques analytiques les plus utilisées en raison de sa simplicité, de sa rapidité d'analyse, de la haute sensibilité des systèmes détecteurs et de l'efficacité des séparations. Ainsi, la CPG a trouvé une large application dans le domaine pharmaceutique, les études environnementales, les industries pétrolières, la chimie clinique, l'analyse des pesticides, les conservateurs alimentaires, etc...(Grob, 1985). L'identification d'un composé est souvent réalisée sur la base de comparaisons des pics avec un standard authentique du composé suspecté.

Le processus de séparation chromatographique résulte, principalement, des interactions entre les molécules de solutés et celles de la phase stationnaire ; les interactions dans la phase gazeuse sont relativement moins importantes, ce qui ne signifie pas que le gaz porteur est un milieu absolument inerte.

I.3.1. Interaction soluté – phase stationnaire (Messadi-Lourici, 2004)

L'interaction d'un soluté avec une phase stationnaire liquide est très complexe, et reste déterminée par différentes forces qui opèrent dans la solution, et qui découlent de la constitution chimique des molécules. La forme et la taille de celles-ci, quoique de peu d'influence en général, peuvent parfois devenir significatives.

Il existe trois forces attractives qui interviennent lorsque la distance entre les molécules est grande, comparativement à leurs tailles

1. Les forces de dispersion de LONDON

La variation très rapide des dipôles formés par les noyaux et les électrons, lors du déplacement des molécules, produit un champ électrique. Les forces de dispersion proviennent de l'action de ce champ électrique sur la polarisation des autres molécules, avec pour conséquence l'apparition de dipôles induits en phase avec les dipôles instantanés qui leur ont donné naissance. Ces forces sont le mieux décrites par la mécanique quantique (Hischfelder, 1964; Mc Lachlan, 1965). Pratiquement, l'énergie de l'interaction de dispersion est donnée par la formule de London (London, 1963):

$$E_L = -\left(\frac{3}{2}\right)\frac{\alpha_1\alpha_2 I_1 I_2}{r^6\left(I_1 I_2\right)} \qquad (86)$$

Les indices 1 et 2 se rapportent aux deux atomes en cohérence, de polarisabilité α et de potentiel d'ionisation I, dont les centres sont distants de r.

Quoique l'équation de London ne s'applique, en toute rigueur, qu'aux atomes, elle peut être utilisée pour les molécules en additionnant les termes de chaque interaction individuelle entre paires d'atomes, un atome de chaque molécule évidemment. Le terme en r^{-6} montre que les forces de dispersion (comme toutes les forces intermoléculaires) décroissent très rapidement quand la distance entre centres en interaction augmente. Avec les molécules organiques complexes, comme celles rencontrées en chromatographie gazeuse, on peut considérer que la cohérence survient uniquement à la surface des molécules (Littelwood, 1963; Dyson et Littelwood, 1967).

Les forces de dispersion qui sont indépendantes de la température, sont toujours présentes dans n'importe quel système soluté – phase stationnaire ; ce sont les seules sources d'attraction entre deux substances non polaires.

2. Interactions dipôle – dipôle ou forces d'orientation

Les forces d'orientation résultent de l'interaction entre deux dipôles permanents. L'énergie moyenne qui en résulte (Keesom, 1921,1922) est :

$$E_k = -\left(\frac{2}{3}\right)\frac{\mu_1^2\,\mu_2^2}{r^6\,k\,T} \qquad (87)$$

μ_1 et μ_2 sont respectivement les dipôles permanents du soluté et du solvant, r la distance entre ces deux dipôles, k la constante de Boltzmann et T la température absolue. Les forces d'orientation décroissent quand on augmente la température, et tendent vers zéro aux très hautes températures, quand toutes les orientations sont équiprobables. Les phases liquides dont la sélectivité aux faibles températures dépend des forces d'orientation, deviennent inefficaces pour les températures élevées.

3. Interactions dipôle – dipôle induit

Les forces d'induction résultent de l'interaction entre un dipôle permanent soit du soluté soit de la phase stationnaire, et un dipôle induit de l'autre. L'énergie moyenne est exprimée par l'équation de Debye (Debye, 1920)

$$E_D = -\left(\frac{1}{r^6}\right)(\alpha_2\mu_1^2 + \alpha_1\mu_2^2) \qquad (88)$$

α étant la polarisabilité.

Les interactions dipôle–dipôle induit ne sont pas isotropes, et dépendent de l'orientation relative des molécules

L'indice de rétention de Kováts (Kováts, 1985) est un système de représentation uniforme des données, utile pour leur report et les comparaisons inter-laboratoires.

Cependant, l'utilisation de séries homologues de *n*-alcanes comme composés de référence pour le calcul des indices de rétention est critiquable à cause du comportement irrégulier de ces hydrocarbures (dû au phénomène d'adsorption sur les colonnes à garnissage polaires (Hawk, 1989; Berezkin et Returnsky, 1984; Lorenz et Roger, 1971; Mathiasson *et al.*, 1978) et sur les colonnes capillaires en verre (Orav, 1993; Krupcik *et al.*, 1982; Matisová et al., 1982], et de leur non détection par de nombreux détecteurs spécifiques. D'où la proposition, avancée par différents auteurs, d'utiliser une série de référence alternative.

L'indice de rétention de Lee (Lee *et al.*, 1979) est basé sur l'utilisation de 4 HAP comme étalons de référence : le naphtalène (I_r = 200), le phénanthrène (I_r = 300), le chrysène (I_r = 400) et le picène ou le benzo[ghi] pérylène (I_r = 500). Les indices de rétention des HAP sont calculés à l'aide de la relation de van den Dool et Kratz (Van den Dool et Kratz, 1963):

$$\frac{I_r}{100} = z + \frac{t_{R_x} - t_{R_z}}{t_{R_{z+1}} - t_{R_z}} \qquad (89)$$

Où t_{R_x} est le temps de rétention de la substance *x* considérée ; t_{R_z} (qui précède t_{R_x}) et $t_{R_{z+1}}$ (qui suit t_{R_x}) sont les temps de rétention des étalons qui encadrent la substance d'intérêt sur le chromatogramme. Le paramètre z représente le nombre de cycles du HAP étalon élué avant la substance *x*. Les indices de rétention des solutés extérieurs à l'intervalle des composés de référence sont obtenus par extrapolation linéaire sur l'intervalle le plus proche.

Cependant, il n'est pas toujours possible d'obtenir des échantillons du matériau standard pur. Ainsi, il est souhaitable de développer des méthodes pour prédire les caractéristiques de rétention du composé inconnu en fonction des caractéristiques structurelles et des propriétés chromatographiques d'autres composés représentatifs. La rétention est un phénomène qui dépend principalement des interactions entre le soluté et la phase stationnaire. Idéalement, chaque soluté présentera des caractéristiques de rétention uniques en fonction de ses propriétés chimiques, structurelles et électroniques.

La méthodologie QSPR est largement acceptée dans divers domaines d'application, qui relient les propriétés d'une molécule à sa structure. Le processus qui met en relation la structure chimique avec la rétention chromatographique comprend un domaine de recherche connu sous le nom de relations quantitatives structure-rétention (QSRR). De nombreuses publications sont apparues dans ce domaine au cours des cinq dernières décennies, y compris un livre de Kaliszan (Kaliszan, 1987) basé sur plusieurs aspects du développement de modèles d'estimation valides et de la signification des paramètres du modèle.

Références bibliographiques

Berezkin V G, Returnsky V N, Calculation of invariant retention indices for a series of aliphatic alcohols and acetates by consideration of adsorption at the interface . J. Chromatogr A. 1984. 292, 9.

Debye P, Van der Waals cohesion forces. Phys. Z. 1920. 21, 178.

Dyson N, Littlewood A B, Effect of organic vapour molecules on the viscosities of hydrogen and helium, Trans. Farad. Soc. 1967. 63, 1895.

Grob R L, Modern Practice of Gas Chromatography; John Wiley & Sons: New York, 1985.

Hawke S J, Uncertainty resulting from inconstancy in the slope of the log plot of homologous series, Chromatographia. 1989. 28, 237.

Hirschfelder J O, Curtiss C F, Bird R B, The Molecular Theory of Gases and Liquids,2nd edn., Wiley, New York, 1964.

Kaliszan R, Quantitative Structure-Chromatographic Retention Relationships; John Wiley & Sons: New York, 1987.

Keesom W H, The calculation of the molecular quadrupole moments from the state equation, Phys. Z. 1922. 23, 225.

Keesom W H, Van der Waals attractive force, Phys. Z. 1921. 22, 129.

Kováts Sz E, Gas-Chromatographische Charakterisierung organischer Verbindungen. Teil 1: Retentionsindices aliphatischer Halogenide, Alkohole, Aldehyde und Ketone, Helv. Chimica Acta. 1958.41, 1915.

Krupcik J, Matisová E, Garaj J, Soják L, Berezkin V G., Contribution of adsorption to retention data in glass capillary gas chromatography part. I. Polar stationary phases, Chromatographia. 1982. 16, 166.

Lee M, Vassilaros D L, White C M, Novotny M, Retention indices for programmed-temperature capillary-column gas chromatography of polycyclic aromatic hydrocarbons, Analytical Chemistry. 1979. 51, 768.

Littlewood A B, The classification of stationary liquids used in gas chromatography, J. Gas Chromatg. 1963. 1, 16.

London F, The general theory of molecular forces, Trans. Faraday Soc. 1937.33, 8b-26.

Lorenz L J, Rogers L B, Specification of gas-chromatographic behavior using Kovats indices and Rohrschneider constants. Analytical Chemistry. 1971. 43, 1593.

Mathiasson L, Jönsson J Å, Olsson A M, Haraldson L, Sensitivity of retention index to variations in column liquid loading and sample size. J. Chromatogr A. 1978. 152, 11.

Matisová E, Krupcik J, Garaj J, Contribution of adsorption to rentension data in capillary gas chromatography part II. Non-polar stationary phases, Chromatographia. 1982. 16, 169.

Mc Lachlan A D, Effect of the medium on dispersion forces in liquids, Discuss. Faraday Soc. 1965. 40, 239.

Messadi-Lourici L, thèse de doctorat "contribution a l'étude des indices de rétention en chromatographie gazeuse" Laboratoire de Sécurité Environnementale et Alimentaire Université BADJI Mokhtar-Annaba-ALGERIE, 2004

Orav A, Kuningas K, Range S, A comparison of different retention index systems with unsaturated and aromatic hydrocarbons in capillary gas chromatography on PEG 20M, Chromatographia. 1993. 37, 411.

Van den Dool H, Kratz P Dec, A generalization of the retention index system including linear temperature programmed gas—liquid partition chromatography. J. Chromatogr A., 1963. 11, 436.

I.4. La solubilité aqueuse (S_w)

La solubilité aqueuse (S_w) des composés organiques est l'un des facteurs clés à prendre en considération lors du classement des produits chimiques organiques significatifs pour l'environnement par rapport à leur mobilité dans le sol et leur volatilité de la surface de l'eau. C'est aussi un paramètre particulier important dans les études sur l'absorption, la distribution, les métabolismes et l'excrétion xénobiotique chez les êtres humains. Cependant, la mesure expérimentale de la solubilité est difficile car elle peut être très longue pour atteindre l'équilibre de solubilité dans le cas des composés apolaires ou nécessiter une grande quantité de produits chimiques dans le cas de molécules hautement hydrophiles. En outre, les valeurs de la solubilité de la majorité des composés organiques restent inconnues (Muir et Howard, 2006).

La prédiction de la solubilité dans l'eau est importante dans les sciences de l'environnement. Les approches les plus utiles pour l'estimation de la solubilité dans l'eau et dans des solutions non aqueuses sont basées sur une contribution de groupes fonctionnels. Le coefficient UNIversel d'ACtivité Fonctionnelle (en anglais UNIFAC) est une approche fiable et rapide pour prédire les coefficients de solubilité / activité aqueuse des non-électrolytes dans le mélange liquide (Fredenslund *et al.*, 1975; Hensen *et al.*, 1991). Ce modèle a été étendu à l'eau comme solvant. Le principe fondamental est qu'un composé comporte des groupes fonctionnels, et chaque groupe apporte une contribution unique à la solubilité aqueuse. Un grand nombre de composés contiennent simplement un nombre limité de groupes fonctionnels, et il est donc possible de prévoir la solubilité de nombreux composés en utilisant un nombre minimal de groupes fonctionnels organiques. Cependant, l'exactitude du modèle proposé par la Fédération internationale des Nations Unies reste controversée (Kan et Tomson, 1996). Une autre approche similaire axée sur la prédiction de la solubilité aqueuse est bien connue sous le nom de coefficients d'activité de groupe fonctionnel aqueux (AQUAFAC) (Myrdal *et al.*, 1993). L'idée fondamentale est d'exprimer les contributions enthalpiques et entropiques à l'énergie excédentaire en additionnant les parties interactives du soluté, les molécules organiques dissoutes et leurs groupes fonctionnels. Certaines méthodes prometteuses pour prédire la solubilité aqueuse sont basées sur des descripteurs moléculaires topologiques, géométriques et électroniques (Mitchell et Jurs, 1998).

Duchowicz *et al.*, (2008) ont développé un modèle QSPR linéaire, généralement applicable, basé sur 147 composés pharmaceutiques contenant trois descripteurs moléculaires.

Kim *et al*.,(2008) ont corrélé la solubilité aqueuse de médicaments peu solubles, tels que l'acide ursodésoxycholique, la diphénylhydrantoïne et le diméthylbiphényldicarboxylate. Trois ensembles de données de 50 composés ont été extraits des données de la littérature en fonction de leur similitude structurelle avec chaque médicament. Des modèles QSPR rapides et prédictifs (R^2> 0,90) ont été développés et validés (R^2> 0,85). Huuskonen *et al*.,(2008) ont extrait un ensemble de calibrage de 191 composés antidrogue de la base de données AQUASOL pour corréler la solubilité aqueuse par un modèle à cinq paramètres (C log P, poids moléculaire, variable indicatrice pour les groupes amines aliphatiques, nombre de liaisons rotatives et un nombre d'anneaux aromatiques) avec les statistiques: R^2 = 0,87 et s = 0,51. Le modèle a été appliqué à un ensemble d'essai de 174 composés antidrogue avec R^2 = 0,80 et s = 0,68. Les résultats de cette étude suggèrent que l'augmentation de la taille moléculaire, la rigidité et la lipophilie diminuent la solubilité alors que l'augmentation de la flexibilité conformationnelle et la présence d'une amine non conjugué augmente la solubilité des composés pharmaceutiques. Du-Cuny *et al*., (2008) visant à modéliser la solubilité aqueuse de composés apparentés aux médicament dans des séries congénères, la lipophilie (C log P) combinée à l'information sur les fragments structurels, les facteurs de correction basés sur des fragments et les indices de séries de congénères ont été utilisés comme descripteurs pour une ACP suivie d'une régression PLS multivariée. Le modèle général résultant (R^2 = 0,84 et rms = 0,51) était basé sur un ensemble de données internes de 1515 composés pharmaceutiques, et la solubilité de l'ensemble d'essai de 958 composés était prédite avec un haut degré de précision, R^2 = 0,81 et s = 0,42. Au cours du développement du modèle, des règles ont été dérivées qui peuvent être utilisées par les médecins chimistes ou les scientifiques intéressés en tant que directive approximative sur la contribution des fragments structurels à la solubilité.

Références bibliographiques

Duchowicz P R, Talevi A., Bruno-Blanch L E, Castro E A., New QSPR study for the prediction of aqueous solubility of drug-like compounds. Bioorg. Med. Chem. 2008. 16, 7944.

Du-Cuny L , Huwyler J, Wiese M, Eur. Kansy M, Computational aqueous solubility prediction for drug-like compounds in congeneric series. J. Med. Chem. 2008, 43, 501.

Fredenslund A, Jones R L, Prausnitz J M, Group-Contribution Estimation of Activity Coefficients in Nonideal Liquid Mixtures. Aiche J., 1975. 21, 1086

Hansen H K , Rasmussen P, Fredenslund A, Schiller M, Gmehling J, Vapor-Liquid Equilibria by UNIFAC Group Contribution. 5. Revision and Extension. Ind. Eng. Chem. Res., 1991. 30, 2352.

Huuskonen J, D. Livingstone J, Manallack D T, Prediction of drug solubility from molecular structure using a drug-like training set. SAR QSAR Environ. Res. 2008, 19, 191.

Kan A T, Tomson M B, UNIFAC Prediction of Aqueous and Nonaqueous Solubilities of Chemicals with Environmental Interest. Environ. Sci. Technol. 1996. 30, 1369.

Kim J, D Jung H, Rhee H, Choi S H, Sung M J., Choi W S, Aqueous solubility of poorly water-soluble drugs: Prediction using similarity and quantitative structure-property relationship models. Korean J. Chem. Eng. 2008, 25, 865.

Mitchell B E, Jurs P C, Prediction of Infinite Dilution Activity Coefficients of Organic Compounds in Aqueous Solution from Molecular Structure. J. Chem. Inf. Comput. Sci. 1998. 38, 200.

Muir D C G, Howard P H, Are there other persistent organic pollutants? A challenge for environmental chemists. Envi. Sci & Tech. 2006. 40 , 7157.

Myrdal P, Ward G H, Simamora P, Yalkowsky S H , AQUAFAC: Aqueous Functional Group Activity Coefficients. SAR QSAR Environ. Res. 1993. 1, 53.

I.5. Le coefficient de partage octanol/carbone organique (Koc)

Le coefficient de sorption du sol mesure la capacité de partage d'un composé entre deux phases, la phase liquide (c'est-à-dire l'eau) et la phase solide (c'est-à-dire les composants du sol). La fraction organique du sol est reconnue comme la partie du sol qui est responsable de la sorption des contaminants. Cela a conduit à la normalisation du coefficient de sorption du sol par la teneur en carbone organique, Koc, qui rend également comparable les données expérimentales provenant de différents sols.

La sorption de produits chimiques par le sol et les sédiments joue un rôle très important dans leur transport et leur mobilité dans l'environnement. En outre, la sorption peut influencer de manière significative sur la transformation chimique et biologique ou sur la dégradation des produits chimiques dans l'environnement (Sabljić *et al.*, 1995).

Ainsi, pour la plupart des produits chimiques organiques non ioniques, la mesure ou l'estimation exacte du (Koc), paramètre largement utilisé pour mesurer la sorption et la conservation d'un produit chimique par la matière organique des sols et des sédiments(Jury,1986), est d'une importance cruciale pour évaluer leur sort et leur exposition potentielle dans l'environnement et, par conséquent, pour l'ensemble du processus d'évaluation des risques environnementaux. En général, les composés ayant des valeurs plus élevées de Koc ont tendance à être moins mobiles que ceux ayant des valeurs plus faibles dans les systèmes sol-eau (Sabljić *et al.*, 1995).

Malheureusement, les valeurs de Koc sont rares à trouver dans la littérature pour de nombreux polluants, et souvent les données rapportées ont une précision et une exactitude médiocres. Cela s'explique principalement par des difficultés expérimentales à partir de procédures telles que la préparation, la manipulation et l'analyse (Ferreira, 2001). En outre, il est peu pratique de mesurer le Koc de chaque produit chimique directement au laboratoire car il existe de nombreux produits chimiques dans la nature.

Plus de 200 modèles QSPR sur la sorption du sol ont été examinés par Gawlik *et al.*, (1997) . Il a été démontré que les valeurs logarithmiques du Koc étaient le plus souvent modélisées avec la solubilité dans l'eau (Sw), le coefficient de partage n-octanol / eau (Kow), le facteur de capacité CLHP-PI (chromatographie liquide haute performance-à phase inverse (k'), les indices topologiques ou les paramètres d'énergie de solvatation linéaire. log Kow était le plus couramment utilisé pour décrire la sorption du sol. La plupart des QSPR partaient sur

des classes chimiques et des sols spécifiques. En raison de l'hétérogénéité des constituants organiques du sol, la sorption peut impliquer des mécanismes spécifiques ou non. Les interactions non spécifiques sont caractérisées par des composés hydrophobes. Les indices topologiques ont donné de bons résultats dans la modélisation des séries homologues de composés. La taille et la ramification, attribuables aux indices topologiques, peuvent affecter physiquement la mobilité du contaminant dans la matrice humique. Les interactions spécifiques de la sorption des sols exposés par des composés polaires ont été couvertes par l'inclusion de descripteurs moléculaires invariables (WHIM), constitutionnels, électrostatiques, quanto chimiques et pondérés. Des modèles plus généraux résultaient de l'incorporation de descripteurs pour les deux mécanismes de sorption. En outre, des réactions chimiques peuvent avoir lieu avec les composants du sol, ce qui diminue la mobilité de ces composés.

Références bibliographiques

Ferreira M M C , Polycyclic aromatic hydrocarbons: a QSPR study. Chemosphere. 2001. 44,125.

Gawlik B M, Sotiriou N, Feicht E A. Schulte-Hostede S, Kettrup, A. Alternatives for the determination of the soil adsorption coefficient, KOC, of non-ionicorganic compounds: a review. Chemosphere. 1997. 34, 2525.

Jury W A. Adsorption of organic chemicals onto soil, in Henn SC, Melancon SM (eds.),Vadose Zone Modeling of Organic Pollutants, Lewis Publisher, 1986. 177.

Sabljić A, Gusten H, Verhaar H, Hermens J. QSAR modelling of soil sorption. Improvements and systematics of log KOC vs. log KOW correlations. Chemosphere.1995. 31, 4489.

II. MÉTHODOLOGIE

1. Traitements des données :

Le développement du modèle nécessite le passage de notre base de composés chimiques par plusieurs logiciels de traitement:

A. **ChemDraw V.7.0** : Utilisé pour dessiner et structurer les composés chimiques (ChemDrawUtra, 2002)
B. **HyperChem Pro V6.03** : Il offre plusieurs méthodes d'optimisation (PM3, MM+, AM1, etc ….) par lesquelles les molécules sont optimisées, des descripteurs moléculaire peuvent être calculés telle que; le volume moléculaire, les énergies HUMO et LUMO…etc. (HyperChem, 1999).
C. **DRAGON V5.3** : Il assure le calcul des descripteurs moléculaires (1664 descripteurs appartenant à différentes classes) (Talet Srl, 2007).
D. **QSARINS V2.2 :** QSARINS (QSAR-INSUBRIA) est un nouveau logiciel pour le développement et la validation des modèles de régression linéaire multiple (MLR) par les moindres carrés ordinaires (MCO) et algorithme génétique pour la sélection de variables. Ce programme est principalement axé sur la validation externe des modèles QSAR. Divers outils pour l'analyse exploratoire des données par Analyse en Composantes Principales, réduction de descripteurs moléculaires d'entrée, division d'ensembles de données en ensembles d'entraînement et de prédiction, détection de valeurs aberrantes et prédictions interpolées ou extrapolées, validation interne et externe par différents paramètres, modélisation consensuelle et diverses graphs pour les visualisations sont implémentées. QSARINS est une plate-forme conviviale pour la modélisation QSAR en accord avec les Principes de l'OCDE et pour l'analyse de la fiabilité des données prédites obtenues (Gramatica *et al.*, 2012) .

2. Optimisation et calcul des descripteurs

Les descripteurs moléculaires théoriques ont été calculés par le processus suivant. Premièrement, les structures moléculaires ont été pré-optimisées par le champ de force de la mécanique moléculaire MM^{+} du logiciel de modélisation moléculaire HyperChem 6.03 (HyperChem, 1999). La géométrie finale de la conformation d'énergie minimale a été obtenue par la méthode semi-empirique PM3 avec un niveau Hartree-Fock restreint sans interaction de configuration, en appliquant un gradient limite standard de 0,001 Å $kcal.mol^{-1}$ comme critère d'arrêt. Les molécules optimisées par HyperChem ont été transférés dans le logiciel Dragon

(Talet Srl, 2007) pour calculer un grand nombre de descripteurs moléculaires à partir de la structure géométrique et électronique des molécules. Les descripteurs quantiques (moment dipolaire, énergies des orbites frontières: Ɛ HOMO, Ɛ LUMO, etc…) ont été obtenus à l'aide du logiciel HyperChem. Après la génération de 1675 descripteurs pour chacun des hydrocarbures aromatiques polycycliques, une présélection de descripteurs a été réalisée dans le but de réduire le nombre de descripteurs.

3. Réduction du nombre de descripteurs et méthode de sélection

a. Réduction du nombre de descripteurs

Un grand nombre de descripteurs différents sont collectés pour la modélisation d'une grandeur donnée (activité ou propriété), car les facteurs déterminants du processus étudié ne sont pas connus. Cependant, les descripteurs envisagés n'ont pas tous une influence significative sur la grandeur modélisée, et les variables ne sont pas toujours indépendantes. De plus, le nombre de descripteurs, c'est-à-dire la dimension de la base de données d'entrée, détermine la dimension du vecteur des paramètres à ajuster. Si cette dimension est trop importante par rapport au nombre d'observations (molécules) de la base d'apprentissage, le modèle risque d'être sur-ajusté à ces exemples, incapable de prédire la grandeur modélisée sur de nouvelles molécules et peut contenir des informations redondantes. Pour que les modèles QSAR/QSPR soient simples et compréhensibles, il faut que les descripteurs employés soient significatifs et interprétables (Mannan, 2005).

Dans une modélisation QSAR/QSPR, il faut que le modèle ait le moins de paramètres possibles tout en expliquant au mieux la propriété ou l'activité (Crawley, 2005).

Avant d'entamer le développement effectif des équations de régression QSAR/QSPR, il est hautement recommandé d'examiner la qualité statistique des données de départ, à la fois les données à corréler (variable dépendante) et les descripteurs utilisés dans la corrélation (variables indépendantes)(Nalimov,1962 ; Katritzky *et al*., 1994; Dagnélie, 1998). Il est nécessaire de calculer le coefficient de corrélation entre chacune des paires de l'ensemble des descripteurs. Si ce coefficient est statistiquement significatif ($R>0,95$), ces deux descripteurs sont considérés comme fortement corrélés et ne peuvent être utilisés simultanément lors de l'analyse QSAR/QSPR (Trinajstiæ *et al.*, 2001) et en pratique, ils seront alors enlevés dans le procédé de sélection. Ce type d'analyse permet de réduire le nombre de descripteurs sans faire participer la variable dépendante (l'activité ou la propriété).

b. Méthode de sélection de descripteurs

Les meilleurs descripteurs sont sélectionnés en explorant la qualité statistique de toutes les combinaisons possibles des descripteurs disponibles, en utilisant la régression par les moindres carrés ordinaires (OLS) et la sélection de sous-ensembles de variables par algorithme génétique (AG-SSV) [le logiciel QSARINS (version 2.2) (Gramatica *et al.*, 2012)]. Cette procédure de «sélection de variables» génère une «population» de modèles, classés selon les valeurs décroissantes de R^2. Les meilleurs modèles ont été choisis en utilisant le Q^2_{LOO} comme valeur d'optimisation et en tenant compte du principe de parcimonie concernant la dimension des modèles qui devrait être aussi petite que possible. En outre, la corrélation entre les descripteurs et la réponse modélisée est vérifiée par la règle *QUIK* (Q Under Influence of K) pour exclure les modèles à forte colinéarité entre les descripteurs (Gramatica *et al.*, 2013).

4. Répartition des échantillons

Pour le choix des composés de calibrage et de validation, nous avons utilisé soit une procédure aléatoire de sélection d'échantillons soit une sélection par algorithme de Kennard et Stone **(**Kennard et Stone,1969).

a) Le choix aléatoire

La sélection aléatoire des échantillons est la technique la plus simple. Elle suppose qu'un groupe d'échantillons extrait aléatoirement à partir d'un lot suffisamment grand suit la distribution statistique du lot entier (Wu *et al* ., 1996) .

b) Algorithme de Kennard et Stone

Une alternative à la sélection aléatoire est l'utilisation de l'algorithme de Kennard et Stone **(**Kennard et Stone, 1969; Dantas et al., 2004). L'algorithme maximise la distance euclidienne minimale entre les échantillons déjà sélectionnés et les échantillons restants.

Cette procédure illustrée par la figure 3, est rappelée ci-après:

- a) sélection des échantillons les plus éloignés. Il s'agit ici des échantillons 1 et 2 qui sont entourés sur la Figure 9a ;
- b) pour chaque échantillon restant, calcul de la distance euclidienne par rapport à l'échantillon le plus proche déjà sélectionné (Figure 9b) ;
- c) sélection de l'échantillon ayant la plus grande distance avec l'échantillon déjà sélectionné. Le troisième échantillon sélectionné est l'échantillon n°4.

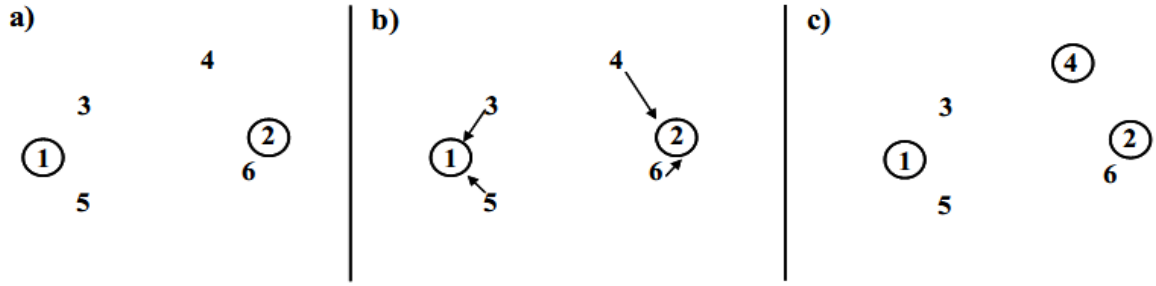


Figure. 9 Répartition des échantillons avec l'algorithme de Kennard et Stone

Cette procédure est répétée jusqu'à ce que le nombre d'échantillons spécifié par l'utilisateur soit atteint. Cependant, le nombre d'échantillons à inclure dans chaque lot n'est pas le problème fondamental. En effet, il est plus important de savoir qu'elles sont les variabilités présentes dans le lot de données et comment sélectionner les échantillons du lot d'entraînement afin de prendre en compte ces variabilités (Fearn, 2005).

Dans l'ensemble de notre travail, nous avons utilisé la régression linéaire multiple MLR, et les réseaux de neurones artificiels ANN pour la construction des modèles QSPR et l'analyse en composante principale (ACP) comme technique pour l'analyse descriptive des données (classification des données).

5. La Régression Linéaire Multiple (MLR)

La régression linéaire multiple MLR est l'une des méthodes de modélisation les plus en vogue grâce à sa simplicité d'utilisation et son interprétation facile. L'avantage important de la régression linéaire multiple est qu'elle est très transparente, puisque l'algorithme est disponible, et que les prédictions peuvent être réalisées facilement (Roy *et al.*, 2015)

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

$$y = a_0 + \sum_{i=1}^{n} a_i x_i \qquad (90)$$

Où: y est la variable dépendante (à expliquer ou à prédire) ; les x_i sont les variables indépendantes (explicatives) ; *n* est le nombre de variables explicatives ; a_0 est la constante de l'équation du modèle ; les a_i représente les coefficients de descripteurs dans l'équation du modèle ;

Les variables indépendantes *xi*, comme leur nom l'indique, sont supposées indépendantes entre elles.

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

6. Les réseaux de neurones artificiels (RNA)

5.1 Les neurones biologiques :

Le cerveau humain est constitué d'un très grand nombre de cellules nerveuses appelées neurones, environs 100 milliards, avec 1000 à 10 000 synapses (connexions) par neurone (Roland et Blouch, 2002).

Le neurone biologique (Figure 10) est une cellule nerveuse spécialisée dans le traitement de l'information (signaux électriques). Il est constitué de trois composantes principales :

Les dendrites : fines prolongations du corps cellulaire entourant celui-ci en une sorte de filet qui capte les oscillations et les informations issues d'autres cellules nerveuses et les transmettent au corps cellulaire.

Le corps cellulaire : qui a pour fonction de recevoir les excitations, les intégrer et les transmettre ou non. Il contient également le noyau qui assure la vie du neurone.

L'axone : Les axones conduisent les signaux électriques de la sortie d'un neurone vers l'entrée à un autre neurone. Le point de contact entre l'axone d'un neurone et la dendrite d'un autre neurone s'appelle la synapse (Torres-Morino, 1992; Fadlallah, 2005).

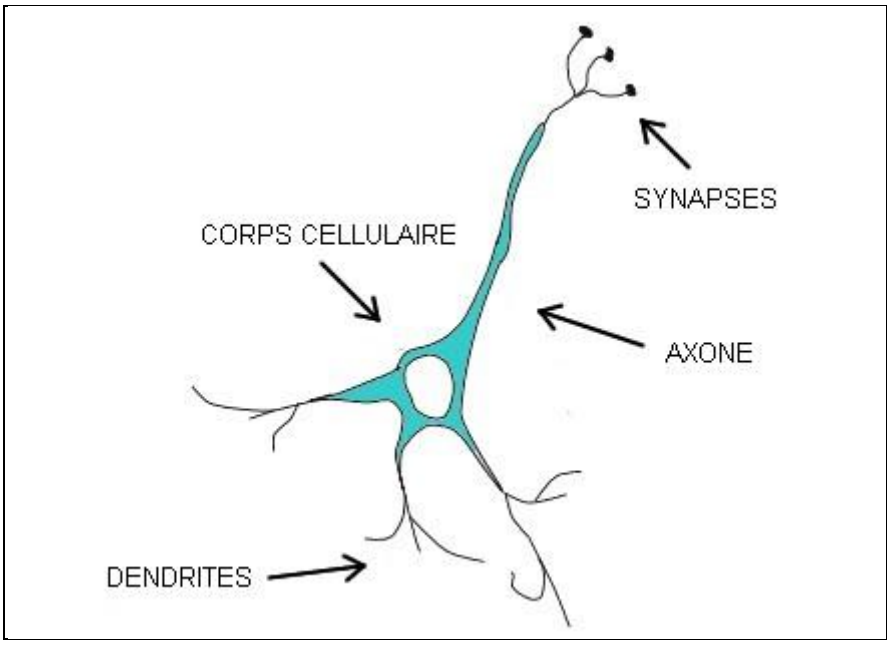


Figure. 10 Le neurone biologique

Au niveau du neurone se produit une intégration (sommation) des signaux reçus et si cette somme dépasse un certain seuil le neurone émet à son tour un signal électrique vers d'autres neurones. Ce signal peut renforcer ou diminuer l'activité des neurones qui le reçoivent selon que les synapses sont excitatrices ou inhibitrices (Chabaa, 2011).

5.2 Les réseaux de neurones artificiels (RNA)

L‘approche par les RNA est analogue aux systèmes de neurones biologiques qui permettent de traiter et de transmettre des informations en faisant circuler des signaux électriques dans un réseau constitué d‘axones. Chaque neurone artificiel est un processeur élémentaire. Il est donc avant tout un opérateur mathématique avec des « entrées » (variables de la fonction mathématique) et des « sorties » (valeurs de la fonction). L‘intérêt des neurones réside dans les propriétés qui résultent de leur association en réseaux, c‘est-à-dire de la composition des fonctions réalisées par chacun des neurones. Il reçoit un nombre variable d‘entrées en provenance de neurones en amont ou des capteurs composant la machine dont il fait partie. A chacune de ses entrées est associé un poids (wi) représentatif de la force de la connexion. Chaque processeur élémentaire est doté d‘une sortie unique, qui se ramifie ensuite pour alimenter un nombre variable de neurones en aval. 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 moléculaires, 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.

La couche de sortie compte autant de neurones que de propriétés modélisées ; dans notre cas une seule. 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 (w_i) associées aux neurones d'entrée sont ajustés itérativement, afin de minimiser l'erreur entre la propriété calculée et la propriété expérimentale.

La sortie d'un neurone, dépend donc de l'entrée du neurone et de sa fonction de transfert. Il existe essentiellement trois types de fonction de transfert qui sont les fonctions à seuil, les fonctions sigmoïdes et les fonctions linéaires (***Figure 11***). La fonction sigmoïde est la plus utilisée car elle représente un bon compromis entre les fonctions seuils et linéaires.

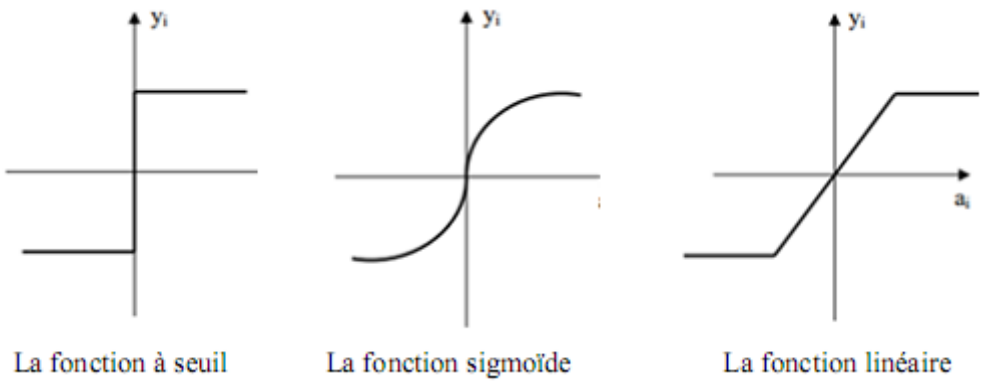


Figure. 11 Différents types de fonction de transfert pour le neurone artificiel

Il existe deux types de réseaux de neurones : les réseaux non bouclés et les réseaux bouclés. Les réseaux de neurone non bouclés réalisent une (ou plusieurs) fonction algébrique de ses entrées, par composition des fonctions réalisées par chacun de ses neurones. Il s'agit donc d'un ensemble de neurones connectés entre eux, l'information circulant des entrées vers les sorties sans retour en arrière possible. On parle souvent de perceptron multicouche à cause de la présence de neurones cachés (***Figure 12***).

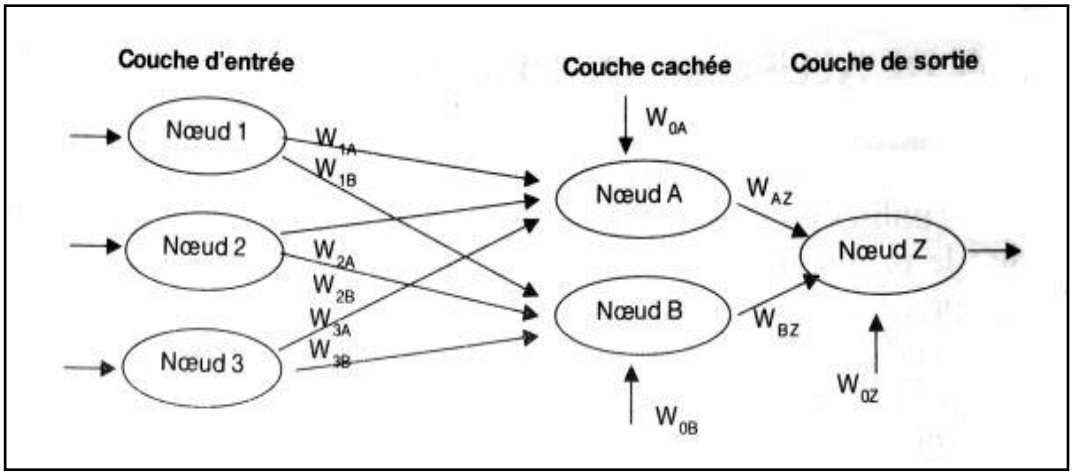


Figure 12 : Topologie d'un réseau de neurones à *n* entrées et une seule sortie

5.3 Apprentissage des réseaux de neurones artificiels

Dans le domaine des réseaux de neurones, l'apprentissage est une phase très importante qui désigne la procédure ou la façon qui consiste à déterminer l'architecture et les paramètres du réseau. En effet, une des propriétés fondamentales d'un réseau neuronal est sa capacité à s'adapter et améliorer sa performance en ajustant les connexions des neurones face à une source d'informations par la procédure d'apprentissage (Ammar, 2007).

L'apprentissage des réseaux de neurones artificiels se fait grâce à des algorithmes d'apprentissage. Dans la majorité des algorithmes actuels, l'apprentissage consiste à modifier les poids de connexions pour que la réponse du réseau s'accorde aux exemples de l'expérience (Torres-Morino, 1992; Fadlallah, 2005).

Après une initialisation aléatoire des poids, des exemples expérimentaux sous formes de couples de vecteurs d'entrées et de sorties sont présentés au réseau. Les poids sont modifiés graduellement à l'aide des algorithmes d'apprentissage en vue de minimiser l'écart entre les sorties calculées (estimées) par le réseau et les observations.

Mise en œuvre :

La base des données est divisée en deux parties :

- L'ensemble d'apprentissage : sur lequel se fait l'optimisation des poids.

- L'ensemble de test : sur lequel on teste la capacité de généralisation du réseau de façon à ce que les poids retenus soient ceux pour lesquels l'erreur obtenue sur cette base est faible.

En effet, si les poids sont ajustés sur toutes les données de l'ensemble d'apprentissage (≈ 70% de la base de données globale), on risque d'avoir le « sur-apprentissage » ou l'apprentissage par cœur, dans ce cas le réseau apprend très bien les données présentées dans la phase d'apprentissage sans pour autant être capable de généraliser le modèle à des données nouvelles.

Pour éviter le « sur-apprentissage » on introduit un nouvel ensemble de données appelé l'ensemble de validation (≈ 30% de la base de données globale). Comme pour l'ensemble de test (≈ 30% de la base de données globale), les éléments de cet ensemble ne participent pas à l'apprentissage. De plus, cet ensemble doit bien sûr avoir les mêmes contraintes que l'ensemble de test quant à sa représentativité et sa taille.

L'ensemble de validation est utilisé de la façon suivante : dès que l'on s'aperçoit que l'erreur sur l'ensemble de validation stagne ou augmente, alors on arrête la procédure d'apprentissage (Ammar, 2007).

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

Références bibliographiques

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

Chabaa S, identification des systèmes non linéaires en utilisant les techniques d'intelligences artificielles et les bases de fonctions de Laguerre pour la modélisation des données du trafic dans les réseaux internet, Thèse de Doctorat, Université Cadi Ayyad, faculté des sciences Semlalia - Marrakech, 2011.

Chem Draw Ultra"ultra-chemical structure drawing standard". Version 7. Copyright Cambridge Soft Coperation. 2002.

Crawley M J, Statistics: an introduction using R, Wiley, Chichester, UK, 2005.

Dagnélie P, Statistique théorique et appliquée. Tomes 1 et 2, De Boeck & Larcier, 1998.

Dantas Filho H A, Harrop Galvao R K, Ugulino Araujo M C, Da Silva E C, Bezerra Saldanha T C, Jose G E, Pasquini C, Raimundo I M, Rodrigues Rohwedder J J, A strategy for selecting calibration samples for multivariate modelling, Chemometr Intell Lab.2004, 72, 83.

Fadlallah N, Contribution à l'optimisation de la synthèse du lobe de rayonnement pour une antenne intelligente. Application à la conception de réseaux à déphasage, Thèse de Doctorat, Université de Limoges, Facultés des Sciences et Techniques, 2005.

Fearn T , Chemometrics: an enabling tool for NIR, NIR news 2005, 16, (7), 17

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

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

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

Katritzky A R, Lobanov V S, Karelson M, CODESSA Reference Manual, Universityof Florida, Gainesville, 1994.

Kennard R W, Stone L A, Computer aided design of experiments. Technometrics 1969, 11, 137.

Mannan S, Lee's Loss Prevention in Process Industries: Hazard Identification, Assessment and Control, Elsevier Butterworth-Heinemann, Burlington, 2005.

Nalimov V Y, the Application of Mathematical Statistics to Chemical Analysis, AddisonWesley, Reading, MA, 1962.

Rolland J, Blouch P, Les bouées météorologiques : L'exemple de Météo-France‖, La Météorologie, 2002, 39, 83

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

Roy P P, Paul S, Mitra I, Roy K, Two novel parameters for validation of predictiveQSAR models, Molecules, 2009, 14, 1660.

Talete Srl. Dragon for windows (Software for Molecular Descriptor Calculation) Version 5.5 Milano, Italy, 2007; software available at: http://www.talete.mi.it.

Topliss JG, Edwards R P, Chance factors in studies of quantitative structure-activity relationships, Journal of Medicinal Chemistry, 1979, 22, 1238.

Torres-Moreno J M, Apprentissage et généralisation par des réseaux de neurones : étude de nouveaux algorithmes constructifs, Thèse de Doctorat, Institut Nationale Polytechnique de Grenoble, 1992.

Trinajstiæ N, Nikoliæ S, Basak S C, Lukovits I, Distances indices and their hyper-counterparts: Intercorrelation and use in the structure-property modeling, SAR and QSAR in Environmental Research, 2001, 12, 31.

Wu W. Massart, D L. Artificial neural networks in classification of NIR spectral data: Selection of the input. Chemometrics and Intelligent Laboratory Systems 1996, 35, 127.

PARTIE III

RÉSULTATS ET DISCUSSIONS

RELATİONS STRUCTURE PROPRİÉTÉS DES 16 HAP PRİORİTAİRES

1. Origine des données

Les valeurs des six propriétés physico-chimiques des HAP rapportées dans (EPA,1990), ont été utilisées pour l'analyse des relations structure / propriétés. Les cinq propriétés physico-chimiques des HAP ont été reliées séparément à des paramètres structuraux (1 et 2 descripteurs).

2. Les modèles MLR

Les meilleurs modèles, de dimension 2 en général, sont basés sur des descripteurs choisis parmi l'ensemble: ***J***, ***RDF060p***, ***X1sol***, ***Mor16m***, ***Gm***, ***Xt***, ***Ss***, ***ITH*** et le **Moment Dipolaire**.

Les modèles retenus ont pour équations respectives:

Ir = - 39,90 (±2,88) + 48,40 (±0,36) ***X1sol*** – 31,90(±4,24) ***Mor16m*** (91)

Log S = -56,50(±2,08) + 24,90 (±1,21) ***J*** + 0,16 (±0,06) ***RDF060p*** (92)

Tfus = 510 (±58,60)+ 128(±31,99) ***Gm*** - 1562(±227,20) ***Xt*** (93)

Teb = - 33,50 (±12,00) + 13,9 (±0,38) ***Ss*** (94)

Log Koc = 3,37(±0,17) – 2,34(±0,35) ***Moment Dipolaire*** + 0,04 (±0,002) ***ITH*** (95)

Tableau.2 Valeurs des paramètres statistiques pour les différents modèles.

	n	***R²***	***Q^2_{Loo}***	***R^2_{adj}***	***$Q^2_{LMO30\%}$***	***s***	***F***	***RMSE***	***PRESS***	***R² Yscr***	***Q² Yscr***
Ir	16	0,999	0,999	0,999	0,998	2,485	11661,529	2,240	144,444	0,114	-0,354
LogS	15	0,977	0,963	0,973	0,961	0,673	254,559	0,602	8,707	0,128	-0,389
Tfus	16	0,793	0,703	0,761	0,684	31,856	24,993	28,715	18932,874	0,122	-0,357
Teb	15	0,990	0,986	0,989	0,985	10,155	1337,670	9,454	1893,581	0,070	-0,244
LogKoc	16	0,950	0,922	0,943	0,916	0,252	125,224	0,227	1,306	0,141	-0,330

Les petites valeurs de Q^2_{LOO}, R^2, et R^2_{adj} pour la température de fusion indiquent que modèles peu robustes et peu stable (***$Q^2_{LMO30\%}$***).

Tableau.3 valeurs des variables dépendantes (expérimentales et calculées) des HAP prioritaires.

composés	Ir Exp	Ir Calc	- logS Exp	- logS Calc	Tfus Exp(°C)	Tfus Calc	Teb Exp(°C)	Teb Cal	logKoc Exp	logKoc Calc
Naphthalene	200	202,615	8,3061	8,473	81	77,528	217,9	234,600	4,35	4,004
Acenaphthylene	244,63	243,727	10,565	10,443	95	88,555	280	280,930	3,65	3,893
Acenaphthene	251,29	249,051	10,715	10,371	93	87,916	270	267,059	3,65	3,436
Fluorene	268,17	270,768	*	*	114	103,277	295	301,737	3,86	4,020
Anthracene	301,69	303,474	15,193	14,456	216	174,352	339,9	336,416	4,14	4,510
Phenanthrene	300	298,702	12,091	12,934	97	118,639	340	336,416	5,41	4,978
Fluoranthene	344,01	341,307	13,795	13,513	110	157,169	384	382,607	4,57	4,738
Pyrene	351,22	350,670	14,241	14,554	150	213,394	404	382,607	4,57	4,539
Benzo(a)anthracene	398,5	400,932	17,504	18,553	158	145,781	438	438,093	6,13	6,131
Chrysene	400	398,344	18,657	17,864	252	233,695	448	438,093	5,30	5,589
Benzo(a)pyrene	453,44	450,997	19,163	19,042	179	172,457	495	484,284	6,74	6,508
Benzo(b)fluoranthene	441,74	439,641	16,805	17,446	168	171,179	481	484,284	5,74	5,900
Dibenzo(ah)anthracene	495,45	496,509	22,438	22,547	267	274,299	524	539,769	6,51	6,554
Benzo(k)fluoranthene	442,56	441,592	17,988	18,706	217	202,742	480	484,284	5,74	5,581
Benzo(ghi)perylene	501,32	500,519	19,791	19,320	278	224,350	525	530,614	6,20	6,323
Indeno(1,2,3-cd)pyrene	481,87	487,034	20,070	19,099	163	192,660	*	*	6,20	6,051

3. La qualité de l'ajustement

La qualité de l'ajustement a été vérifiée en procédant à une validation croisée par « leave-one-out ». La figure 13, qui reproduit les valeurs calculées de chaque propriété physico-chimique en fonction de celles observées fait ressortir, sauf pour la température de fusion, une faible dispersion caractéristique d'un bon ajustement, confirmé par les grandes valeurs de Q^2_{LOO}.

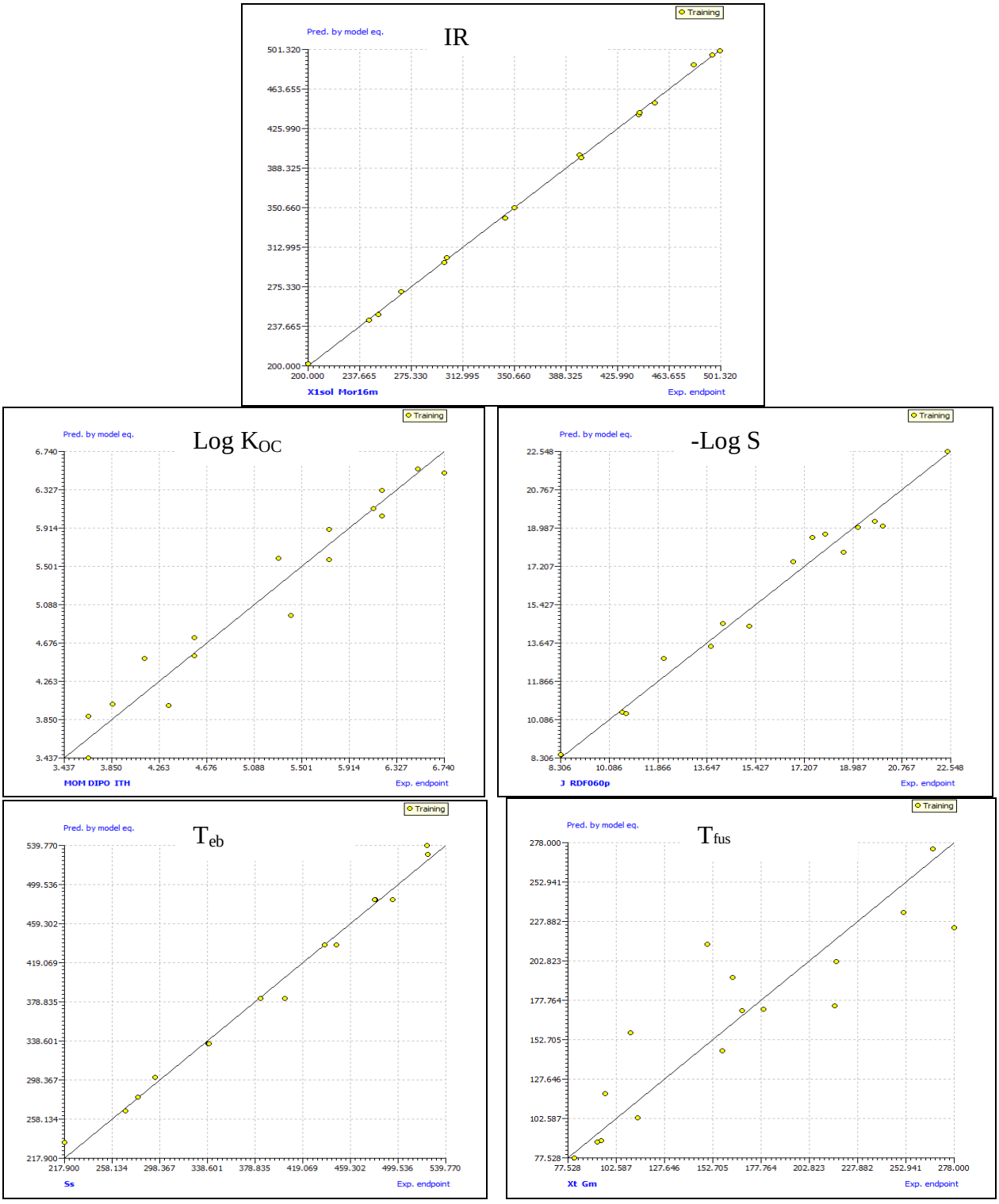


Figure. 13 Graphes des valeurs calculées en fonction des valeurs expérimentales pour chacune des cinq propriétés testées.

4. Test de randomisation

La figure 14 montre le résultat de 100 permutations effectuées sur la colonne des Y (IR, T_{eb}, T_{fus}, Log K_{OC} et -Log S). Il est clair que les statistiques obtenues pour les vecteurs modifiés de toutes les propriétés sont plus petites que celles du modèle réel correspondant, ce qui permet d'affirmer que les modèles proposés ne sont pas aléatoires.

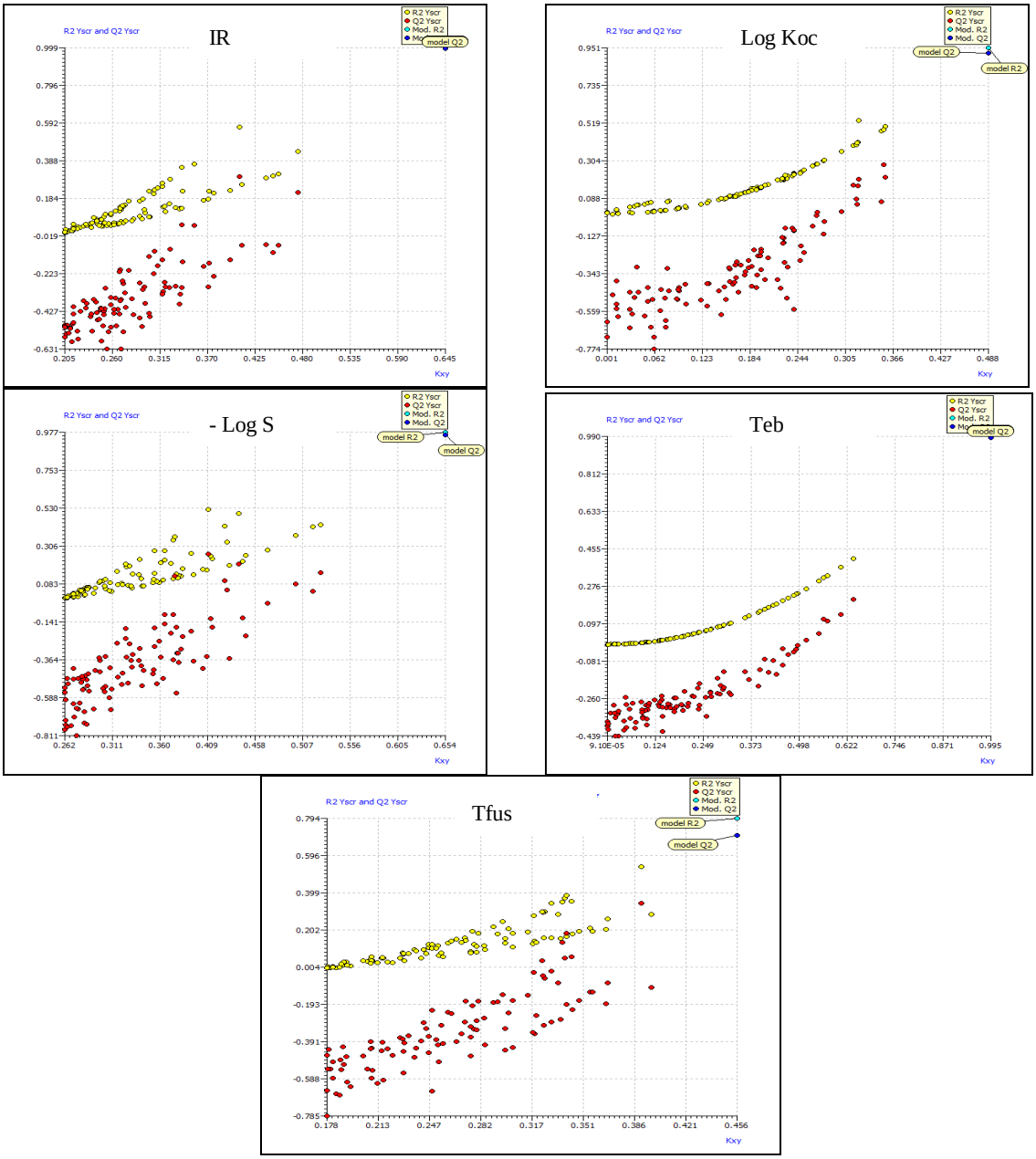


Figure.14Test de randomisation pour chacun des modèles calculés.

Tableau 4 : Les descripteurs intervenant dans les modèles choisis.

Descripteurs	*classe*	*Signification*
J	Indice topologique	Indice de connectivité de distance de Balaban
Xt	Indice topologique	Indice de connectivité de structure totale.
X1sol	Indice de connectivité	Indice de connectivité de solvatation d'ordre 1.
Mor16m	Descripteur 3D-MoRSE	Signale 16 / pondéré par la masse.
Gm	Descripteur de WHIM	Indice de symétrie totale / pondéré par la masse.
RDF060p	descripteur RDF	Fonction de distribution radiale - 060 / pondérée par la polarisabilité.
Ss	Descripteur constitutionnel	Somme des états électrotopologiques de Kier-Hall.
Mom Dipo	Descripteur électronique 3D	Comportement et orientation d'une molécule dans un champ électrostatique.
ITH	descripteur GETAWAY	Contenu total de l'information sur l'égalité de levier.

Mom Dipo: Moment dipolaire

Conclusion

Nous avons relié cinq propriétés physico-chimiques des 16 HAPs prioritaires, à des descripteurs moléculaires. Les modèles RSP ont été établis en utilisant l'analyse de régression multilinéaire associant l'approche algorithme génétique pour la sélection de sous ensembles de variables significatives en maximisant la valeur du coefficient de prédiction Q^2_{LOO} . La taille de chaque modèle a été fixée à 1 et 2 descripteurs. A chaque fois, la qualité de l'ajustement a été vérifiée en procédant à la représentation des valeurs calculées en fonction de celles observées. Les grandes valeurs de Q^2 obtenues reproduisent pratiquement celles du coefficient de détermination multiple correspondant, ce qui fait ressortir la qualité de modèle obtenu. Le test de randomisation montre, dans tous les cas, que seul le vecteur réel des observations conduit à des valeurs élevées des statistiques R^2 et Q^2, ce qui prouve que les modèles obtenus ne sont pas dus au hasard. Dans les parties qui suivent nous avons augmenté la gamme pour chaque propriété dans le but de développer des modèles puissants et prédictifs.

Référence

US-EPA Report, EPA-600/8-90/-003, 1990.

APPLİCATİON DE L'APPROCHE QSPR DANS LA MODÉLİSATİON DE LA TEMPÉRATURE D'ÉBULLİTİON DE HAP

La température d'ébullition de 61 hydrocarbures aromatiques polycycliques a été modélisée par régression multilinéaire (MLR). Le modèle a été validé par validations interne et externe, puis appliqué sur un jeu de données constitué de 57 nouveaux HAP dont on ne connaît pas les valeurs des températures d'ébullition.

1. Collecte de données

Les valeurs expérimentales des températures d'ébullition comprises entre 491 et 869 K ont été prélevées dans (Bjorseth, 1983; Karcher *et al*., 1988). Les numéros de CAS des composés étudiés, ainsi que les valeurs expérimentales de Teb, sont réunis dans le tableau 5.

2. Division des données

Trois modes de fractionnement ont été appliqués dans ce travail: (a) choix aléatoire, (b) choix selon l'ordre de réponse (c) similitude structurelle ordonnée par le premier axe de l'analyse en composantes principales. Dans cette partie nous présenterons les résultats relatifs au choix aléatoire. Quant aux résultats associés aux choix (b) et (c) ils ont été réunis dans l'annexe I (page 177).

3. Résultats et discussion

Nous avons opté pour le modèle avec les descripteurs les plus significatifs, dont l'équation est:

$$Teb = 509{,}51(\pm 28{,}04) + 28{,}213(\pm 3{,}091)\ \boldsymbol{HOMO} + 34{,}101(\pm 0{,}283)\ \boldsymbol{EPS0} \qquad (96)$$

$\boldsymbol{N_{tr}}$ = 42, $\boldsymbol{R^2}$= 99,83%, $\boldsymbol{Q^2_{LOO}}$ = 99,80%, $\boldsymbol{R^2_{ext}}$ = 99,77%, $\boldsymbol{Q^2_{LMO30\%}}$ = 99,80%, $\boldsymbol{Q^2_{F1}}$ = 99,79%,

$\boldsymbol{Q^2_{F2}}$= 99,72%, $\boldsymbol{Q^2_{F3}}$ = 99,74%, $\boldsymbol{CCC_{ext}}$ = 99,86%, $\boldsymbol{RMSE_{tr}}$ = 4,42,

$\boldsymbol{RMSE_{cv}}$ = 4,77, $\boldsymbol{RMSE_{pr}}$ =5,42, S = 4,58.

Les paramètres statistiques montrent une forte corrélation entre les 2 variables (descripteurs) sélectionnées et la propriété étudiée, caractérisée par des paramètres statistiques excellents, en plus d'une très grande valeur du F de Fisher (= 11423,29), ce qui indique l'excellence du modèle dans la prédiction des valeurs du point d'ébullition, et une valeur minimale de l'erreur standard (s = 4,587). La valeur de R^2_{adj} = 99,82% indique un excellent accord entre la corrélation et la variation des données, alors que la faible valeur de R^2ys assure que le modèle obtenu n'est pas dû au hasard. Tous les paramètres statistiques du modèle sont

satisfaisants et on peut conclure que celui-ci est stable, robuste et prédictif. Le tableau 5 montre les résultats de la prédiction par le modèle développé et par le logiciel EPIsuite .

Tableau.5 Numéro de CAS des composés étudiés. Valeurs expérimentales et prédites des points d'ébullition. Pour les noms se reporter à l'annexe I.

CAS	Exp. Teb (K)	Préd.par le modèle.	ei (1)*	Prédite par EPIsuite	ei (2)*	CAS	Exp. Teb (K)	Préd.par le modèle.	ei (1)*	Prédite par EPIsuite	ei (2)*
000091-57-6	514	518,567	4,567	522,6	8,6	000205-99-2	754	751,9209	2,079	715,75	38,25
000090-12-0	518	521,892	3,892	522,6	4,6	000207-08-9	754	757,5619	3,561	715,75	38,25
000581-42-0	535	536,66	1,66	539,66	4,66	000192-97-2	769	761,1749	7,825	715,75	53,25
000582-16-1	535	535,390	0,390	539,66	4,66	000213-46-7	792	803,5801	11,58	743,09	48,91
000575-37-1	536	539,116	3,116	539,66	3,66	000050-32-8	769	767,6623	1,337	715,75	53,25
111495-85-3	538	539,426	1,426	539,66	1,66	000198-55-0	770	770,9932	0,993	715,75	54,25
000575-43-9	539	538,975	0,024	539,66	0,66	000053-70-3	808	804,6564	3,343	743,09	64,91
000581-40-8	541	537,733	3,266	539,66	1,34	000191-24-2	815	816,0489	1,048	759,31	55,69
000571-58-4	541	543,118	2,118	539,66	1,34	000191-07-1	863	861,1951	1,804	802,87	60,13
000571-61-9	542	541,877	0,123	539,66	2,34	000192-65-4	865	864,2395	0,760	786,65	78,35
000575-41-7	544	540,974	3,025	539,66	4,34	000189-55-9	867	866,7829	0,217	786,65	80,35
002245-38-7	558	557,803	0,196	555,81	2,19	000191-30-0	868	865,9605	2,039	786,65	81,35
000829-26-5	559	555,346	3,653	555,81	3,19	000189-64-0	869	871,8613	2,861	786,65	82,35
001430-97-3	591	593,149	2,149	580,25	10,75	000205-12-9	679	683,403	4,403	643,44	35,56
000832-71-3	625	619,092	5,907	612,75	12,25	000191-26-4	820	824,9121	4,912	759,31	60,69
002531-84-2	628	619,205	8,794	612,75	15,25	000224-41-9	804	803,9793	0,020	743,09	60,91
000883-20-5	628	618,699	9,300	612,75	15,25	000193-43-1	804	803,0724	0,927	759,31	44,69
000613-12-7	632	630,015	1,984	612,75	19,25	000193-39-5	807	813,2558	6,255	759,31	47,69
000832-69-9	632	620,730	11,26	612,75	19,25	000215-58-7	808	805,7525	2,247	743,09	64,91
000610-48-0	636	632,155	3,844	612,75	23,25	000205-82-3	753	758,2407	5,240	715,75	37,25
001576-67-6	636	636,796	0,796	624,49	11,51	027208-37-3	712	710,1049	1,895	673,85	38,15
003353-12-6	683	682,742	0,257	656,45	26,55	000091-20-3	491	500,7226	9,722	504,64	13,64
003442-78-2	683	680,010	2,989	656,45	26,55	000208-96-8	543	543,8939	0,893	547,85	4,85
002381-21-7	683	684,012	1,012	656,45	26,55	000083-32-9	552	557,0131	5,013	545,72	6,28
000243-17-4	675	678,097	3,097	643,44	31,56	000086-73-7	567	574,0128	7,012	565,57	1,43
000238-84-6	680	681,061	1,061	643,44	36,56	000085-01-8	611	600,9999	10,00	600,31	10,69
000203-12-3	705	700,536	4,463	688,41	16,59	000120-12-7	613	613,0736	0,073	600,31	12,69
000056-55-3	708	708,441	0,441	672,19	35,81	000203-64-5	632	628,6046	3,395	616,02	15,98
000217-59-4	712	699,467	12,53	672,19	39,81	000206-44-0	656	652,5746	3,425	644,85	11,15
000218-01-9	714	705,481	8,518	672,19	41,81	000129-00-0	666	664,1969	1,803	644,85	21,15
000092-24-0	723	719,562	3,437	672,19	50,81						

$MAE_{QSPR\ modèle}$= 3,541 MAE_{EPIWIN} =29,173

$RMSE_{QSPR\ modèle}$ =4,756 $RMSE_{EPIWIN}$ =37,647

* $e_{i\ (1)}$: Teb (Exp) – Teb prédite (par le modèle), $e_{i\ (2)}$: Teb (Exp) – Teb prédite par EPISUITE

La bonne qualité de l'ajustement, la robustesse et la capacité prédictive ont été confirmées par les valeurs de R^2, Q^2_{LOO}, R^2_{ext} et celles des erreurs ($RMSE_{tr}$, $RMSE_{cv}$, $RMSE_{pr}$) relativement faibles. De plus, la figure 15 qui reproduit les valeurs des températures d'ébullition prédites en fonction des valeurs observées montre une bonne corrélation entre les valeurs observées et prédites ce qui confirme la bonne qualité du modèle. Comme les valeurs des erreurs ($RMSE_{tr}$, $RMSE_{pr}$) sont proches et petites, nous pouvons conclure que le modèle

n'est pas surestimé. Cela signifie que le modèle prédit correctement non seulement pour les composés d'entraînement mais également pour d'autres composés (externes).

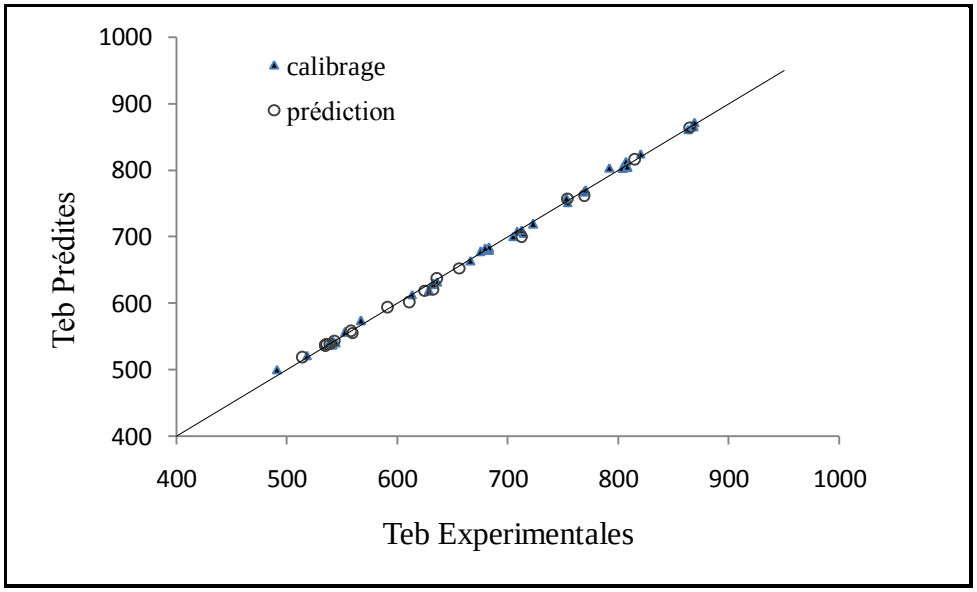


Figure .15 Graphe des valeurs Teb prédites en fonction des valeurs expérimentales.

La figure 16 présente les erreurs standardisées de prédiction en fonction des valeurs des leviers (*hii*). On constate que tous les résidus se situent dans la plage (± 3 SD) (lignes horizontales), Nous notons l'absence de point aberrant et /ou influent pour l'ensemble de calibrage et de prédiction (validation externe), *h** étant égal à 0,214, ce qui signifie que le modèle a une bonne capacité prédictive.

Ainsi, le modèle proposé pourrait être utilisé pour examiner les bases de données existantes ou des structures chimiques virtuelles pour identifier le point d'ébullition des HAP, le domaine d'application servira, alors, comme un outil précieux pour filtrer les structures chimiques dissemblables.

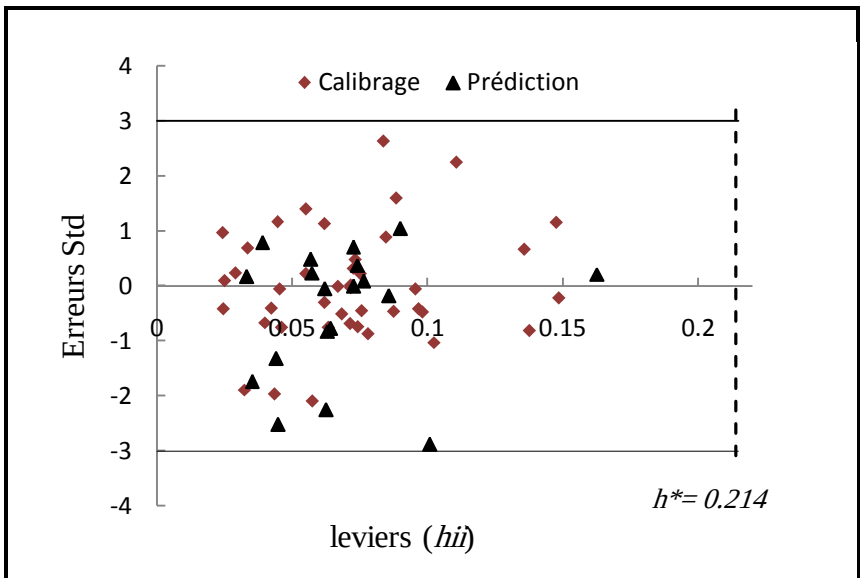


Figure .16 Diagramme de Williams: résidus standardisés en fonction des leviers. Les lignes pleines indiquent les limites ± 3SD, la ligne pointillée indique la valeur seuil h^* = 0,214.

Les résultats (R^2ys et Q^2ys en fonction de Kxy (Todeschini *et al.*, 1999)) sont reproduits dans la figure 17, où Kxy est la corrélation totale dans les variables du modèle (y inclus). Les petites valeurs de R^2ys (= 0,0492) et Q^2ys (= -1,1309) indiquent que les bons résultats du modèle original ne sont pas dus à une corrélation par chance ou à la dépendance structurelle de l'ensemble de calibrage (d'apprentissage).

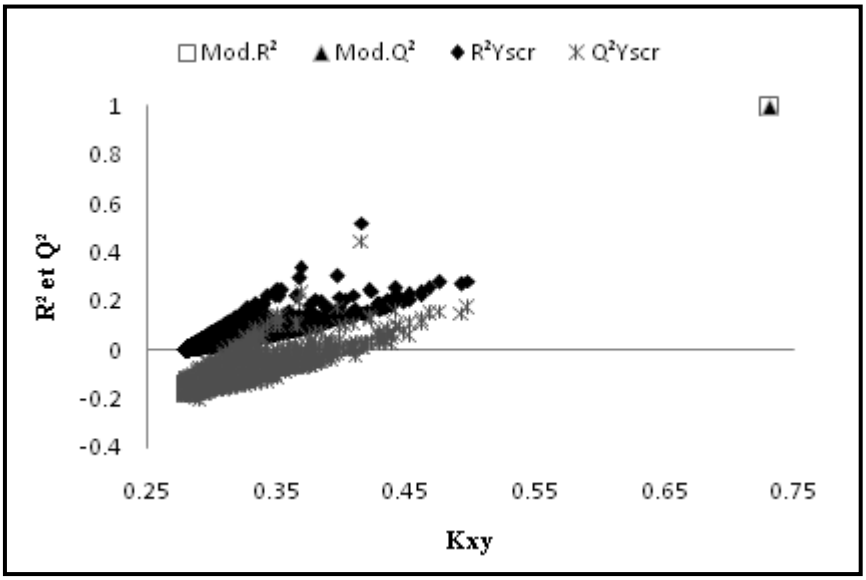


Figure.17 R^2 et Q^2 obtenus en utilisant des données de réponse permutées en fonction de Kxy.

Afin de vérifier si tous les produits chimiques de l'ensemble de prédiction (produits chimiques, qui n'ont pas de valeurs expérimentales pour la température d'ébullition) sont à l'intérieur du domaine du modèle, nous avons appliqué l'Insubria graph (Gramatica *et al.*, 2012). Le graphique de la figure 18 reproduit l'effet de levier pour l'ensemble de prédiction par rapport aux valeurs prédites. Avec le graphe Insubria, nous avons défini la zone de prédiction fiable du modèle, basée sur la similarité structurelle avec les composés d'entraînement (valeur de levier) et la valeur prédite du point d'ébullition. Nous supposons que les résultats prédits sont fiables, si les deux conditions: *hi* <*h** et *Ymin* <*Ypred* <*Ymax* (*Ymin* et *Ymax* sont la valeur minimale et la valeur maximale de Teb dans l'ensemble d'apprentissage). Nous avons trouvé que tous les composés de l'ensemble de prédiction étaient situés dans le domaine d'application du modèle. Ceci nous permet de dire que le modèle obtenu dans ce travail a une grande applicabilité pour de futures expériences pour combler le manque de données.

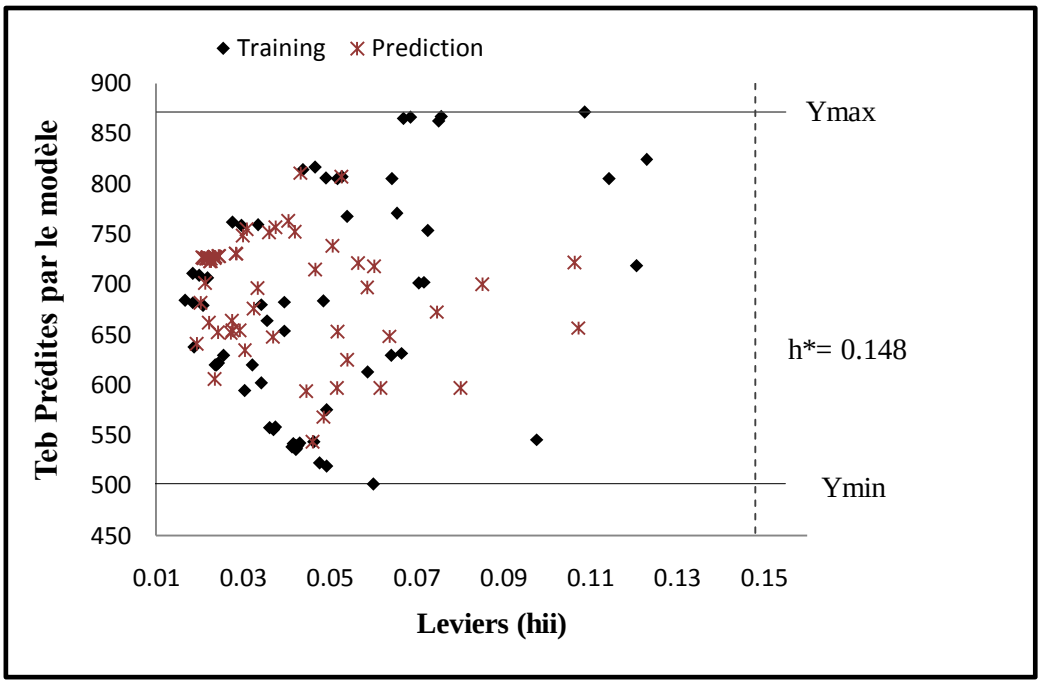


Figure .18 Graphe d'Insubria (graphique des valeurs de levier par rapport aux valeurs prédites pour l'ensemble des HAP).

3.1. Analyse de contributions des descripteurs et interprétation

Les contributions relatives des deux descripteurs ont été déterminées et reproduites dans la figure 19. L'importance des descripteurs impliqués dans le modèle diminue dans l'ordre suivant: EPS0 (indice de connectivité de bord d'ordre 0) (91,56%)> HOMO (8,43%). Il convient de noter que la différence dans la contribution des descripteurs au modèle est

significative, nous avons trouvé que le descripteur ***EPS0*** lui-même donnait un modèle à une variable ***Teb = 252,530 + 35,75EPS0*** avec **R^2= 99,71%** et une erreur standard **s = 7,1002 K** pour n_{tr} = 42. Cela signifie que le descripteur ***EPS0*** utilisé est un descripteur important pour décrire l'influence de la structure moléculaire sur la valeur la température d'ébullition des HAP. Cependant, un inconvénient de l'indice de connectivité de bord d'ordre 0 est sa dégénérescence, c'est-à-dire que les isomères ont des valeurs numériques identiques. Par conséquent, le modèle développé uniquement en utilisant l'indice de connectivité des arêtes, n'est pas suffisamment discriminant pour les HAP. Pour améliorer la description, un second régresseur, HOMO, a été ajouté comme indiqué ci-dessus (équation (96)).

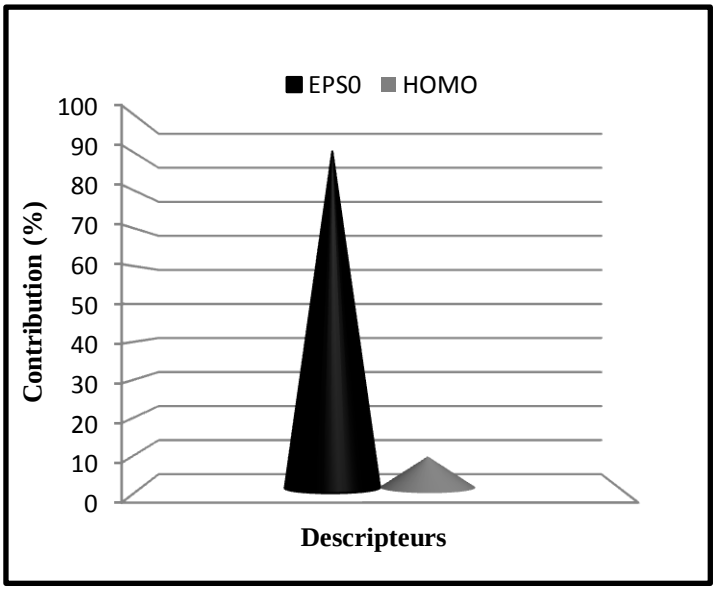


Figure. 19 Contribution relative des descripteurs sélectionnés.

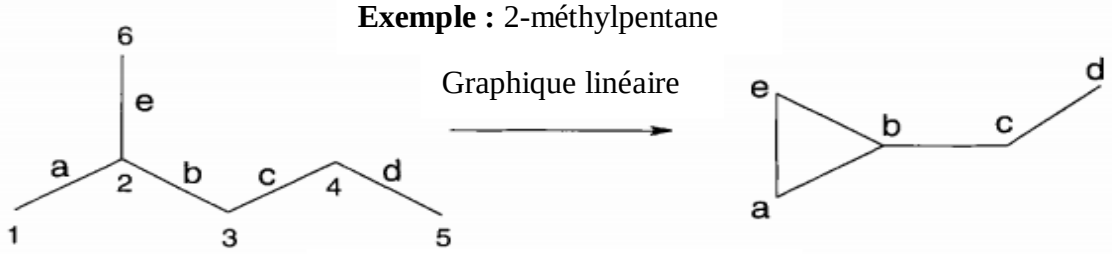
Selon White (White, 1986), les propriétés des HAP sont une fonction directe de leur taille et de leur topologie. La taille est une fonction du nombre d'électrons π, tandis que la topologie est liée aux systèmes d'anneaux annelés ou péri-condensés. La topologie est aussi une fonction de l'annélation linéaire et angulaire. La taille et la topologie des HAP affectent l'énergie (HOMO), qui à son tour est un prédicteur raisonnable de leurs propriétés.

L'importance du descripteur EPS0 (Estrada *et al.*, 1998) sur les valeurs des points d'ébullition est évidente, puisque le descripteur EPS0 contribue pour 91,568% dans le modèle (8,432% pour HOMO). Le descripteur EPS0 est fortement corrélé aux valeurs expérimentales du point d'ébullition (R = 0,997). Le coefficient et les valeurs positives de ce descripteur indiquent que les HAP avec des grandes valeurs de ce descripteur auront des points d'ébullition élevés. Ce type de descripteur moléculaire est défini comme suit:

$$\varepsilon(G) = \sum_S \left[\delta(e_i).\delta(e_j)\right]_\delta^{-1/2} \quad (105)$$

Où $\delta(e_i)$ degré du côté (e_i) ; la sommation porte sur toutes les paires d'arêtes adjacentes dans le graphe. Puisque le nombre de connexions est sensible aux différentes caractéristiques de la structure moléculaire telles que la taille, la ramification, la cyclicité et les liaisons multiples; ce paramètre moléculaire, *EPS0*, est connu sous l'appellation « indice de complexité moléculaire » (Bagheri *et al.*, 2013). Les graphes les plus compliqués se caractérisent par de grandes valeurs du descripteur *EPS0* et expliquent pourquoi les points d'ébullition sont grands; la figure 20 montre un exemple de calcul de EPS0.

Exemple : 2-méthylpentane



Matrice d'adjacence des arêtes E

Liaison	a	b	c	d	e	ε_i
a	0	1	0	0	1	2
b	1	0	1	0	1	3
c	0	1	0	1	0	2
d	0	0	1	0	0	1
e	1	1	0	0	0	2

$$A_E = \sum_{i=1}^{5}\sum_{j=1}^{5}[E]_{ij} = \sum_{i=1}^{5}\varepsilon_i = 2 \cdot N_2 = 10$$

$$\varepsilon = (\varepsilon_a \cdot \varepsilon_b)^{-1/2} + (\varepsilon_b \cdot \varepsilon_c)^{-1/2} + (\varepsilon_c \cdot \varepsilon_d)^{-1/2} + (\varepsilon_a \cdot \varepsilon_e)^{-1/2} + (\varepsilon_b \cdot \varepsilon_e)^{-1/2} =$$
$$= (2 \cdot 3)^{-1/2} + (3 \cdot 2)^{-1/2} + (2 \cdot 1)^{-1/2} + (2 \cdot 2)^{-1/2} + (3 \cdot 2)^{-1/2} = 2.432$$

Figure .20 Exemple de calcul de EPS0 pour le 2-méthylpentane.

3.2. Comparaisons aux modèles de la littérature

On peut faire une comparaison entre les modèles les plus importants modèles publiés (White, 1986; Ferreira, 2001; Todeschini *et al.*, 1995; Ribeiro et Ferreira, 2003) pour la

prédiction du point d'ébullition des HAP; le tableau 6 reproduit les résultats obtenus par différents auteurs.

Tableau 6. Comparaison entre les travaux antérieurs et ce travail pour le point d'ébullition

Works	N	taille du modèle	Type des descripteurs	R^2 (%)	R	Q^2_{LOO} (%)	$RMSE_{tr}$ $RMSE_{cv}$ ($RMSE_{pr}$)	S	Q^2_{F1}(%) Q^2_{F2}(%) Q^2_{F3}(%)
C.M.White[5]	47	1	The first-order valence molecular connectivity ^1X_v,	-	0,994	-	-	8.59	-
Todeschini R, *et al.*[9]	53	4	WHIM descriptors	95,9	-	95,0	(17,4)	-	-
M. Márcia *et al.* [8]	23	3	*EA, Xv, Log W*	-	0,999	-	-	6,68	-
		3	*EA, Xe, Log W*	-	0,999	-	-	5,52	-
		4	*EA, Xe, Xv, Log W*	-	0,999	-	-	5,70	-
		4	*EA, Xe, SArea, Log W*	-	0,999	-	-	4,40	-
C. Ferreira *et al.* [10]	36	3	Volume (*V*), molecular weight (*MW*) and Randic connectivity index(*R*)	PLS model: 99,5	-	99,42	(7,756)	-	-
				PCR Model: 99,5	-	99,38	(8,474)	-	-
Norte travail	61	2	*EPS0 HOMO*	99,83	0,998	99,80	4,420 4,770 (5,425)	4,587	99,79 99,72 99,74

Il est facile de vérifier si un modèle est meilleur que d'autres, en tenant compte de la taille de l'ensemble de données étudié, des paramètres statistiques pris en considération et de la complexité du modèle (c'est-à-dire le nombre de descripteurs impliqués et la méthode de modélisation utilisée). Le modèle développé dans ce travail comprend plus de composés et moins de descripteurs. En outre, il a été évalué avec davantage de paramètres statistiques comparativement à d'autres modèles de la littérature (white, 1986; Ferreira, 2001; Todeschini *et al.*, 1995; Ribeiro et Ferreira, 2003), qu'il dépasse en qualité et performance.

Le tableau 5 (page 130) présente les valeurs de Teb calculées en utilisant notre modèle et le modèle EPISuit (une série de programmes de propriétés physiques et chimiques et d'évaluation du devenir environnemental mis au point par EPA et Syracuse Research Corp. (SRC)) (https://www.epa.gov), en se basant sur l'erreur absolue moyenne (MAE) et l'erreur quadratique moyenne (RMSE) fournies par chacun des modèles. La comparaison tourne à l'avantage du modèle que nous avons développé. En effet, nous notons que :

$$MAE_{(notre\ modèle)} = 3,541 << MAE_{(EPIWIN)} = 29,173$$

$$RMSE_{(notre\ modèle)} = 4,756 << RMSE_{(EPIWIN)} = 37,647$$

Ce qui nous permet d'affirmer que le modèle développé est prédictif et utile pour le calcul du point d'ébullition des HAP.

Conclusion

Le point d'ébullition de 61 hydrocarbures aromatiques polycycliques a été corrélé avec leur structure moléculaire par l'approche QSPR en exploitant le logiciel QSARINS. Un modèle bidimensionnel a été calculé, à partir des ensembles d'apprentissage obtenus avec différentes procédures de fractionnement (*cf* : Annexe 1). Les valeurs très élevées des paramètres statistiques (Q^2_{LOO} , Q^2_{LMO} , Q^2_{F1} , Q^2_{F2} , Q^2_{F3} , CCC_{ext}), montrent que le modèle proposé est robuste et possède une bonne capacité prédictive, et peut être utilisé pour estimer le point d'ébullition des HAP sans données expérimentales connues. La validité des prédictions du modèle est également garantie par la vérification des Teb de 57 HAP sans valeurs expérimentales, en considérant celles appartenant au domaine d'applicabilité avec l'approche par effet de levier. Le modèle QSAR présenté dans notre travail a montré de meilleures valeurs des paramètres statistiques et de meilleurs résultats de prédiction comparativement à ceux obtenus auparavant par d'autres auteurs.

Références bibliographiques

Bagheri M, Borhani T N G, Zahedi G, Simple yet accurate prediction of liquid molar volume via their molecular structure. Fluid. Phase. Equilibr. 2013. 337, 183.

Bjorseth A, (Ed.), Handbook of Polycyclic Aromatic Hydrocarbons; Marcel Dekker: New York, 1983; Appendix.

Estrada E, Guevara N, Gutman I, Extension of edge connectivity index. Relationships to line graph indices and QSPR applications, J. Chem. Inform. Comput. Sci .1998. 38, 428 .

Ferreira M M C, polycyclic aromatic hydrocarbons: QSPR study, Chemosphere. 2001. 44, 125.

Https://www.epa.gov/tsca-screening-tools/epi-suitetm-estimation-program-interface.

Karcher W, Fordham R J, Dubois J J, P. Glaude G J M, Lighart J A M, Spectral Atlas of Polycyclic Aromatic Compounds, Kluwer Academic Publishers, Dordrecht, 1988.

Ribeiro F A L, Ferreira M M C, QSPR models of boiling point, octanol–water partition coefficient and retention time index of polycyclic aromatic hydrocarbons, J Mol Struc-Theochem. 2003. 663 , 109.

Todeschini R, Consonni V, Maiocchi A, The K correlation index: theory development and its application in chemometrics, Chemometr. Intell. Lab. 1999. 46, 13.

Todeschini R, Gramatica P, Provenazi R, Marengo E, Weighted holistic invariant molecular descriptors. Part 2. Theory development and applications on modeling physicochemical properties of polyaromatic hydrocarbons Chemometr. Intell. Lab. 1995. 27, 221.

White C M, Prediction of the boiling point, heat of vaporization, and vapor pressure at various temperatures for polycyclic aromatic hydrocarbons. J. Chem. Eng. Data. 1986.31, 198.

PRÉDİCTİON DE LA TEMPÉRATURE DE FUSİON DES HAP À L'AİDE DES MÉTHODES MLR ET RNA

Nous avons calculé deux modèles QSPR pour la température de fusion en utilisant la méthode de régression linéaire multiple (MLR), et une autre méthode non-linéaire par approcheRNA.

1. Résultats du modèle MLR

Les valeurs expérimentales de la température de fusion de 77 HAP prélevés dans (Todeschini *et al.*, 1995) ont été séparées par l'algorithme CADEX en deux sous-ensembles, respectivement de 55 composés pour le calibrage (calcul du modèle) et des 22 composés restants pour la prédiction (validation externe du modèle).

Une étape concerne la définition du nombre optimal de descripteurs à introduire dans le modèle final de sorte qu'il soit le plus robuste possible tout en évitant sa sur-paramétrisation (surestimation). La figure 21 présente l'influence du nombre de descripteurs (taille du modèle) sur les valeurs de Q^2, R^2 et $Q^2_{ext.}$

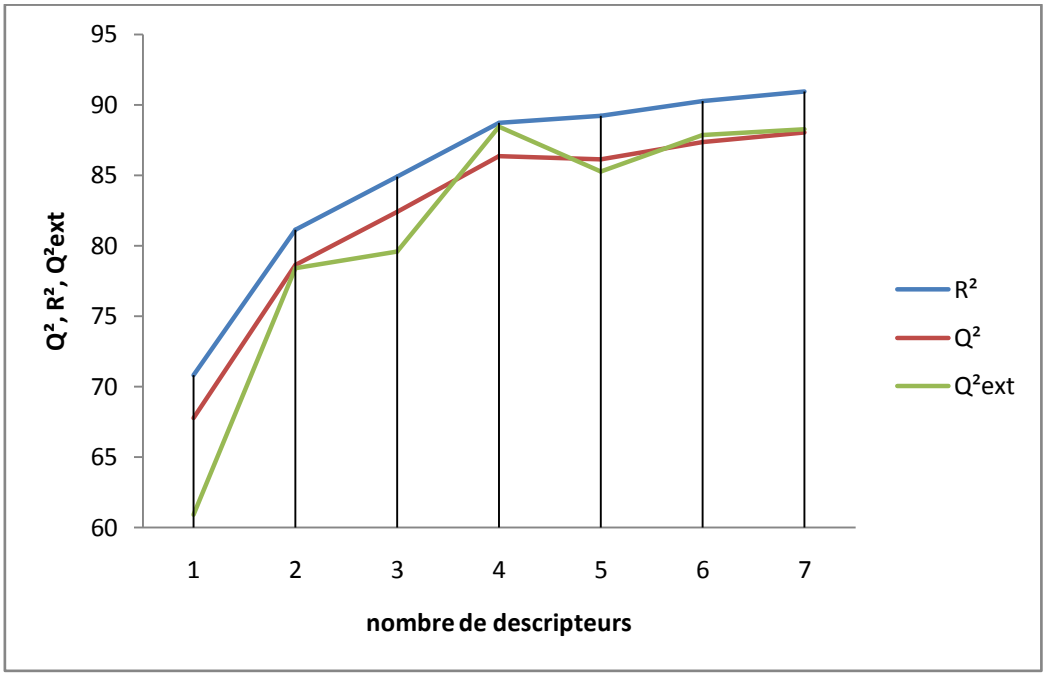


Figure.21 Variation de R^2,Q^2, Q^2_{ext} en fonction de la taille du modèle.

Comme le montrent la figure 21 et le tableau 7, il n'y a pas une différence significative entre les valeurs de Q^2, R^2 et Q^2_{ext} pour les modèles à 5,6 et 7 descripteurs. Cependant le modèle à 4 descripteurs présente la plus grande valeur de Q^2ext qui est d'ailleurs proche de celles de Q^2 et R^2. Aussi avons-nous opté pour le modèle de dimension 4.

Tableau.7 : Comparaison entre R^2, Q^2 et Q^2_{ext} des modèles de différentes tailles

Dimension	R^2	Q^2	Q^2ext	ΔQ^2	ΔR^2	ΔQ^2ext
1	70.84	67.79	60.93			
2	81.16	78.66	78.43	10.87	10.32	17.5
3	84.92	82.4	79.58	3.74	3.76	1.15
4	88.72	86.38	88.43	3.98	3.8	8.85
5	89.22	86.15	85.27	-0.23	0.5	-3.16
6	90.28	87.36	87.84	1.21	1.06	2.57
7	90.96	88.05	88.27	0.69	0.68	0.43

Où, k= {1 à 7}, ΔR^2, ΔQ^2, ΔQ^2_{ext} ont été calculés respectivement par l'application des formules suivantes

$$\Delta R^2 = R^2_{(k+1)} - R^2_{(k)} \quad (98\ a)$$

$$\Delta Q^2 = Q^2_{(k+1)} - Q^2_{(k)} \quad (98\ b)$$

$$\Delta Q^2_{ext} = Q^2_{ext\,(k+1)} - Q^2_{ext(k)} \quad (98\ c)$$

- $\boldsymbol{\Delta R^2}$: représente la différence entre deux coefficients de déterminations de dimensions successives (k+1 et k).
- $\boldsymbol{\Delta Q^2}$: représente la différence entre deux coefficients de prédictions de dimensions successives.
- $\boldsymbol{\Delta Q^2 ext}$: représente la différence entre deux coefficients de validation externe de dimensions successives.

2. Equation et analyse de régression :

L'équation optimale du modèle a la forme suivante :

$$\boldsymbol{Tfus(k)} = -567(\pm 98{,}841) + 122\,(\pm 10{,}305)\boldsymbol{AMW} + 2{,}99(\pm 0{,}413)\boldsymbol{TIC1} - 342(\pm 41{,}685)\boldsymbol{SIC4} + 281(\pm 46{,}048)\boldsymbol{P1m} \quad (\mathbf{99})$$

$\boldsymbol{N_{tr}}$ = 55, $\boldsymbol{R^2}$= 88,72%, $\boldsymbol{Q^2_{LOO}}$ = 86,38%, $\boldsymbol{R^2_{ext}}$ = 82,49%, $\boldsymbol{Q^2_{LMO30\%}}$ = 85,90%, $\boldsymbol{Q^2_{F1}}$ = 89,84%, $\boldsymbol{Q^2_{F2}}$= 77,92%, $\boldsymbol{Q^2_{F3}}$ = $\boldsymbol{Q^2}_{\text{ext}}$ = 88,43%, $\boldsymbol{CCC_{ext}}$ = 86,64%, $\boldsymbol{RMSE_{tr}}$ =34,88,

$\boldsymbol{RMSE_{pr}}$ = 35,32, S = 36,58.

Tableau.8 Noms, valeurs mesurées et prédites de la température de fusion ; valeurs des descripteurs calculés du modèle.

N	NOM	Tfus (k) exp	Prédite par MLR	Prédite par RNA	AMW	TIC1	SIC4	P1m
1	1-methylnaphthalene	251	261,500	281,602	6,770	46,181	0,905	0,626
2	1-ethylnaphthalene	259	248,379	275,695	6,510	57,126	0,902	0,572
3	2,3,5-trimethylnaphthalene	298	310,052	297,063	6,310	65,124	0,827	0,702
4	1-phenylnaphthalene	318	355,686	318,100	7,30	44,567	0,911	0,756
5	9-methylfluorene	320	386,263	343,257	6,930	62,211	0,797	0,699
6	4-methylphenanthrene	323	361,380	323,160	7,120	58,568	0,895	0,686
7	1,5-dimethylnaphthalene	353	335,152	324,430	6,510	57,510	0,695	0,625
8	1-methylfluorene	360	365,315	363,547	6,930	64,544	0,928	0,759
9	9-methylphenanthrene	364	364,367	315,191	7,120	58,568	0,869	0,665
10	Acenaphthylene	366	366,500	368,434	7,610	40,929	0,792	0,554
11	2,7-dimethylnaphthalene	370	374,163	355,830	6,510	57,510	0,714	0,787
12	Azulene	373	309,701	357,000	7,120	25,059	0,76	0,694
13	2-phenylnaphthalene	377	392,271	418,956	7,30	44,567	0,896	0,868
14	2,6-dimethylnaphthalene	383	389,644	391,497	6,510	57,510	0,695	0,819
15	Fluoranthene	384	426,377	399,785	7,780	40,263	0,771	0,675
16	4H-cyclopenta[def] phenanthrene	389	442,627	395,330	7,610	56,096	0,793	0,665
17	Fluorene	390	413,629	389,196	7,230	50,042	0,789	0,786
18	2-methylpyrene	417	459,662	445,047	7,460	63,558	0,796	0,715
19	4-methylpyrene	421	395,527	402,579	7,460	63,558	0,876	0,584
20	Benzo[ghi]fluoranthene	422	479,084	449,675	8,080	44,167	0,706	0,612
21	Pyrene	429	468,186	424,528	7,780	40,263	0,624	0,645
22	1-methylchrysene	434	451,460	476,668	7,340	70,093	0,919	0,818
23	Benz[a]anthracene	435	450,735	443,544	7,610	45,658	0,803	0,817
24	Indeno[1,2,3-cd]pyrene	436	489,075	472,464	8,130	53,715	0,875	0,730
25	Benzo[j]fluoranthene	439	454,561	423,649	7,890	49,961	0,870	0,745
26	Benzo[b]fluoranthene	441	468,656	437,920	7,890	49,961	0,823	0,738

Tableau.8 suite.

N	NOM	Tfus(k) exp	Prédite. par MLR	Prédite par RNA	AMW	TIC1	SIC4	P1m
27	Benzo[a]pyrene	450	460,460	443,864	7,890	49,961	0,870	0,766
28	Benzo[e]pyrene	452	449,209	469,414	7,890	49,961	0,741	0,569
29	3-methylcholanthrene	453	505,530	459,006	7,250	93,775	0,934	0,816
30	9,10-dimethylanthracene	456	475,786	471,462	6,880	71,697	0,596	0,694
31	Benzo[a]fluorene	463	440,433	456,185	7,460	66,443	0,972	0,830
32	Triphenylene	472	474,468	467,344	7,610	45,658	0,473	0,500
33	Dibenz[a,c]anthracene	478	495,968	496,672	7,730	55,511	0,691	0,685
34	2-methylanthracene	482	416,050	464,652	7,120	58,568	0,869	0,849
35	Benzo[b]fluorene	482	457,084	475,477	7,460	66,443	0,943	0,854
36	Anthracene	489	470,404	504,474	7,430	35,601	0,600	0,825
37	Aenzo[k]fluoranthene	490	498,875	497,879	7,890	49,961	0,788	0,803
38	Chrysene	529	473,547	471,783	7,610	45,658	0,742	0,824
39	Naphthacene	530	536,877	542,158	7,610	45,658	0,606	0,884
40	6-methylchrysene	530	451,123	497,805	7,340	70,093	0,883	0,773
41	Dibenzo[def,mno]chrysene	534	525,748	531,358	8,130	53,715	0,757	0,717
42	Dibenz[a,h]anthracene	543	529,866	541,921	7,730	55,511	0,734	0,858
43	Pentacene	544	590,803	553,300	7,730	55,511	0,605	0,918
44	Perylene	551	511,603	513,312	7,890	49,961	0,625	0,65
45	Benzo[ghi]perylene	556	519,535	567,616	8,130	53,715	0,656	0,572
46	Benzo[b]chrysene	567	507,705	552,000	7,730	55,511	0,812	0,874
47	Phenalene	358	307,466	364,225	7,230	60,532	0,962	0,507
48	2,6-dimethylanthracene	523	479,026	518,558	6,880	71,697	0,732	0,871
49	Dibenzo[a,i]anthracene	537	517,676	550,481	7,730	55,511	0,782	0,873
50	Hexaphene	581	575,254	579,994	7,820	65,258	0,725	0,866
51	Coronene	633	632,265	636,199	8,340	57,059	0,371	0,500
52	Indene	271	260,737	275,157	6,830	43,977	0,971	0,701
53	Ovalene	746	687,364	745,564	8,660	72,419	0,555	0,618
54	Quaterrylene	756	773,746	754,259	8,340	95,098	0,605	0,884

Tableau.8 suite.

N	NOM	Tfus(k) exp	Préd. par MLR	Prédi. par RNA	AMW	TIC1	SIC4	P1m
55	Dibenzo[a,e]pyrene	507	470,443	492,133	7,960	59,596	0,848	0,642
56	1,7-dimethylnaphthalene*	259	287,000	281,709	6,510	57,51	0,877	0,675
57	1,3,7-trimethylnaphthalene	287	311,387	301,515	6,310	65,124	0,842	0,725
58	2-ethylnaphthalene	266	327,560	325,722	6,510	57,126	0,884	0,832
59	1,2-dimethylnaphthalene	269	295,989	289,426	6,510	57,51	0,877	0,707
60	2-methylphenanthrene	329	387,713	407,037	7,120	58,568	0,91	0,798
61	3-methylphenanthrene	338	375,144	357,132	7,120	58,568	0,895	0,735
62	1-methylpyrene	343	418,840	384,976	7,460	63,558	0,876	0,667
63	Naphthalene	354	388,671	378,065	7,120	25,059	0,547	0,716
64	1-methylanthracene	359	392,409	398,433	7,120	58,568	0,879	0,777
65	Acenaphthene	369	338,456	362,181	7,010	53,088	0,755	0,54
66	2,3,6-trimethylnaphthalene	374	347,672	347,286	6,310	65,124	0,795	0,797
67	Phenanthrene	374	409,561	399,207	7,430	35,601	0,709	0,741
68	2-methylfluorene	377	380,572	354,213	6,930	64,544	0,945	0,834
69	3,6-dimethylphenanthrene	414	432,118	418,587	6,880	71,697	0,732	0,704
70	2,7-dimethylanthracene	514	471,494	513,516	6,880	71,697	0,745	0,860
71	Pentaphene	536	527,848	532,542	7,730	55,511	0,721	0,835
72	4-methylfluorene	344	353,237	343,081	6,930	64,544	0,928	0,716
73	3,4-benzofluorene	398	414,310	436,605	7,460	66,443	0,972	0,737
74	2-methylnaphthalene	308	310,028	295,382	6,770	46,181	0,883	0,772
75	1-methylphenanthrene	396	374,231	366,675	7,120	58,568	0,91	0,75
76	2,3-dimethylanthracene	525	473,408	520,169	6,880	71,697	0,732	0,851
77	3-methylfluorene	361	368,775	359,065	6,930	64,544	0,945	0,792

La Figure 22, reproduit les valeurs prédites des températures de fusion en fonction de celles mesurées. La dispersion des points autour de la première bissectrice montre que les valeurs prédites (pour l'ensemble de validation (prédiction)) et calculées (pour l'ensemble de calibrage) sont en adéquation avec les valeurs expérimentales. Les valeurs proches de Q^2_{LOO} et R^2 indiquent que pour la majorité des composés, le modèle proposé a un bon ajustement.

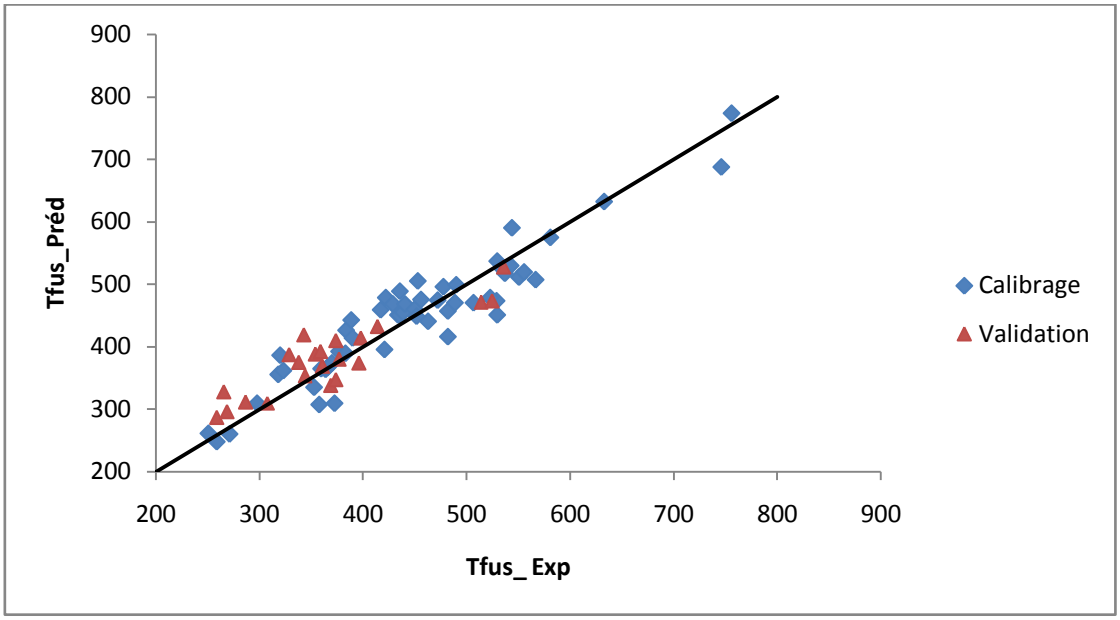


Figure. 22 Droite d'ajustement (Tfus prédites en fonction de celles observées).

Comme le montre le diagramme de Williams de la figure 23, la majorité des composés de l'ensemble de données se situent dans le domaine d'application, à l'exception du quaterrylène qui dépasse le seuil h^*= 0,273. Nous avons vérifié (en recalculant le modèle sans utiliser ce composé) que le quaterrylène a une influence positive sur la qualité du modèle.

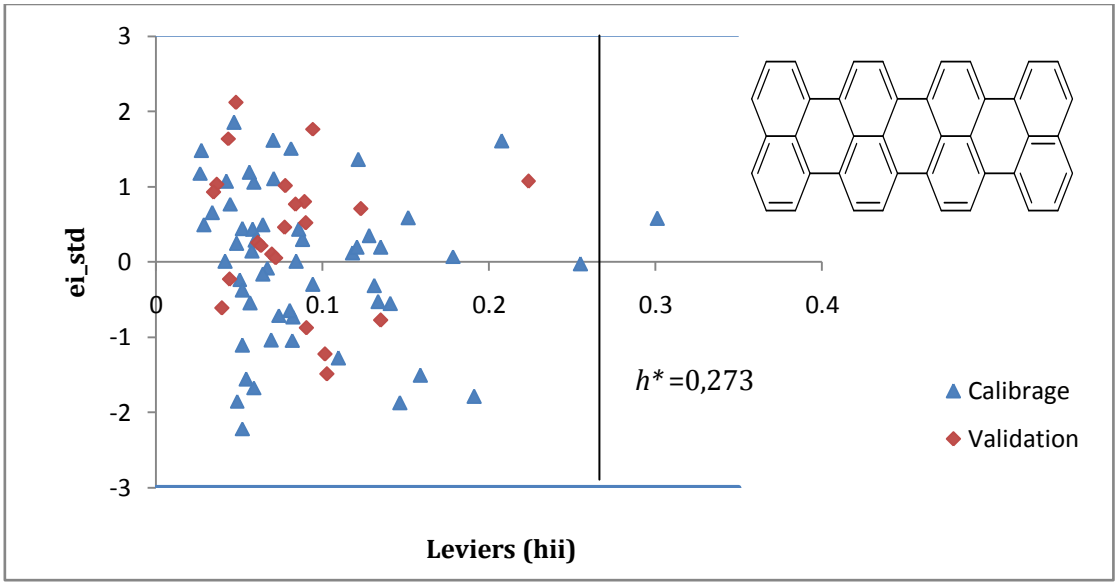


Figure.23 Diagramme de Williams.

Pour éviter les corrélations par chance et valider le modèle MLR calculé, le test de randomisation a été appliqué. Deux mille (2 000) modèles ont été développés, et la figure 24

montre les faibles valeurs de Q^2 et R^2 obtenues après chaque mélange (itération), faisant ressortir que les résultats du modèle original ne sont pas aléatoires.

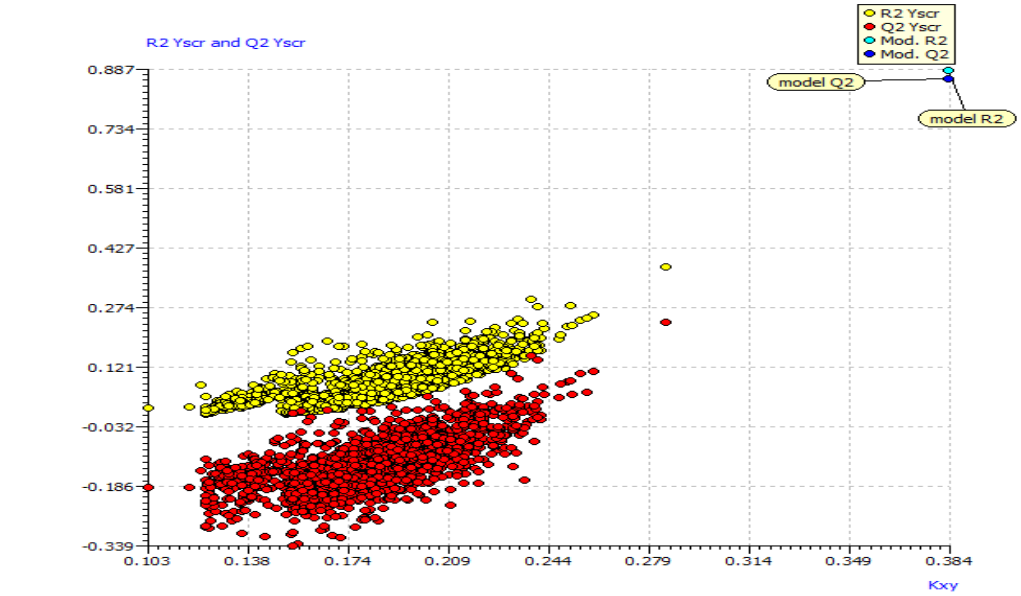


Figure. 24 Tests de randomisation.

3. Analyse et interprétation des contributions des descripteurs

Les contributions relatives des quatre descripteurs au modèle MLR ont été déterminées et reproduites sur la figure 25. L'importance des descripteurs impliqués dans le modèle MLR diminue dans l'ordre suivant: **AMW** (30,847%)> **SIC4** (25,318%)> **TIC1** (22,943%)> **P1m** (20,890%).

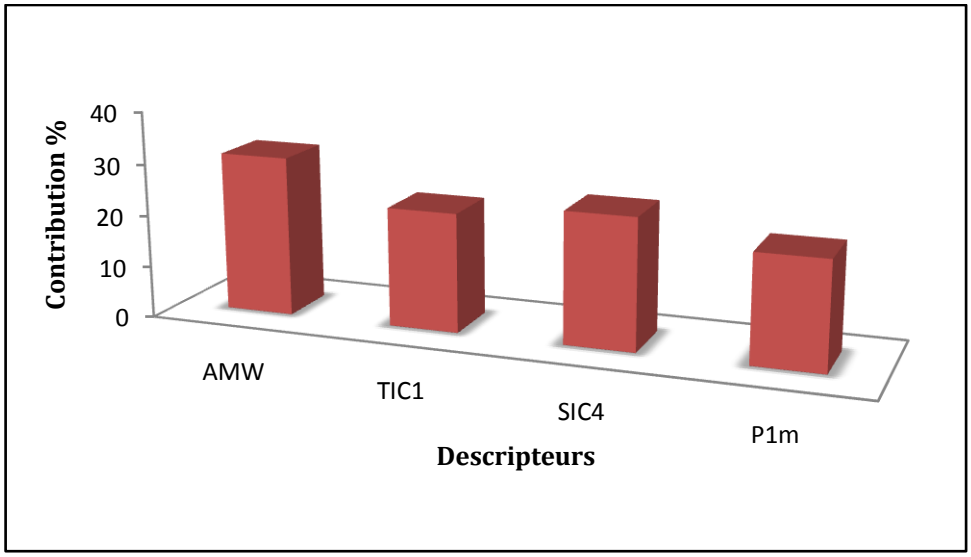


Figure. 25 Contribution (%) des descripteurs au modèle.

Le descripteur *AMW (= MW/ A* = Masse moléculaire /nombre d'atomes dans la molécule) qui est le plus corrélé (R= 0,732) avec la variable à expliquer (température de fusion) est le plus significatif dans le modèle calculé. C'est un descripteur constitutionnel définissant la masse moléculaire moyenne et comportant des informations sur la composition atomique.

Le descripteur ***TIC1*** présente une corrélation de (R = 0,351) avec la température de fusion. Ce descripteur est un indice d'information totale de la symétrie d'ordre 1.

Le contenu d'information totale du voisinage (***TICk***) est calculé par ***nAT*** fois ***ICk***, ***nAT*** étant le nombre total d'atomes de la molécule. Ce descripteur représente une mesure de la complexité du graphe moléculaire.

ICk Le contenu d'information de voisinage (***ICk***) est calculé comme: ''contenu d'information moyen'' selon la relation :

$$ICk = -\sum_{g=1}^{G} \frac{A_g}{nAT} \, log_2 \frac{A_g}{nAT} \qquad (100)$$

Où g parcourt les classes d'équivalence, A_g est la cardinalité (l'origine) de la classe d'équivalence g*th* et nAT est le nombre total d'atomes. Cet indice représente une mesure de la complexité structurelle par sommet.

La corrélation du troisième descripteur ***SIC4*** avec la température de fusion est de (R= - 0,566). Le coefficient de ce descripteur est négatif, ce qui indique que les valeurs de la température de fusion sont inversement liées à ce descripteur, par conséquence l'augmentation des valeurs de ce descripteur provoque une réduction des valeurs de la température de fusion. Ce descripteur appartient à la classe des indices d'information structurelle de la symétrie d'ordre 4.

Le contenu de l'information structurelle (***SICk)*** est calculé sous une forme normalisée du contenu d'information IC*k* pour supprimer l'influence de la taille du graphe:

$$\boldsymbol{SICk} = \frac{ICk}{log_2 nAT} \qquad (101)$$

Le quatrième descripteur ***P1m*** présente une corrélation avec la température de fusion de R = 0,266 c'est un descripteur WHIM signifie la 1^{ere} composante directionnelle de forme de l'indice de WHIM / pondérée par les masses atomiques (1st component shape directional

WHIM index / weighted by atomic masses). Les descripteurs WHIM sont construits de manière à capturer des informations moléculaires 3D pertinentes concernant la taille moléculaire, la forme, la symétrie et la distribution des atomes par rapport aux cadres de référence invariants.

4. Analyse des variables

La valeur du facteur d'inflation de la variance (VIF), est alors calculée selon l'équation de définition :

$$VIF = \frac{1}{1-r^2} \qquad (102)$$

Où r^2 est le coefficient de corrélation de l'équation de régression multiple entre les descripteurs du modèle. Si le VIF est égal à 1, aucune inter-corrélation n'existe pour chaque variable; Si le VIF est compris entre 1 et 5, le modèle associé est acceptable et si le VIF est supérieur à 10, le modèle associé est instable et une revérification est nécessaire (Jaiswal *et al*., 2004; Shapiro et Guggenheim, 1998) . La valeur du VIF de chaque descripteur (***AMW***, ***TIC1***, ***SIC4***, ***P1m***) utilisé dans le modèle MLR (tableau 9) est inférieure à 5 et proche de l'unité, ce qui indique une légère corrélation d'un descripteur donné avec les trois autres.

Tableau. 9 Caractéristiques des descripteurs du modèle

Predicteur	***Coef***	***SE Coef***	***T***	***P***	***VIF***
Constante	-567,370	98,840	-5,740	0.000	
AMW	121,770	10,310	11,820	0.000	1,161
TIC1	2,986	0,413	7,220	0.000	1,051
SIC4	-341,740	41,690	-8,200	0.000	1,195
P1m	280,880	46,050	6,100	0.000	1,072

La valeur de t pour un descripteur est liée à sa signification statistique. Les valeurs élevées de *t* en valeur absolues indiquent que chaque coefficient de régression est significativement plus grand que l'écart type associé. La probabilité de t d'un descripteur donne sa signification statistique lorsqu'il est combiné avec d'autres descripteurs dans le modèle QSPR (c'est-à-dire, renseigne sur les interactions entre descripteurs). Les descripteurs qui ont des valeurs de probabilités de t inférieures à 0,05 sont considérés comme statiquement significatifs pour un modèle donné, c'est-à-dire que leur influence sur la variable dépendante (la réponse) n'est pas due au hasard. Les valeurs des probabilités de t pour les quatre

descripteurs du modèle développé sont toutes égales à 0,000, ce qui indique que les descripteurs choisis sont très significatifs**.**

5. Résultats du modèle RNA

L'adaptabilité mathématique de la méthode RNA la recommande comme un outil puissant pour la classification des modèles et la construction de modèles prédictifs. Un avantage particulier de RNA est sa capacité inhérente à incorporer des dépendances non linéaires entre les variables dépendantes et indépendantes sans utiliser de fonction mathématique explicite. Dans cette étude, l'algorithme de rétro-propagation (BP-RNA) a été utilisé pour développer des modèles non linéaires. Les quatre descripteurs du modèle MLR ont été utilisés comme entrées dans le réseau.

La figure 26 reproduit les valeurs des paramètres statistiques en fonction du nombre de neurones de la couche cachée pour le modèle. Les valeurs des RMSE des ensembles d'apprentissage, de validation et de test sont proches et les plus petites lors de l'utilisation de cinq neurones dans la couche cachée.

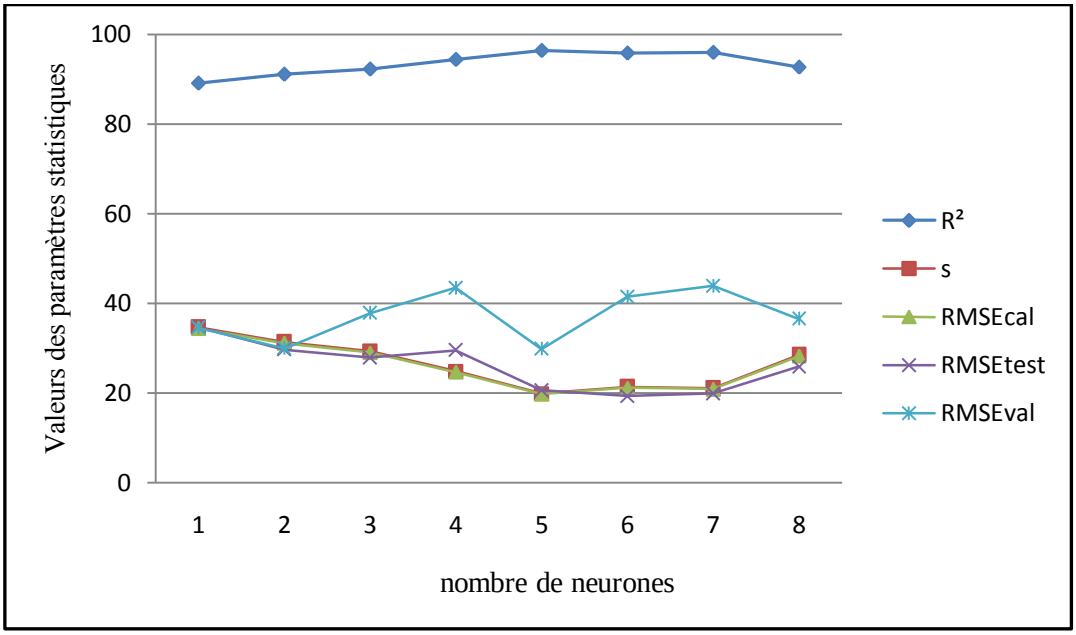


Figure .26 Différents paramètres statistiques en fonction du nombre de neurones dans la couche cachée.

Le nombre de neurones dans la couche cachée est un paramètre important qui influe sur les performances des modèles RNA. Le nombre de neurones cachés ne devrait pas être supérieur à 8 avec 55 échantillons dans l'ensemble d'apprentissage (Qi *et al.*, 2002). De

meilleurs résultats pourraient être obtenus en utilisant cinq neurones cachés après optimisation de l'architecture réseau par rapport au nombre de neurones cachés. Ainsi, une architecture (4-5-1) a été choisie, avec comme résultats statistiques un R^2 = 96,387%, $RMSE_{ext}$= 29,808, $RMSE_{test}$= 20,559, $RMSE_{cal}$=19,742, et s = 19,878 pour l'ensemble d'apprentissage.

La qualité de l'ajustement a été vérifiée par la représentation des valeurs prédites de la température de fusion en fonction de celles expérimentales. La figure 27 montre une faible dispersion des points autour de la première bissectrice, ce qui indique la bonne concordance entre ces valeurs (R^2_{train}=0,964, R^2_{val}= 0,87 et R^2_{test} =0,951).

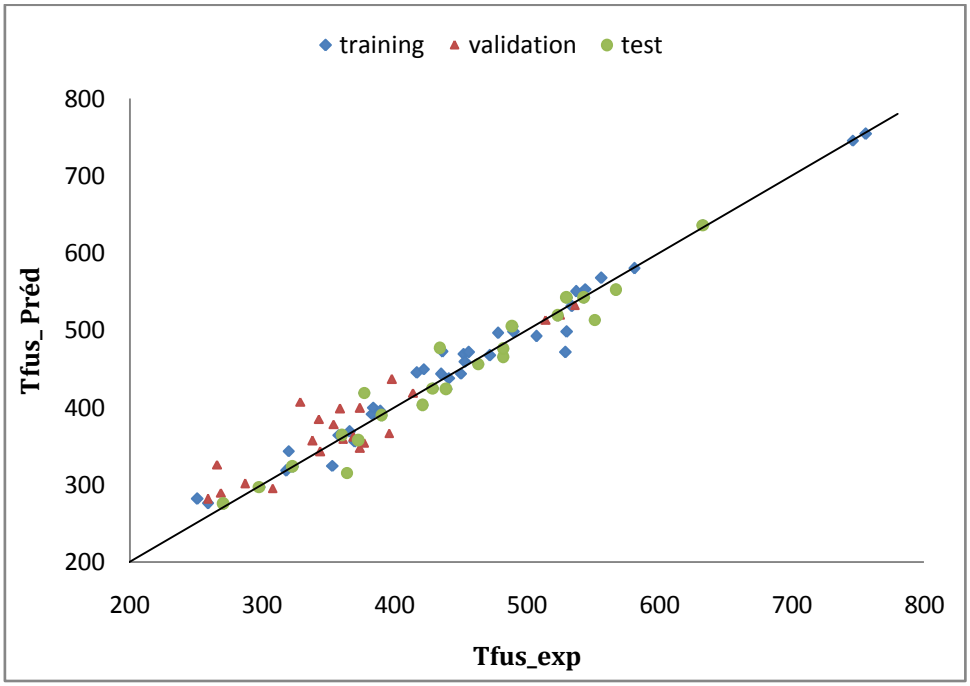


Figure.27 Diagramme des valeurs prédites pour les ensembles de calibrage (training), validation et test en fonction des valeurs expérimentales.

Les résultats de prédiction du modèle RNA pour l'ensemble des données sont reproduits dans le tableau 8 et la figure 27. On note une certaine modification par rapport aux résultats du modèle MLR, ce qui confirme la relation non linéaire entre l'information structurelle et les valeurs des températures de fusion des composés. Le modèle RNA proposé qui satisfait les conditions de l'ensemble de test est prédictif:

r^2 = 0.8713 r_0^2 = 0,9765 $r_0'^2$ = 0,9755

$(r^2 - r_0^2)/r^2$ = -0,1208 $(r^2 - r_0'^2)/r^2$ = -0,1196

$R^2_{cv,ext}$ = 0,8471 0,85 ≤ k = 1,0274 ≤ 1,15 0,85 ≤ k' = 0,9682 ≤1,15

6. Comparaison MLR et RNA

Nous avons procédé à une comparaison entre les résultats obtenus par les deux méthodes. La figure 28 établit que les performances des deux méthodes sont en général bonnes mais avec un avantage pour le modèle non linéaire. Sur la base des résultats obtenus pour les des deux modèles, on peut dire que la technique des réseaux de neurones artificiels donne de meilleurs résultats que la MLR.

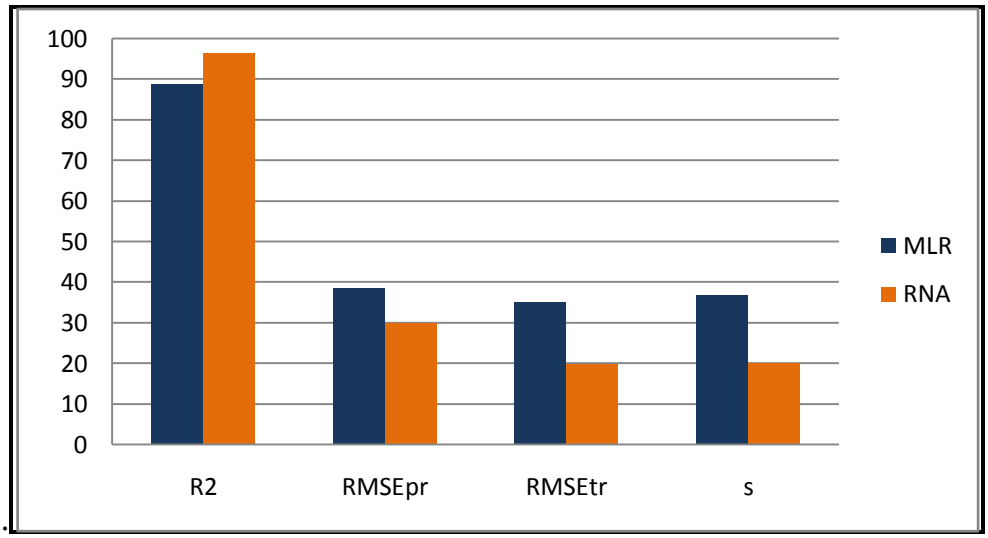


Figure.28 Comparaison des performances des modèles MLR et RNA.

Conclusion

Les Méthodes MLR et RNA exploitées pour le calcul de modèles (linéaire et non linéaire) de la température de fusion d'une série de HAP. Les deux méthodes semblent être utiles, bien que leur comparaison soit à l'avantage de la RNA. La supériorité des résultats de la RNA indique que la température de fusion des HAP possède certaines caractéristiques non linéaires. La méthode MLR est une technique appropriée pour choisir les entrées pour la modélisation RNA et aussi plus puissante dans la sélection des paramètres importants, les résultats de ce travail montrent que l'introduction des réseaux de neurones améliore la qualité de la prédiction de la température de fusion.

Références bibliographiques

Jaiswal M, Khadikar P V, Scozzafava A, Supuran C T, Carbonic anhydrase inhibitors: the first QSAR study on inhibition of tumor-associated isoenzyme IX with aromatic and heterocyclic sulfonamides. Bioorg. Med. Chem. Lett. 2004. 14, 3283.

Qi Y H, Zhang Q Y, Xu L, Correlation analysis of the structures and stability constants of gadolinium (III) complexes. J. Chem. Inf. Comput Sci. 2002, 42, 1471.

Shapiro S, Guggenheim B, Inhibition of oral bacteria by phenolic compounds. Part 1. QSAR analysis using molecular connectivity. Quant. Struct. Act. Relat. 1998. 17, 327.

Todeschini R, Gramatica P, Provenzani R, Marengo E, Weighted holistic invariant molecular descriptors. Part 2. Theory development and applications on modeling physicochemical properties of polyaromatic hydrocarbons, Chemometr Intell Lab. 1995, 27, 221.

MODÉLISATION DE LA SOLUBILITÉ AQUEUSE

L'une des principales caractéristiques des Hydrocarbures Aromatiques Polycycliques est leur faible solubilité aqueuse (hydrophobie) qui, avec leur coefficient de partage octanol-eau élevé, est essentielle pour évaluer et modéliser leur devenir et leur distribution dans l'environnement.

Les logarithmes des valeurs mesurées de la solubilité aqueuse (tableau 10) de 72 HAP, dont les structures moléculaires sont reproduites dans la figure 29, ont été extraites de (Ruelle et Kesselring, 1997). L'ensemble des données a été, au préalable, séparé à l'aide de l'algorithme Cadex (Kennard & stone) en deux ensembles disjoints de calibrage (51 composés) et de validation externe (21 composés).

Il y'a peu de modèles plusieurs modèles pour la prédiction de la solubilité aqueuse des HAP. Parmi les tentatives de modélisation publiées citons; (Ferreira, 2001; Gui-Ning Lu *et al.*, 2007; Qi Jun *et al.*, 2010, 2011) ont étudié cette propriétés par approche QSPR. Ferreira (2001) a utilisé la méthode PLS pour développer un modèle en utilisant 5 descripteurs, tandis que Gui-Ning Lu *et al.*, (2007) ont utilisé la même méthode et 11 descripteurs pour 31 molécules. Plus tard, Qi Jun *et al.*, (2010) ont utilisé deux méthode différentes (machine à support vecteur et réseaux de neurones à base radiale) et un jeu de données de 46 molécules et 6 descripteurs et comparé les résultats ainsi obtenus; en 2011 les mêmes auteurs ont comparé les résultats du travail précédent avec les résultats obtenus avec une autre méthode (moindres carrés partiels) en gardant le même ensemble de données. Cependant, ces modèles n'ont pas été validés par un jeu de validation externe.

Le but de ce travail est de développer un modèle QSPR simple et robuste en vue de prédire la solubilité aqueuse pour un ensemble de HAP en utilisant une méthode simple (régression linéaire multiple), donc établir une relation linéaire entre des descripteurs moléculaire et la solubilité.

Tableau.10 Noms, valeurs expérimentales et prédites de la solubilité ainsi que les valeurs des descripteurs du modèle.

ID	Nom	Statut	-Log S exp	-Log S pred	Km	BIC0	qpmax	ID	Nom	Statut	-LogS exp	LogS pred	Km	BIC0	qpmax
1	Indan	Tr	3.040	2.601	0.554	0.221	0.105	37	1,2,3,6,7,8-Hexahydropyrene	Tr	5.960	6.261	0.482	0.187	0.105
2	1-Ethylnaphthalene	Tr	4.170	3.666	0.495	0.203	0.127	38	Benzo[e]pyrene	Tr	7.600	7.839	0.500	0.171	0.111
3	1,4-Dimethylnaphthalene	Tr	4.140	4.238	0.495	0.203	0.112	39	Benzo[a]pyrene	Tr	7.800	8.082	0.649	0.171	0.111
4	1,4,5-Trimethylnaphthalene	Tr	4.910	4.863	0.493	0.197	0.113	40	6-Methylbenzo[a]pyrene	Tr	8.530	7.808	0.551	0.171	0.114
5	1,2,3,4-Tetrahydronaphthalene	Tr	4.370	3.764	0.563	0.211	0.104	41	Indeno[1,2,3,cd]pyrene	Tr	9.160	8.886	0.595	0.164	0.108
6	2-Methylnaphthalene	Tr	3.770	3.989	0.658	0.209	0.108	42	Chrysene	Tr	8.060	7.446	0.736	0.178	0.111
7	Biphenyl	Tr	4.310	4.473	0.752	0.205	0.111	43	6-Methylchrysene	Tr	6.570	6.681	0.664	0.178	0.128
8	4-Methylbiphenyl	Tr	4.620	5.123	0.810	0.200	0.111	44	5,6-Dimethylchrysene	Tr	7.010	6.774	0.608	0.176	0.129
9	4,4'-Dimethylbiphenyl	Tr	6.020	5.780	0.849	0.195	0.110	45	Triphenylene	Tr	6.740	7.061	0.500	0.178	0.111
10	Diphenylmethane	Tr	4.080	4.995	0.802	0.200	0.114	46	Cholanthrene	Tr	7.850	7.569	0.677	0.175	0.114
11	Diphenylethane	Tr	4.620	5.325	0.617	0.195	0.112	47	3-Methylcholanthrene	Tr	7.920	7.795	0.724	0.174	0.113
12	trans-Stillbene	Tr	5.800	5.690	0.862	0.196	0.110	48	Naphthacene	Tr	8.600	7.707	0.826	0.178	0.108
13	Acenaphthylene	Tr	3.960	4.354	0.500	0.201	0.115	49	Benzo[g,h,i]perylene	Tr	8.700	8.617	0.500	0.164	0.111
14	Fluorene	Tr	5.000	5.172	0.679	0.199	0.107	50	Coronene	Tr	9.330	9.623	0.500	0.157	0.105
15	1-Methylfluorene	Tr	5.220	5.401	0.638	0.196	0.108	51	Picene	Tr	7.870	8.615	0.795	0.168	0.112
16	Benzo[b]fluorene	Tr	8.040	7.002	0.781	0.183	0.110	52	1,5-Dimethylnaphthalene	Pr	4.740	4.238	0.495	0.203	0.112
17	4-Methylenephenanthrene	Tr	5.240	5.328	0.533	0.191	0.120	53	1-Methylnaphthalene	Pr	3.700	3.574	0.497	0.209	0.112
18	Anthracene	Tr	6.350	5.897	0.738	0.193	0.108	54	Acenaphthene	Pr	4.630	4.353	0.494	0.204	0.106
19	2-Methylanthracene	Tr	6.960	6.289	0.774	0.190	0.108	55	10-Methylbenzo[a]anthracene	Pr	6.640	7.332	0.736	0.178	0.114
20	9, 10-Dimethylanthracene	Tr	6.570	6.128	0.541	0.187	0.111	56	2-Ethylnaphthalene	Pr	4.290	4.725	0.747	0.203	0.110
21	Benzo[a]anthracene	Tr	7.210	7.314	0.725	0.178	0.114	57	1,3-Dimethylnaphthalene	Pr	4.290	4.276	0.495	0.203	0.111
22	7-Methylbenzo[a]anthracene	Tr	7.350	7.176	0.664	0.178	0.115	58	5-Methylchrysene	Pr	6.590	7.206	0.659	0.178	0.114
23	9-Methylbenzo[a]anthracene	Tr	6.560	7.432	0.774	0.178	0.113	59	Naphthalene	Pr	3.600	3.300	0.574	0.215	0.105
24	7-Ethylbenzo[a]anthracene	Tr	6.800	6.789	0.617	0.176	0.129	60	2,3-Dimethylnaphthalene	Pr	4.720	4.636	0.646	0.203	0.108
25	10-Ethylbenzo[a]anthracene	Tr	6.780	7.525	0.718	0.176	0.114	61	2,6-Dimethylnaphthalene	Pr	4.890	4.808	0.728	0.203	0.107
26	10-Butylbenzo[a]anthracene	Tr	7.550	7.974	0.721	0.172	0.114	62	Benzo[a]fluorene	Pr	6.680	7.020	0.745	0.183	0.108
27	12-Butylbenzo[a]anthracene	Tr	7.520	7.334	0.445	0.172	0.119	63	Phenanthrene	Pr	5.260	5.577	0.612	0.193	0.111
28	10-Pentylbenzo[a]anthracene	Tr	8.570	8.234	0.744	0.170	0.114	64	1-Methylphenanthrene	Pr	5.850	5.894	0.625	0.190	0.112
29	9,10-Dimethylbenzo[a]anthracene	Tr	6.770	7.643	0.767	0.176	0.113	65	2-Methylphenanthrene	Pr	5.840	6.049	0.697	0.190	0.111
30	4,5-Dimethylenebenzo[b]anthracene	Tr	7.960	7.407	0.665	0.178	0.109	66	4,5-Dimethylenephenanthrene	Pr	5.240	6.211	0.500	0.186	0.110
31	Dibenzo[a,h]anthracene	Tr	8.660	8.526	0.787	0.168	0.114	67	9-Methylanthracene	Pr	5.890	5.945	0.633	0.190	0.111
32	Fluoranthene	Tr	6.000	6.679	0.513	0.183	0.107	68	1-Methylbenzo[a]anthracene	Pr	6.640	7.047	0.678	0.178	0.119
33	Benzo[b]fluoranthene	Tr	8.230	8.052	0.607	0.171	0.110	69	12-Methylbenzo[a]anthracene	Pr	6.680	7.227	0.695	0.178	0.115
34	Benzoljlfluoranthene	Tr	8.000	7.699	0.465	0.169	0.119	70	7,12-Dimethylbenzo[a]anthracene	Pr	7.020	7.318	0.638	0.176	0.116
35	Benzo[k]fluoranthene	Tr	8.490	8.211	0.705	0.171	0.110	71	Dibenzo[a,j]anthracene	Pr	7.370	8.133	0.686	0.168	0.120
36	Pyrene	Tr	6.190	6.696	0.500	0.183	0.106	72	Perylene	Pr	8.800	7.839	0.500	0.171	0.111

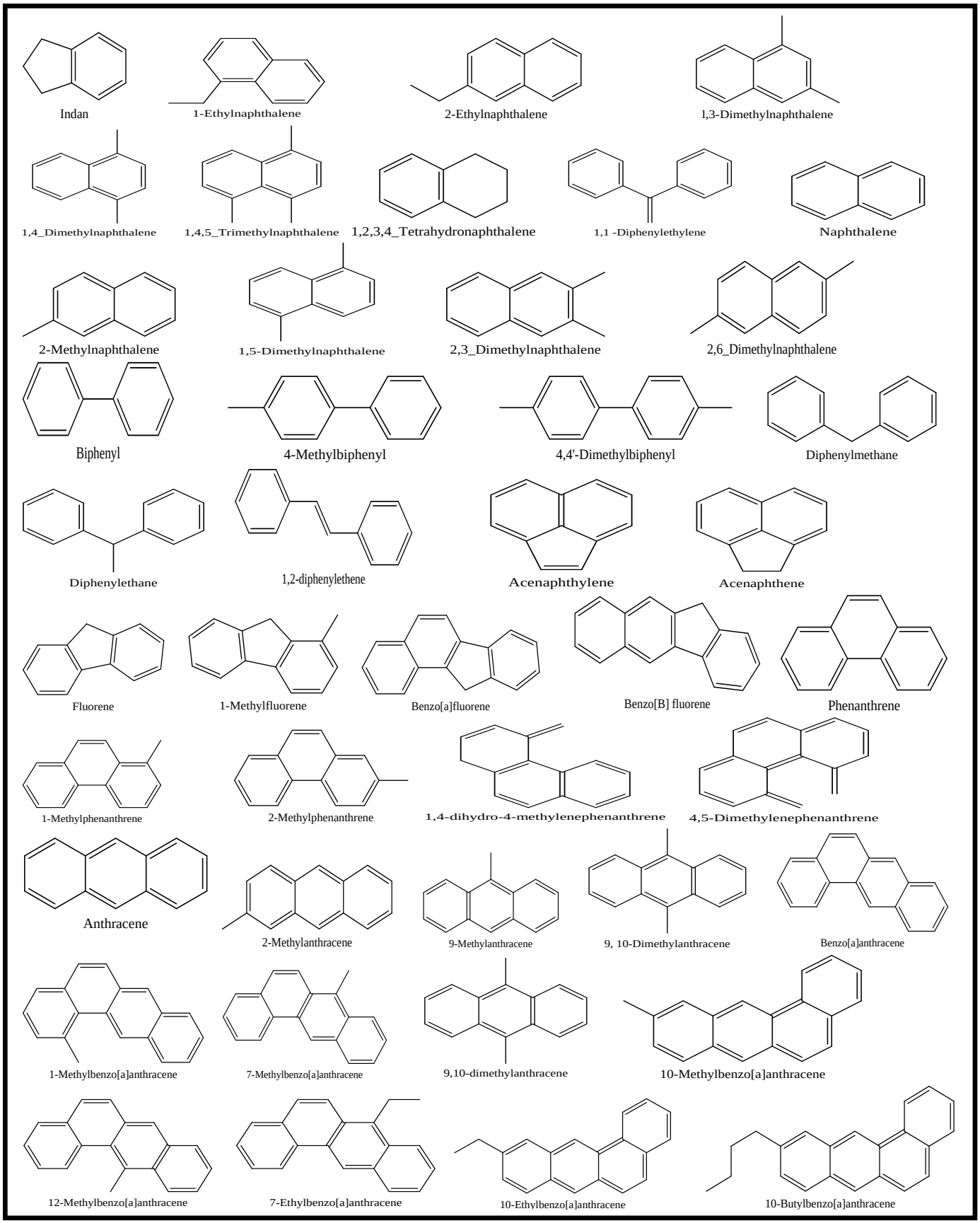


Figure. 29 Les structures des HAP étudiés.

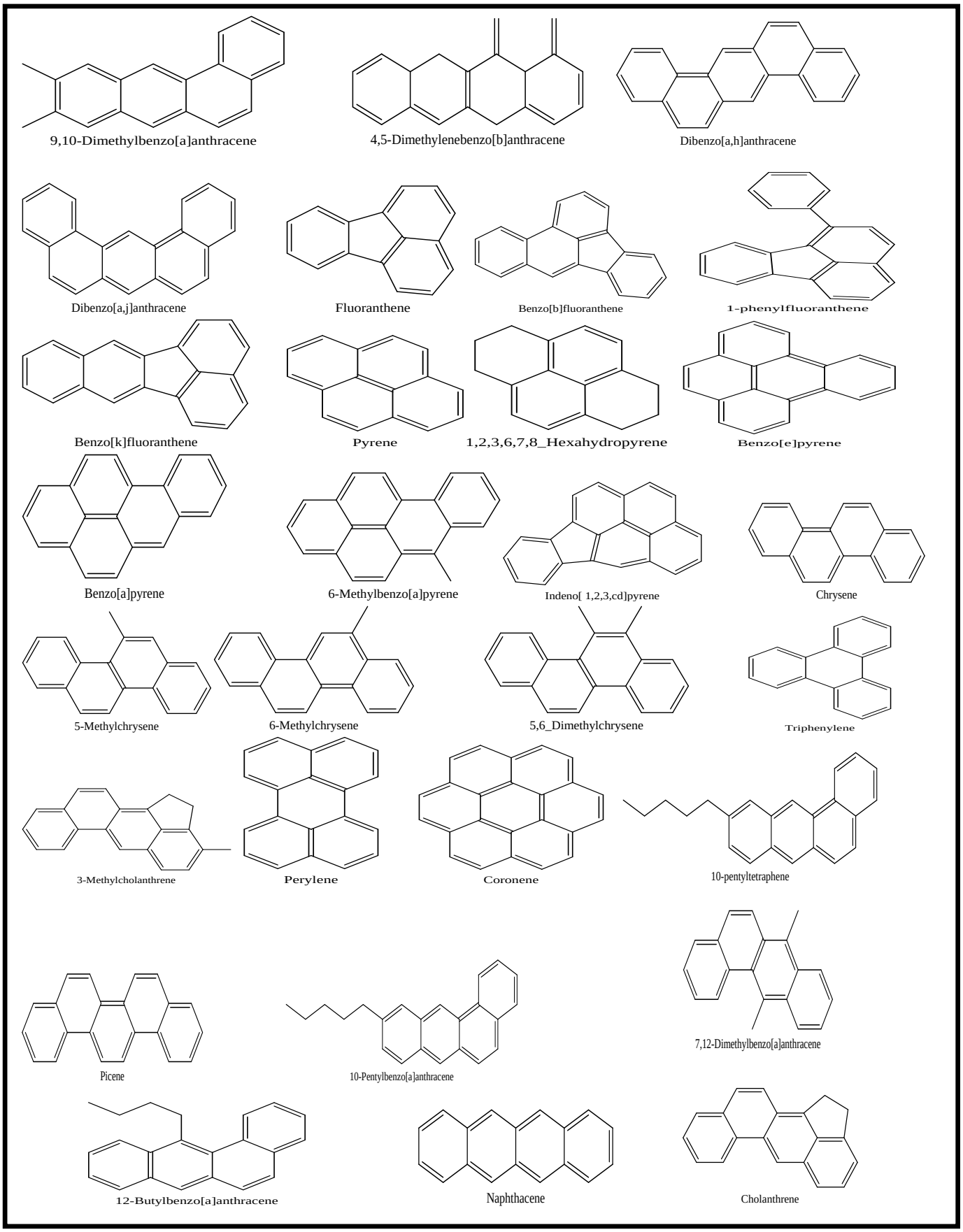


Figure .29 **(suit et fin)**

1. Le modèle MLR

Le modèle QSPR sélectionné est représenté par l'équation :

$$\textbf{-log S} = 30{,}25(\pm 1{,}84)+1{,}63(\pm 0{,}58)\boldsymbol{Km} - 111{,}09(\pm 4{,}95)\boldsymbol{BIC0} - 38{,}08(\pm 12{,}16)\boldsymbol{qpmax} \quad (103)$$

Avec les valeurs des paramètres statistiques ci-après :

N_{tr} = 51, R^2= 91,57%, Q^2_{LOO} = 90, 24%, R^2_{ext} = 90,43%, Q^2_{LMO} = 89,87%, Q^2_{F1} = 91,10%, Q^2_{F2}= 86, 46%, Q^2_{F3} = Q^2_{ext} = 91,38%, CCC_{ext} = 93,96%, $RMSE_{tr}$ =0,47, $RMSE_{pr}$ =0,47, S = 0,49.

Tableau 11: Définition des descripteurs moléculaires intervenant dans la modélisation de la solubilité

Descripteurs	Classe	Définition
Km	descripteur WHIM	Indice global de forme K / pondéré par la masse
BIC0	Indices d'information	Indice du contenu de l'information de liaison (symétrie de voisinage d'ordre 0)
qpmax	Descripteur de charge	Charge positive maximum

La matrice de corrélation reproduite dans le tableau 12 précise les valeurs des coefficients de corrélation pris deux à deux, pour les trois descripteurs et la valeur de la corrélation de chaque descripteur avec la variable à expliquer (logS(-)). Il est à remarquer que chaque descripteur est faiblement corrélé avec chacun des deux autres et que le descripteur ***BIC0*** a une forte corrélation négative avec logS(-). La corrélation négative d'un descripteur indique que les valeurs de la solubilité sont inversement liées à ce descripteur puisque ses valeurs sont positives.

Tableau.12: Matrice de corrélation.

	log S(-)	***Km***	***BIC0***
Km	0,159		
BIC0	-0,942	-0,075	
qpmax	0,136	-0,033	-0,292

Les valeurs absolues de T indiquées dans le tableau 13 expriment que les coefficients de régression des descripteurs impliqués dans le modèle MLR sont significativement plus grands que l'écart-type. Nous remarquons aussi que tous les descripteurs sont significatifs (p <0,05). La valeur du VIF pour chaque descripteur (***Km***, ***BIC0***, ***qpmax***) utilisé proche de 1 ce qui indique une faible multi-colinéarité (*cf*. la matrice de corrélation).

Tableau.13 Caractéristiques des descripteurs du modèle

Prédicteur	*Coef*	*SE Coef*	*T*	*P*	*VIF*
Constante	30,250	1,840	16,440	0,000	
Km	1,630	0,582	2,800	0,007	1,009
BIC0	-111,097	4,958	-22,410	0,000	1,033
qpmax	38,080	12,170	-3,130	0,003	1,042

2. Qualité de l'ajustement

Les valeurs prédites de la solubilité pour les composés dans les ensembles de calibrage et de validation externe en utilisant l'équation (103) ont été présentées en fonction des valeurs expérimentales dans la figure 30. D'après le tableau 10 et la figure 30, les valeurs prédites pour la solubilité sont en bon accord avec celles des valeurs expérimentales.

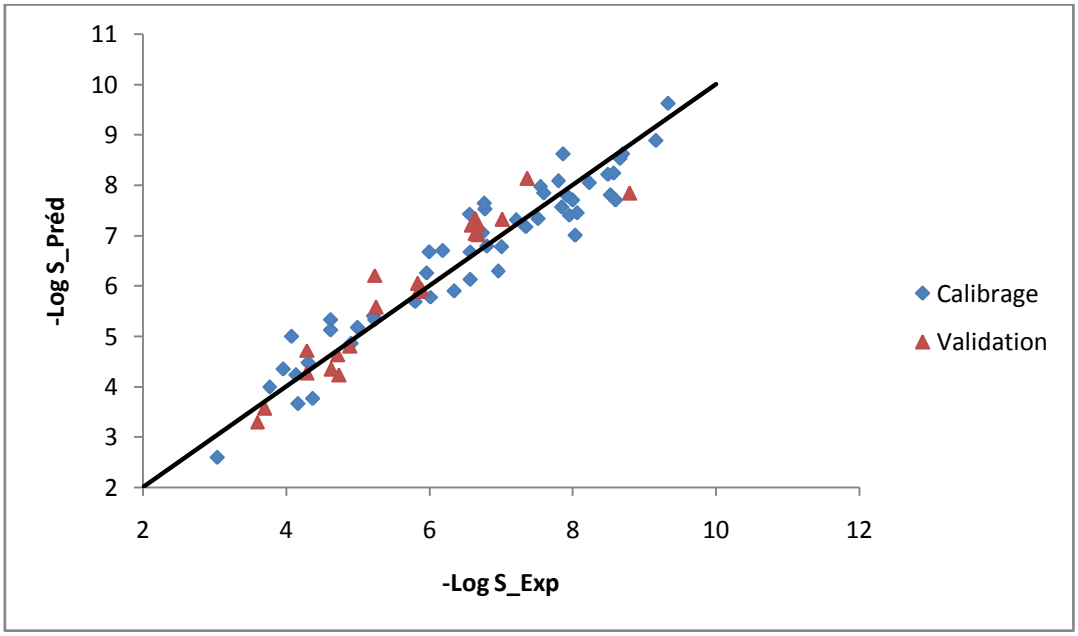


Figure.30 La solubilité prédite des HAP en fonction de celle observée.

3. Domaine d'application

Le Domaine d'application du modèle a été analysé à l'aide du diagramme de Williams (figure 31). La Figure montre clairement la bonne prédictivité pour les ensembles d'apprentissage et de validation. Tous les composés de calibrage et de validation sont dans le domaine chimique, ce qui suggère qu'il n'y a pas de valeurs aberrantes (en X et/ou en Y) et que la prédiction du modèle est fiable.

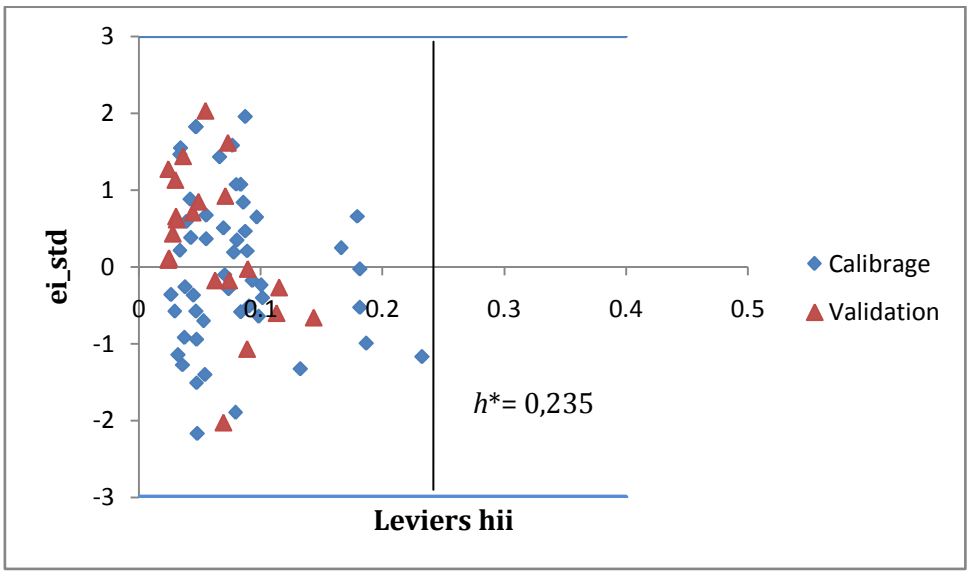


Figure. 31 Diagramme de Williams.

4. Test de randomisation

Le modèle a été validé en utilisant un test de randomisation (figure 32). Il est clair que les statistiques obtenues pour les vecteurs modifiés de (- log S) sont inférieures à celles du modèle réel (cercle bleu et vert). Ceci permet d'assurer que le modèle établi n'est pas dû au hasard.

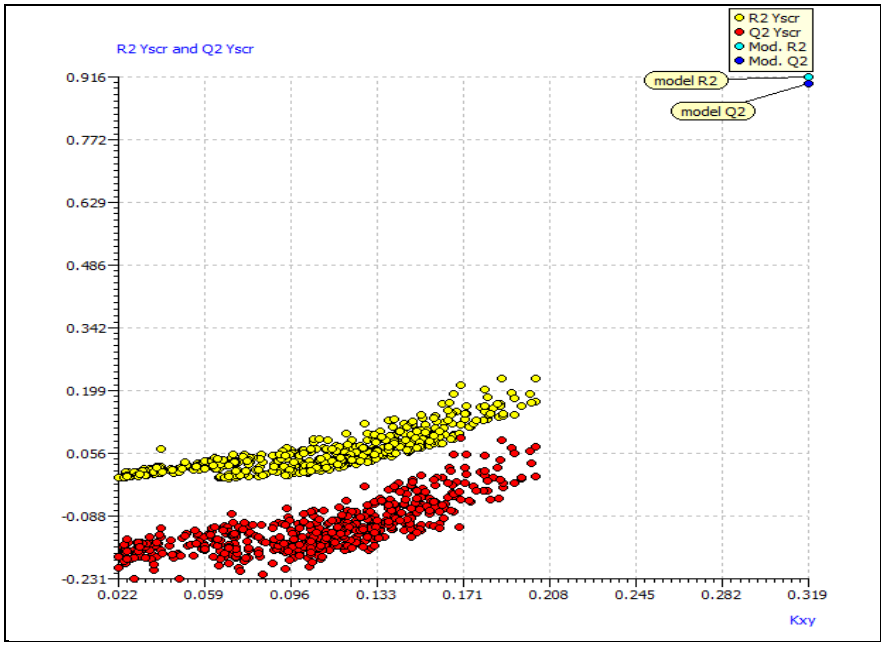


Figure.32 Test de randomisation

Conclusion

L'approche hybride GA/RLM a été exploitée pour développer un modèle à trois descripteurs concernant la solubilité aqueuse de 72 HAP. Les paramètres statistiques relatifs à la validation interne et externe du modèle démontrent que le modèle est simple, fiable avec une bonne capacité prédictive. Puisque ce modèle a été développé sur la base de descripteurs moléculaires théoriques calculés exclusivement à partir de la structure moléculaire, le modèle proposé pourrait estimer et fournir des informations sur la solubilité des HAP.

Références bibliographiques

Ferreira M M C. Polycyclic aromatic hydrocarbons: a QSPR study. Chemosphere. 2001. 44, 125.

Gui-Ning Lua, Zhi Danga, Xue-Qin Taob, Chen Yanga, Xiao-Yun Yi, Estimation of Water Solubility of Polycyclic Aromatic Hydrocarbons Using Quantum Chemical Descriptors and Partial Least Squares, QSAR Comb. Sci. 2008. 27, 618.

Qi Jun, Sun Chang-Hong, Wei Jia, Comparison of Genetic Algorithm Based Support Vector Machine and Genetic Algorithm Based RBF Neural Network in Quantitative Structure-Property Relationship Models on Aqueous Solubility of Polycyclic Aromatic Hydrocarbons. Procedia. Environ. Sci. 2010. 2, 1429.

Qi Jun, Wei Jia, Sun Chang-Hong, Tao PAN. A comparative QSPR study on aqueous solubility of polycyclic aromatic hydrocarbons by GA-SVM, GA-RBFNN and GA-PLS. Front. Earth. Sci. 2011. 5, 245.

Ruelle P, Kesselring U W, aqueous solubility prediction of environmentally important chemicals from the mobile order thermodynamics. Chemosphere.1997. 34, 275.

QSRR DE L'INDICE DE RÉTENTION DE 209 HAP SÉPARÉS PAR CHROMATOGRAPHIE GAZEUSE À TEMPÉRATURE PROGRAMMÉE

Les relations quantitatives structure-rétention (QSRR) relient quantitativement la structure chimique avec l'indice de rétention.

Plusieurs modèles QSRR ont été établis pour la prédiction de l'indice de rétention des 209 HAP. Shushen *et al.*,(2002) ont proposé un vecteur distance - électronégativité moléculaire (molecular electronegativity–distance vector, MEDV) pour décrire la structure des HAP et la relier avec leurs Ir. Ils ont appliqué la régression linéaire multiple (MLR) en combinaison avec la technique de validation croisée (VC); un modèle QSRR à quatre paramètres pour 209 HAP a été développé avec un coefficient de corrélation (R) de 0,9812 et une erreur quadratique moyenne (RMS) de 15,533 entre les Ir estimés et expérimentaux. En 2010, Drosos et *al.*, ont développé une étude QSRR pour certains HAP utilisant la mécanique quantique et d'autres sources de descripteurs estimés par différentes approches. Une bonne relation linéaire a été trouvée entre l'indice de rétention en chromatographie gazeuse et les descripteurs électroniques ou topologiques par analyse de régression linéaire pas à pas. Touhami *et al.*(2015) ont développé un modèle QSRR pour un ensemble de 248 (209+39) HAP avec R^2= 0,9708; et R^2_{ext} =0,9427.

Le but de cette étude est de proposer un modèle QSRR simple, robuste avec une capacité prédictive externe élevée et facile à interpréter.

1. Conditions expérimentales

Nous résumerons les conditions d'analyses chromatographiques, décrites en détail dans (Lee *et al.,* 1979). Les séparations ont été réalisées sur une colonne capillaire en verre (longueur : 12m ; diamètre intérieur : 0,3 mm) dont les parois internes sont recouvertes d'un mince film (0,17µm d'épaisseur) d'une phase stationnaire méthylphénylsilcone SE-52 (5% phényl) peu polaire. La colonne a été montée sur un chromatographe Varian, modèle 3700, dont la température du four est élevée de 50 à 250° C à raison de 2°C/min.

Le débit de l'hélium vecteur varie de 1 à 3 ml/min ; l'acquisition des données est facilitée par l'utilisation d'un intégrateur (Varian CDS101).

Base de données: Les indices de rétention des 209 HAP étudiés ont été prélevés de la référence (Lee *et al.*, 1979).

Pour définir l'ensemble de calibrage et de prédiction (validation externe), nous avons utilisé l'algorithme de Kennard et Stone pour diviser l'ensemble complet en 146 composés pour le calibrage et 63 composés pour la prédiction.

2. Résultats et discussion

Parmi les descripteurs pouvant être en relation avec Ir, un sous-ensemble de quatre descripteurs sera vraisemblablement mieux adapté pour la modélisation par RLM. Les quatre descripteurs optimaux sont présentés dans le tableau 14 :

Tableau .14 Ensemble optimal pour la modélisation de l'indice de rétention des 209 HAP

N°	Descripteur	Classe	Signification
1	X5A	Indice de connectivité	indice de connectivité moyen d'ordre 5
2	X1sol	Indice de connectivité	indice de connectivité de solvatation d'ordre 1
3	MATS4m	Autocorrélation- 2D	Auto-corrélation de Moran de décalage 4 / pondérée par la masse
4	Mor21m	Descripteur 3D-MoRSE	3D-MoRSE - signal 21 / pondéré par les masses atomiques

2.1.Equation et analyse de régression

Le modèle, construit sur l'ensemble de calibrage, a pour équation :

$$\boldsymbol{Ir} = 155\ (\pm 23{,}43) - 1599\ (\pm 232{,}80)\ \boldsymbol{X5A} + 36{,}7\ (\pm 1{,}28)\ \boldsymbol{X1sol} - 13{,}74(\pm 3{,}06)$$

$$\boldsymbol{MATS4m} - 45{,}54(\pm 8{,}22)\ \boldsymbol{Mor21m} \qquad (104)$$

Avec : $\boldsymbol{n_{tr}} = 146$, $\boldsymbol{R^2} = 95{,}73\%$, $\boldsymbol{Q^2_{LOO}} = 95{,}34\%$, $\boldsymbol{R^2_{ext}} = 97{,}32\%$, $\boldsymbol{Q^2_{LMO}} = 95{,}08\%$,

$\boldsymbol{Q^2_{F1}} = 98{,}02\%$, $\boldsymbol{Q^2_{F2}} = 97{,}13\%$, $\boldsymbol{Q^2_{F3}} = \boldsymbol{Q^2}_{\text{ext}} = 97{,}69\%$, $\boldsymbol{CCC_{ext}} = 98{,}56\%$, $\boldsymbol{RMSE_{tr}} = 16{,}49$,

$\boldsymbol{RMSE_{pr}} = 12{,}14$, $S = 16{,}78$.

Les paramètres statistiques reproduits dans le tableau 15 permettent d'évaluer les descripteurs du modèle. Les valeurs des probabilités de t pour les quatre descripteurs sont toutes très inferieures à 0,05, ce qui indique qu'ils sont très significatifs. Les valeurs des facteurs d'inflation de la variance (FIV, toutes inférieures à 3) indiquent que ces descripteurs sont faiblement corrélés entre eux.

Tableau.15 Caractéristiques des descripteurs sélectionnés dans le modèle MLR.

Régresseur	Coef	SE Coef	T	P	VIF
Constante	154,76	23,43	6,61	0,000	
X5A	-1599,10	232,80	-6,87	0,000	1,862
X1sol	36,733	1,277	28,77	0,000	2,462
MATS4m	-13,739	3,056	-4,50	0,000	1,230
Mor21m	-45,544	8,225	-5,54	0,000	2,141

Les contributions relatives des quatre descripteurs du modèle ont été déterminées et présentées dans la figure 33; les valeurs des contributions renseignent sur l'importance des descripteurs dans le modèle, qui sont classés dans l'ordre : **X1sol** (45,073%) > **X5A** (19,151%), > **Mor21m** (18,310%) > **MATS4m** (17,466). L'importance de l'indice de connectivité de solvatation (X1sol) est évidente ce descripteur contribue pour 45,037% dans la construction du modèle, la différence de contribution entre les trois autres descripteurs n'est pas significative ce qui indique que tous ces descripteurs sont nécessaires pour la construction du modèle.

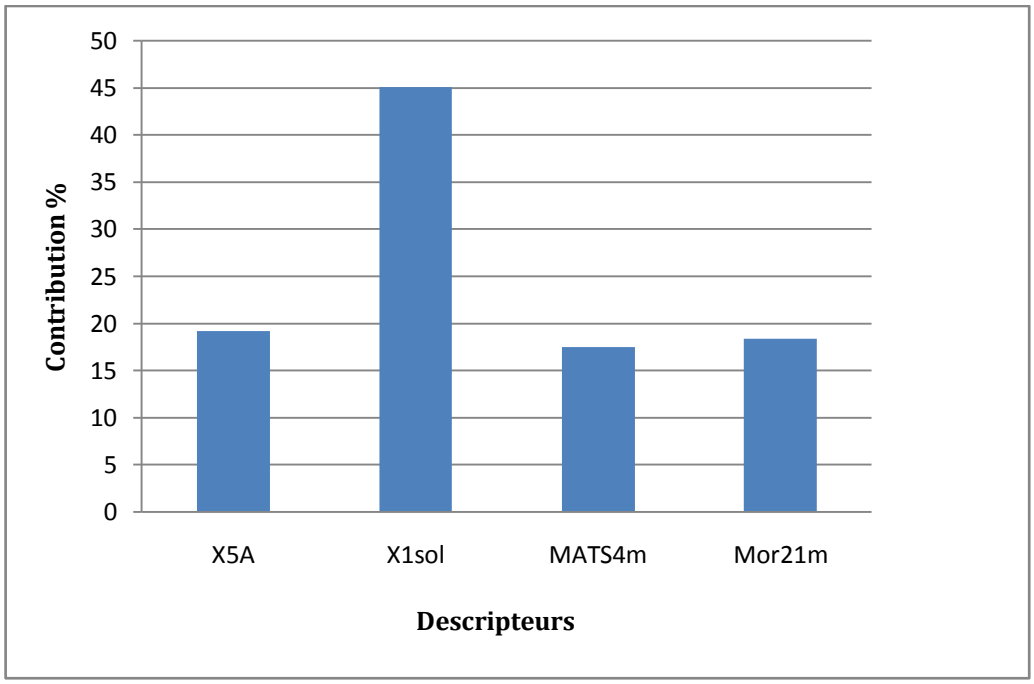


Figure. 33 Importance(%) des descripteurs dans le modèle.

Les paramètres statistiques suivants obtenus pour l'ensemble de validation vérifient les conditions d'acceptabilité de Golbraikh et Tropsha.(2002) établissant ainsi la capacité prédictive du modèle calculé:

r^2_{ext} = 0,9732 > 0,5 r^2_0 = 0,9731 r'^2_0 = 0,9730

$(r^2 - r^2_0)/r^2$ = -0,0253 $(r^2 - r'^2_0)/r^2$ = -0,0258

$R^2_{cv.ext}$ = 0,9713 0,85 ≤ k = 1,0094 ≤ 1,15 0.85 ≤ k' = 0,9892 ≤ 1,15

2.2. Domaine d'application

Le diagramme de Williams de la figure 34 permet de visualiser les composés qui n'appartiennent pas au même domaine chimique. D'après cette figure aucun des composés de l'ensemble de prédiction n'est hors du domaine d'application, nous avons seulement les composés: ***o-quaterphényl***, ***m-quaterphényl***, ***p-quaterphényl*** et ***9-phénylanthracène*** de l'ensemble de calibrage qui sont en dehors du domaine.

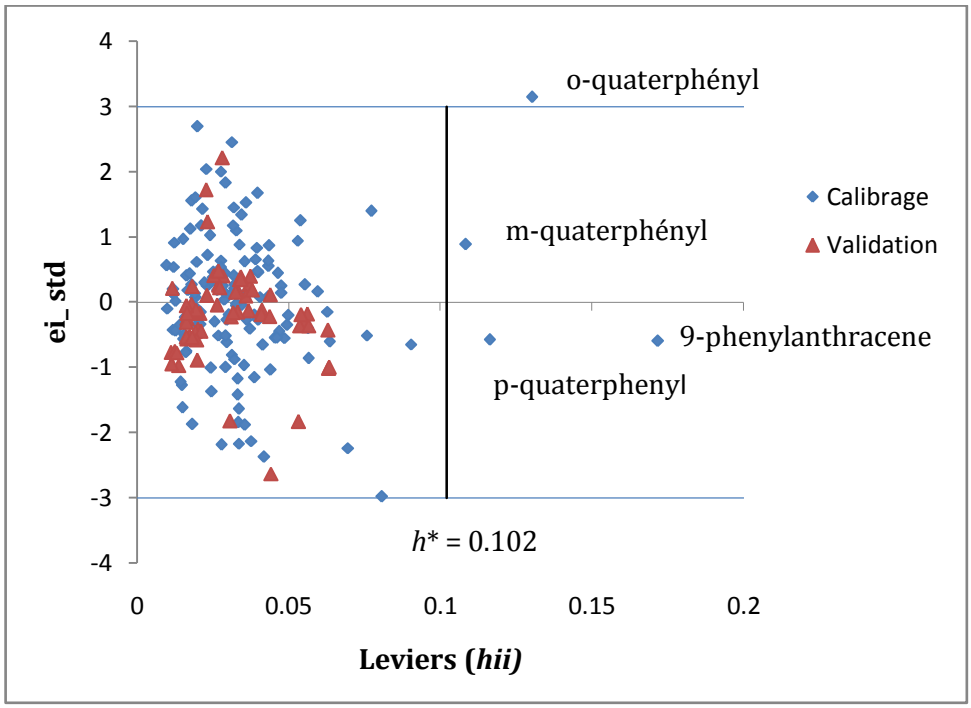


Figure. 34 Diagramme de Williams pour les 209 HAP.

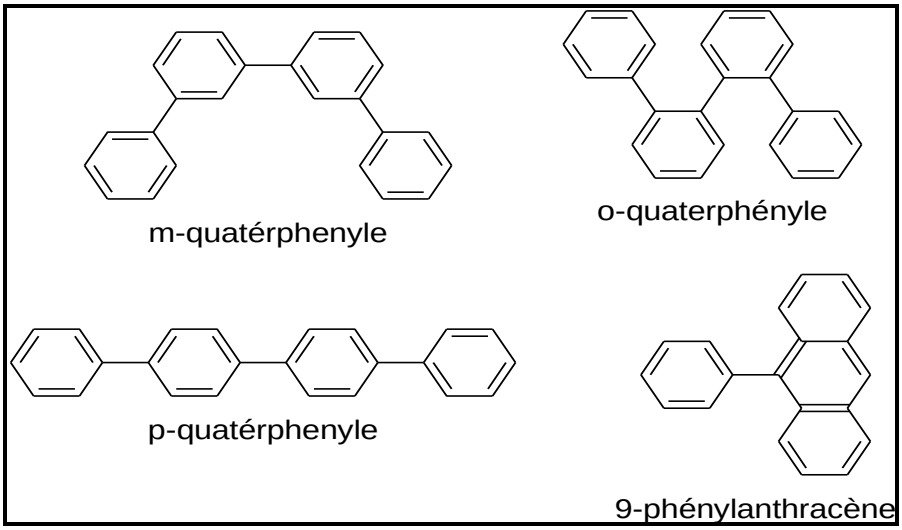


Figure.35 Structures chimiques des composés atypiques.

2.3.Vérification de la qualité de l'ajustement

La figure 36 présente le graphique des valeurs de l'indice de rétention prédites en fonction des valeurs expérimentales, les points s'alignent bien sur la droite de régression (R^2= 95,73%) cela signifie que les indices de rétention sur la colonne SE-52 sont bien prédits par le modèle développé.

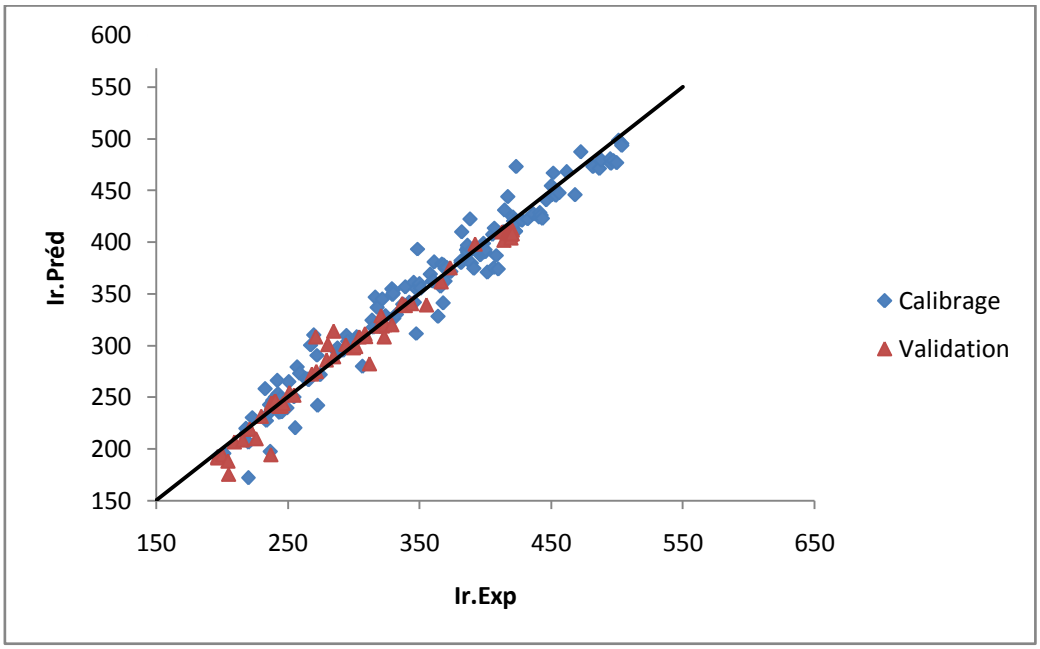


Figure. 36 Graphe des valeurs (IR) prédites en fonction des valeurs expérimentales.

2.4. Le test de randomisation

La figure 37 compare les résultats obtenus pour les modèles randomisés au modèle réel de départ. Il est clair que les statistiques *R²Yscr* et *Q²Yscr* sont plus petites que celles du modèle QSPR réel. Ce qui permet de s'assurer que le modèle établi a une base réelle et qu'il n'est donc pas fortuit.

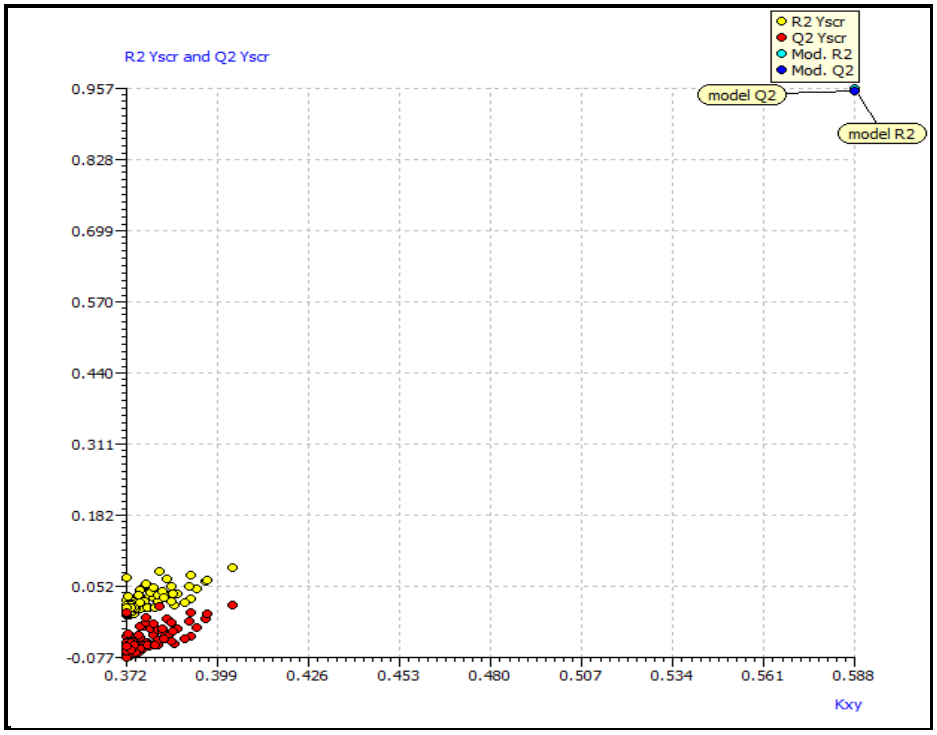


Figure. 37 Test de randomisation.

2.5 Définition des descripteurs du modèle

La disponibilité d'un grand nombre de descripteurs théoriques contenant diverses sources d'informations chimiques serait utile pour mieux comprendre la relation entre la structure moléculaire et les données expérimentales. L'indice de connectivité de solvatation (***X1sol***) apparaissant dans le modèle MLR montre principalement les caractéristiques topologiques. Ceux-ci indiquent que les interactions de dispersion et l'étendue de la ramification des molécules affectent le comportement de rétention des HAP sur la colonne. Les indices de connectivité de solvatation (Xn) sont définis par l'éq. (105) :

$$ {}^{m}x_q^s = \frac{1}{2^{m+1}} \sum_{k=1}^{k} \frac{(\prod_{a=1}^{n} L_a)_k}{(\prod_{a=1}^{n} \delta_a)_k^{1/2}} \qquad (105) $$

Où La est le nombre quantique principal (2 pour les atomes C, N, O, et 3 pour Si, S, Cl,...) du $t^{ème}$ atome dans le $k^{ième}$ sous-graphe et δa le degré du sommet correspondant; *k* est le nombre total de sous-graphes d'ordre *m*; *n* est le nombre de sommets dans le sous-graphe.

Le facteur de normalisation 1 / (2^{m+1}) est défini de telle sorte que les indices ${}^{m}X$ et ${}^{m}X^{s}$ pour les composés ne contenant que des atomes de la deuxième rangée coïncident. Le descripteur X1sol a un coefficient de régression de 36,733, qui est le plus grand parmi les descripteurs apparaissant dans le modèle. Ce paramètre peut être considéré comme une entropie de solvatation et indique en quelque sorte les interactions de dispersion (*cf*.supra) se produisant dans les solutions.

X5A est une mesure de la ramification des molécules. Les grandes contributions de ce paramètre dans le comportement de la rétention des HAP sont en accord avec la contribution que l'on pourrait attendre de l'interaction entre la phase stationnaire et les HAP. ***X5A*** est l'indice de connectivité moyen qui peut être obtenu à partir du graphe moléculaire.

MATS4m (autocorrélation Moran 4 / pondérée par les masses atomiques) un des descripteurs d'autocorrélation 2D (Fernandez *et al.*, 2006), également obtenu à partir des graphes moléculaires, en sommant les produits des poids atomiques des atomes terminaux de tous les chemins de la longueur du chemin considéré (le retard).

Le descripteur ***MATS4m*** est lié à la masse atomique d'une molécule, ce descripteur concerne la taille moléculaire, qui influe sur la rétention du composé. Les descripteurs moléculaires basés sur la fonction d'autocorrélation AC_l définie par:

$$AC_l = \int_b^a f(x)f(x+1)dx \qquad (106)$$

Où *f(x)* est une fonction de la variable x et 1 est le retard représentant un intervalle de x; a et b définis l'intervalle total étudié de la fonction. La fonction *f(x)* est habituellement une fonction dépendante du temps, comme un signal électrique qui dépend du temps, ou une fonction dépendant de l'espace telle que la densité de population dans l'espace.

Mor21m (3D MoRSE -signal 21 / pondéré par les masses atomiques) fait partie des descripteurs 3D-MoRSE. Le descripteur 3D-MoRSE (Saiz-Urra *et al.*, 2006) (***Mor21m***) est une projection d'atomes moléculaires selon différents angles, comme en diffraction d'électrons. Ils représentent différentes vues de la structure de la molécule entière, bien que leur signification ne soit pas trop claire. Les descripteurs 3D-MoRSE sont basés sur l'idée

d'obtenir des informations à partir des coordonnées atomiques 3D par la transformée utilisée dans les études de diffraction d'électrons pour la préparation de courbes de diffusion théoriques. Les descripteurs 3D-MoRSE sont importants car ils prennent en compte la disposition 3D des atomes sans ambiguïté (contrairement à ceux issus des graphes chimiques), et aussi parce qu'ils ne dépendent pas de la taille moléculaire, donc applicables à un grand nombre de molécules avec une grande variance structurelle et présente une caractéristique commune à tous.

Tous ces descripteurs sont individuellement importants pour affecter les temps de rétention mais l'effet final n'est pas le résultat de chaque descripteur individuellement et tous les descripteurs sélectionnés en tant que groupe peuvent influencer sur la variable dépendante.

Conclusion

Nous avons appliqué la méthodologie QSRR pour relier les indices de rétention de 209 HAP à des descripteurs moléculaires théoriques. La base de données a été devisée en deux sous ensembles selon l'algorithme de Kennard et Stone (146 composé pour l'ensemble de calibrage et 63 pour la validation externe). Le modèle a été établi en appliquant l'approche hybride algorithme génétique / régression linéaire multiple. Le diagramme de Williams détecte quatre composés en dehors du domaine d'application du modèle. Le test de randomisation associé au modèle obtenu permet d'assurer qu'une relation structure / rétention réelle a été établie. Le modèle QSRR présenté est robuste, avec de bonnes capacités prédictives internes et externes, et une bonne qualité de l'ajustement.

Références bibliographique

Lee M, Vassilaros D L, White CM, Novotny M, Retention indices for programmed-temperature capillary-column gas chromatography of polycyclic aromatic hydrocarbons, *Analytical Chemistry*. 1979.51, 768.

Shushen L A, Chunsheng Y B, Shaoxi Ca, Zhiliang LA,.. Molecular structural vector description and retention index of polycyclic aromatic hydrocarbons. *Chemometrics and Intelligent Laboratory Systems* . 2002.61, 3.

Fernandez M, Caballero J, Tundidor-Camba A, Linear and nonlinear QSAR study of N-hydroxy-2 [(phenylsulfonyl) amino] acetamide derivatives as matrix metalloproteinase inhibitors. Bioorg.Med.Chem. 2006. 14, 4137.

Saiz-Urra L, Gonzalez MP, Teijeira M: QSAR studies about cytotoxicity of benzophenazines with dual inhibition toward both topoisomerases I and II: 3D-MoRSE descriptors and statistical considerations about variable selection. Bioorg Med Chem. 2006. 14, 7347.

Drosos J C, Viola-Rhenals M, Vivas-Reyes R, Quantitative structure–retention relationships of polycyclic aromatic hydrocarbons gas-chromatographic retention indices J. Chromatogr. A. 2010. 1217,4411.

Liu S, Yin C, Cai S, Li Z, Molecular structural vector description and retention index of polycyclic aromatic hydrocarbons. Chemometr. Intell. Lab. 2002.61, 3.

Touhami Imen, thèse de doctorat. Calcul des équations de prédictions : des indices de rétention en chromatographie gazeuse (HAP comportant un hétéro atome (O,S,N)), pyrazines, et des propriétés physico-chimiques pour une série de pesticides. Laboratoire de Sécurité Environnementale et Alimentaire Université BADJI Mokhtar-Annaba-ALGERIE. 2017.

CONCLUSION GÉNÉRALE

Les hydrocarbures aromatiques polycycliques (HAP) sont des molécules organiques issues de la combustion incomplète des matières carbonées suite à des processus naturels, mineurs, et des processus anthropiques, majoritaires. Les HAP sont libérés dans tous les compartiments de l'environnement. Il est important, pour comprendre la présence et les devenir des HAP dans l'environnement, d'approfondir la connaissance des sources elles-mêmes mais d'étudier également les mécanismes de dispersion, de transformation et de transport. Les propriétés physico-chimiques jouent un rôle fondamental dans le comportement des HAP, la connaissance de ces propriétés est nécessaire pour la prédiction du devenir de ces polluants dans les différentes phases environnementales. Malheureusement la disponibilité des données sur les propriétés physico-chimiques est rare dans la littérature.

Dans ce contexte, l'objectif de cette thèse était de développer des modèles QSPR fiables pour la prédiction de quelques propriétés physicochimiques des hydrocarbures aromatiques polycycliques. Un grand nombre de descripteurs moléculaires a été calculé (Descripteurs constitutionnels, électroniques, topologiques, géométriques, physico-chimiques,...). Des méthodes statistiques ont été utilisées dans la construction de ces modèles (MLR, ANN...). Les principales techniques de validation ont été utilisées (les tests statistiques standards, la validation interne, la validation externe, la randomisation des Y, les domaines d'applications...). Ces modèles ont été développés en accord avec les cinq principes de l'OCDE pour la validation des modèles QSPR, (à savoir : une propriété ciblée définie avec un protocole expérimental identifié; un algorithme sans équivoque ; un domaine d'applicabilité défini ; des mesures appropriées de la qualité d'ajustement, de robustesse et du pouvoir prédictif ; et si possible, une interprétation des mécanismes sous-jacents).

La méthodologie basée sur la MLR, a été utilisée principalement dans la prédiction. Cette méthode permet d'extraire de manière efficace des modèles QSPR. Ces modèles sont à la fois fiables, explicatifs, prédictifs et interprétables en choisissant des descripteurs pertinents pour expliquer et interpréter la propriété des composés étudiés du point de vue statistique et chimique.

La méthode des réseaux de neurones artificiels témoigne de l'existence d'une relation non-linéaire entre la propriété étudiée et les structures moléculaires. Mais le problème majeur

est qu‘il n'existe pas une forme explicite expliquant et analysant la relation entre les entrées et les sorties pour la RNA. Cela cause des difficultés d‘interprétation des résultats (modèles) obtenus par cette méthode.

La température d’ébullition, la solubilité aqueuse et l’indice de rétention ont été reliés à des paramètres structuraux, des modèles de différentes tailles (2,3 et 4 descripteurs respectivement) ont été construits la qualité des résultats obtenus montre la bonne relation linéaire entre la température d’ébullition et les descripteurs choisis.

Pour la température de fusion deux différentes méthodes ont été utilisées (MLR et RNA), un modèle à quatre descripteur à été développé par la méthode MLR, la non satisfaction des résultats obtenu par cette méthode linéaire fait intervenir l’utilisation d’une méthode non linéaire (RNA), cette dernière fait ressortir de bons résultats. La comparaison entre la qualité des modèles MLR et RNA pour la température de fusion montre qu’il n y a une différence tant par la qualité de l’ajustement, la robustesse ou la capacité prédictive.

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

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

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

Le test de randomisation est un outil puissant pour vérifier et s’assurer que les modèles obtenus ne sont pas dus au hasard.

Finalement, bien que les objectifs principaux de cette thèse aient été remplis, et afin de poursuivre notre chemin de recherche dans cette discipline nous prévoyons de :

Reprendre les mêmes bases de données et élaborer des modèles en utilisant d‘autres méthodes telles que : les algorithmes génétiques (GA), les graphs machines, les séparateurs à vaste marge (SVM).

ANNEXES

ANNEXE I

Température d'ébullition

I. *Séparation aléatoire des données*

Tableau,I Valeurs des descripteurs (choix aléatoire)

	composé	T eb(k)	HOMO	EPS0	status		composé	T eb(k)	HOMO	EPS0	status
1	2-methylnaphthalene	514	-8,744	7,5	pr	32	benzo[b]fluoranthene	754	-8,663	14,276	Tr
2	1-methylnaphthalene	518	-8,689	7,552	Tr	33	benzo[k]fluoranthene	754	-8,399	14,223	pr
3	2,6-dimethylnaphthalene	535	-8,643	7,947	pr	34	benzo[e]pyrene	769	-8,335	14,276	pr
4	2,7-dimethylnaphthalene	535	-8,688	7,947	pr	35	picene	792	-8,477	15,637	Tr
5	1,7-dimethylnaphthalene	536	-8,62	8	pr	36	aq Benzo[a]pyrene	769	-8,041	14,223	Tr
6	1,3-dimethylnaphthalene	538	-8,609	8	Tr	37	Perylene	770	-7,987	14,276	Tr
7	1,6-dimethylnaphthalene	539	-8,625	8	pr	38	Dibenz[a,h]anthracene	808	-8,376	15,585	Tr
8	2,3-dimethylnaphthalene	541	-8,669	8	Tr	39	Benzo[ghi]perylene	815	-8,139	15,723	pr
9	1,4-dimethylnaphthalene	541	-8,541	8,052	Tr	40	Coronene	863	-8,289	17,171	Tr
10	1,5-dimethylnaphthalene	542	-8,585	8,052	Tr	41	Naphtho[1,2,3,4-def]chrysene	865	-8,14	17,137	pr
11	1,2-dimethylnaphthalene	544	-8,617	8,052	Tr	42	Benzo[rst]pentaphene	867	-7,987	17,085	Tr
12	2,3,5-trimethylnaphthalene	558	-8,562	8,5	pr	43	Dibenzo[def,p]chrysene	868	-8,079	17,137	Tr
13	2,3,6-trimethylnaphthalene	559	-8,585	8,447	pr	44	Dibenzo[b,def]chrysene	869	-7,807	17,085	Tr
14	2-methylfluorene	591	-8,704	9,654	pr	45	Benzo[c]fluorene	679	-8,424	12,069	Tr
15	3-methylphenanthrene	625	-8,639	10,361	pr	46	Dibenzo[def,mno]chrysene	820	-7,762	15,671	Tr
16	2-methylphenanthrene	628	-8,635	10,361	Tr	47	Dibenz[a,j]anthracene	804	-8,4	15,585	Tr
17	9-methylphenanthrene	628	-8,717	10,414	Tr	48	Indeno[1,2,3-cd] fluoranthene	804	-8,663	15,776	Tr
18	2-methylanthracene	632	-8,189	10,309	Tr	49	Indeno[1,2,3,cd]pyrene	807	-8,238	15,723	Tr
19	1-methylphenanthrene	632	-8,645	10,414	pr	50	Benzo[b]triphenylene	808	-8,4	15,637	Tr
20	1-methylanthracene	636	-8,176	10,361	Tr	51	Benzo[j]fluoranthene	753	-8,439	14,276	Tr
21	3,6-dimethylphenanthrene	636	-8,553	10,809	pr	52	Cyclopenta[cd]pyrene	712	-8,372	12,809	Tr
22	4-methylpyrene	683	-8,196	11,861	Tr	53	naphthalene	491	-8,835	7,052	Tr
23	2-methylpyrene	683	-8,23	11,809	Tr	54	acenaphthylene	543	-9,055	8,5	pr
24	1-methylpyrene	683	-8,151	11,861	Tr	55	acenaphthene	552	-8,59	8,5	Tr
25	benzo[b]fluorene	675	-8,548	12,016	Tr	56	fluorene	567	-8,842	9,207	Tr
26	benzo[a]fluorene	680	-8,507	12,069	Tr	57	phenanthrene	611	-8,74	9,914	pr
27	benzo[ghi]fluoranthene	705	-8,774	12,861	Tr	58	anthracene	613	-8,248	9,861	Tr
28	benz[a]anthracene	708	-8,327	12,723	Tr	59	4H-cyclopenta[def]phenanthrene	632	-8,656	10,654	Tr
29	triphenylene	712	-8,772	12,828	pr	60	fluoranthene	656	-8,725	11,414	pr
30	chrysene	714	-8,496	12,776	Tr	61	pyrene	666	-8,249	11,361	Tr
31	naphthacene	723	-7,87	12,671	Tr						

Paramètres statistiques du modèle (séparation aléatoire)

Variable	Coeff,	Std, coeff,	Std, err,	(+/-) Co, int, 95%	p
Intercept	509,508		28,035	56,707	0,000
HOMO	28,213	0,072	3,090	6,251	0,000
EPS0	34,101	0,956	0,283	0,574	0,000

Critères de fitness

R2: 0,9983	R2adj: 0,9982	R2-R2adj: 0,0001
Kxx: 0,5578	Delta K: 0,1728	RMSE tr: 4,4202
RSS tr: 820,6080	s: 4,5871	CCC tr: 0,9991

Critère de validation interne

Q2loo: 0,9980	R2-Q2loo: 0,0003	RMSE cv: 4,7709
PRESS cv: 955,9981	CCC cv: 0,9990	

Critère de validation externe

RMSE ext: 5,4251 MAE ext: 3,8831 PRESS ext: 559,2084

II. Séparation des données selon l'ordre de réponse

Paramètres statistiques du modèle

Variable	Coeff,	Std, coeff,	Std, err,	(+/-) Co, int, 95%	p
Intercept	491,979		28,101	56,841	0,000
HOMO	26,507	0,069	3,062	6,194	0,000
EPS0	34,376	0,955	0,290	0,588	0,000

Critères de fitness

R2: 0,9984	R2adj: 0,9983	R2-R2adj: 0,0001	LOF: 24,9553
Kxx: 0,6082	Delta K: 0,1515	RMSE tr: 4,5198	MAE tr: 3,3294
RSS tr: 857,9866	CCC tr: 0,9992	s: 4,6904	F: 12139,0217

Critère de validation interne

Q2loo: 0,9981	R2-Q2loo: 0,0003	RMSE cv: 4,8759	MAE cv: 3,5848
PRESS cv: 998,5324	CCC cv: 0,9991	Q2Yscr: -0,1064	RMSE AV Yscr: 110,1245
Q2LMO: 0,9980	R2Yscr: 0,0473		

Critère de validation externe

RMSE ext: 5,1545	MAE ext: 3,9185	PRESS ext: 504,8070	R2ext: 0,9979
Q2-F1: 0,9974	Q2-F2: 0,9974	Q2-F3: 0,9979	CCC ext: 0,9987
r2m aver,: 0,9811	r2m delta: 0,0024		

Angle de régression des données externes calculé à partir de la diagonale: 0,3919°, cet angle est utilisé pour évaluer le biais dans les prédictions externes (Chirico, N., Gramatica, P. 2012. Real External Predictivity of QSAR Models. Part 2. New intercomparable thresholds for different validation criteria and the need for scatter plot inspection. *J.Chem. Inf. Model.* 52, 2044-2058).

Tableau II Valeurs de (Teb) et statut des composés

ID	Nom	Status	Exp Teb,	ID	Nom	Status	Exp Teb,
1	2-methylnaphthalene	tr	514	32	benzo[b]fluoranthene	tr	754
2	1-methylnaphthalene	tr	518	33	benzo[k]fluoranthene	tr	754
3	2,6-dimethylnaphthalene	pr	535	34	benzo[e]pyrene	pr	769
4	2,7-dimethylnaphthalene	tr	535	35	picene	tr	792
5	1,7-dimethylnaphthalene	tr	536	36	aq Benzo[a]pyrene	pr	769
6	1,3-dimethylnaphthalene	pr	538	37	Perylene	tr	770
7	1,6-dimethylnaphthalene	tr	539	38	Dibenz[a,h]anthracene	tr	808
8	2,3-dimethylnaphthalene	tr	541	39	Benzo[ghi]perylene	pr	815
9	1,4-dimethylnaphthalene	pr	541	40	Coronene	tr	863
10	1,5-dimethylnaphthalene	tr	542	41	Naphtho[1,2,3,4-def]chrysene	pr	865
11	1,2-dimethylnaphthalene	pr	544	42	Benzo[rst]pentaphene	tr	867
12	2,3,5-trimethylnaphthalene	tr	558	43	Dibenzo[def,p]chrysene	tr	868
13	2,3,6-trimethylnaphthalene	pr	559	44	Dibenzo[b,def]chrysene	tr	869
14	2-methylfluorene	tr	591	45	Benzo[c]fluorene	tr	679
15	3-methylphenanthrene	tr	625	46	Dibenzo[def,mno]chrysene	tr	820
16	2-methylphenanthrene	pr	628	47	Dibenz[a,j]anthracene	tr	804
17	9-methylphenanthrene	tr	628	48	Indeno[1,2,3-cd] fluoranthene	tr	804
18	2-methylanthracene	tr	632	49	Indeno[1,2,3,cd]pyrene	pr	807
19	1-methylphenanthrene	pr	632	50	Benzo[b]triphenylene	tr	808
20	1-methylanthracene	tr	636	51	Benzo[j]fluoranthene	pr	753
21	3,6-dimethylphenanthrene	pr	636	52	Cyclopenta[cd]pyrene	pr	712
22	4-methylpyrene	pr	683	53	naphthalene	tr	491
23	2-methylpyrene	tr	683	54	acenaphthylene	tr	543
24	1-methylpyrene	tr	683	55	acenaphthene	tr	552
25	benzo[b]fluorene	pr	675	56	fluorene	tr	567
26	benzo[a]fluorene	tr	680	57	phenanthrene	pr	611
27	benzo[ghi]fluoranthene	pr	705	58	anthracene	tr	613
28	benz[a]anthracene	tr	708	59	4H-cyclopenta[def]phenanthrene	tr	632
29	triphenylene	tr	712	60	fluoranthene	tr	656
30	chrysene	tr	714	61	pyrene	tr	666
31	naphthacene	tr	723				

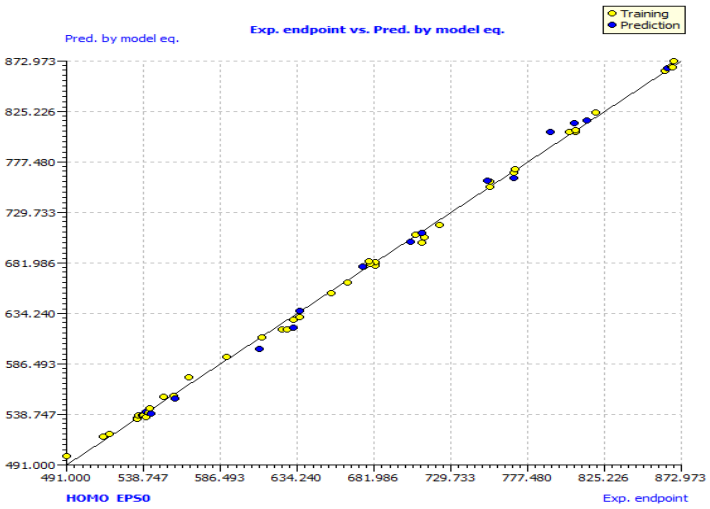


Figure.1 Droite d'ajustement

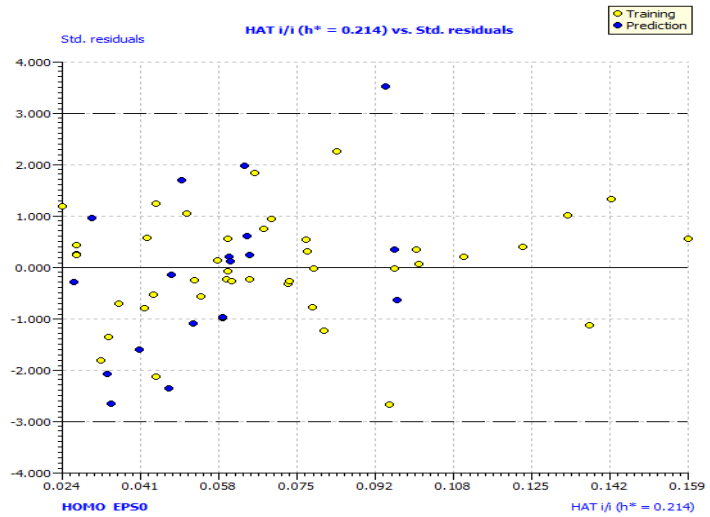


Figure 2.Diagramme de Williams

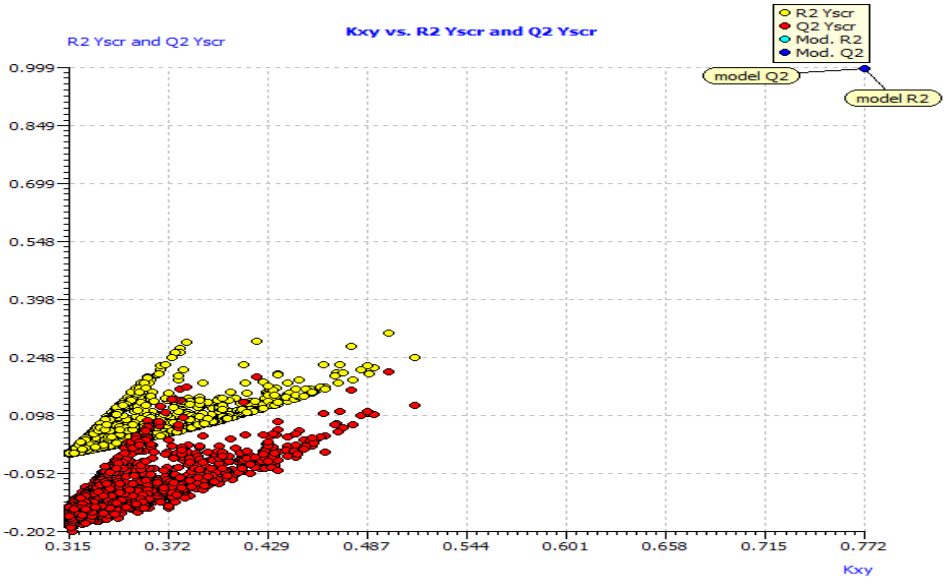


Figure 3. Test de randomisation

III, Séparation selon le premier axe de l'analyse en composantes principales

Paramètre statistique du modèle

Variable	Coeff,	Std, coeff,	Std, err,	(+/-) Co, int, 95%	p
Intercept	459,726		31,759	64,239	0,000
HOMO	22,723	0,061	3,440	6,958	0,000
EPS0	34,495	0,9576	0,332	0,673	0,000

Critères de fitness

R2: 0,9981	R2adj: 0,9980	R2-R2adj: 0,0001	LOF: 27,5490
Kxx: 0,6629	Delta K: 0,1279	RMSE tr: 4,7488	MAE tr: 3,8112
RSS tr: 947,1598	CCC tr: 0,9991	s: 4,9281	F: 10435,9994

Critère de validation interne

Q2loo: 0,997	R2-Q2loo: 0,000	RMSE cv: 5,092
PRESS cv: 1089,028	CCC cv: 0,998	MAE cv: 4,093
Q2LMO: 0,997	R2Yscr: 0,048	Q2Yscr: -0,104

Critère de validation externe

RMSE ext: 4,747	MAE ext: 3,192	PRESS ext: 428,215	R2ext: 0,998
Q2-F1: 0,998	Q2-F2: 0,998	Q2-F3: 0,998	
r2m aver,: 0,997	r2m delta: 0,001	CCC ext: 0,999	

Aver = moyenne

Angle de régression des données externes calculé à partir de la diagonale: 0,0494°

Tableau III Valeurs de (Teb) et statut des composés

ID	Nom	Status	Teb Exp,	ID	Nom	Status	Teb Exp,
1	2-methylnaphthalene	Tr	514	32	benzo[b]fluoranthene	Pr	754
2	1-methylnaphthalene	Tr	518	33	benzo[k]fluoranthene	Tr	754
3	2,6-dimethylnaphthalene	Tr	535	34	benzo[e]pyrene	Tr	769
4	2,7-dimethylnaphthalene	Tr	535	35	picene	Pr	792
5	1,7-dimethylnaphthalene	Pr	536	36	aq Benzo[a]pyrene	Pr	769
6	1,3-dimethylnaphthalene	Tr	538	37	Perylene	Tr	770
7	1,6-dimethylnaphthalene	Tr	539	38	Dibenz[a,h]anthracene	Tr	808
8	2,3-dimethylnaphthalene	Pr	541	39	Benzo[ghi]perylene	Pr	815
9	1,4-dimethylnaphthalene	Tr	541	40	Coronene	Tr	863
10	1,5-dimethylnaphthalene	Pr	542	41	Naphtho[1,2,3,4-def]chrysene	Tr	865
11	1,2-dimethylnaphthalene	Tr	544	42	Benzo[rst]pentaphene	Tr	867
12	2,3,5-trimethylnaphthalene	Pr	558	43	Dibenzo[def,p]chrysene	Pr	868
13	2,3,6-trimethylnaphthalene	Tr	559	44	Dibenzo[b,def]chrysene	Tr	869
14	2-methylfluorene	Tr	591	45	Benzo[c]fluorene	Pr	679
15	3-methylphenanthrene	Pr	625	46	Dibenzo[def,mno]chrysene	Tr	820
16	2-methylphenanthrene	Tr	628	47	Dibenz[a,j]anthracene	Tr	804
17	9-methylphenanthrene	Tr	628	48	Indeno[1,2,3-cd] fluoranthene	Pr	804
18	2-methylanthracene	Tr	632	49	Indeno[1,2,3,cd]pyrene	Tr	807
19	1-methylphenanthrene	Tr	632	50	Benzo[b]triphenylene	Pr	808
20	1-methylanthracene	Tr	636	51	Benzo[j]fluoranthene	Tr	753
21	3,6-dimethylphenanthrene	Pr	636	52	Cyclopenta[cd]pyrene	Tr	712
22	4-methylpyrene	Pr	683	53	naphthalene	Tr	491
23	2-methylpyrene	Tr	683	54	acenaphthylene	Tr	543
24	1-methylpyrene	Tr	683	55	acenaphthene	Pr	552
25	benzo[b]fluorene	Tr	675	56	fluorene	Pr	567
26	benzo[a]fluorene	Tr	680	57	phenanthrene	Tr	611
27	benzo[ghi]fluoranthene	Tr	705	58	anthracene	Tr	613
28	benz[a]anthracene	Pr	708	59	4H-cyclopenta[def]phenanthrene	Pr	632
29	triphenylene	Tr	712	60	fluoranthene	Tr	656
30	chrysene	Tr	714	61	pyrene	Tr	666
31	naphthacene	Tr	723				

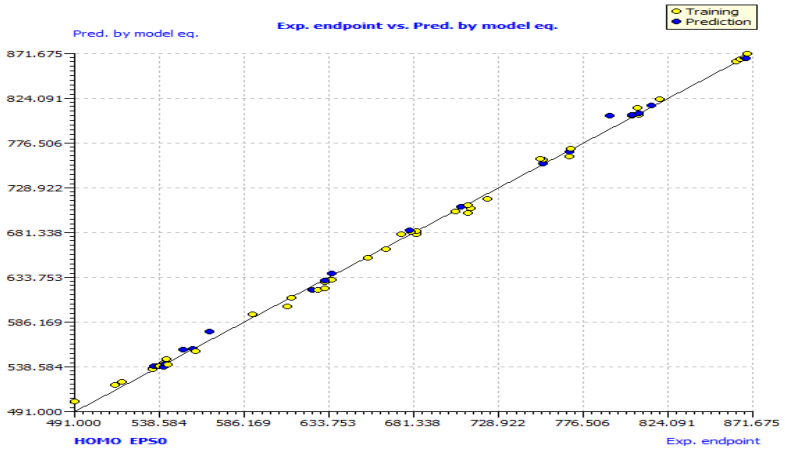


Figure 4. Droite d'ajustement

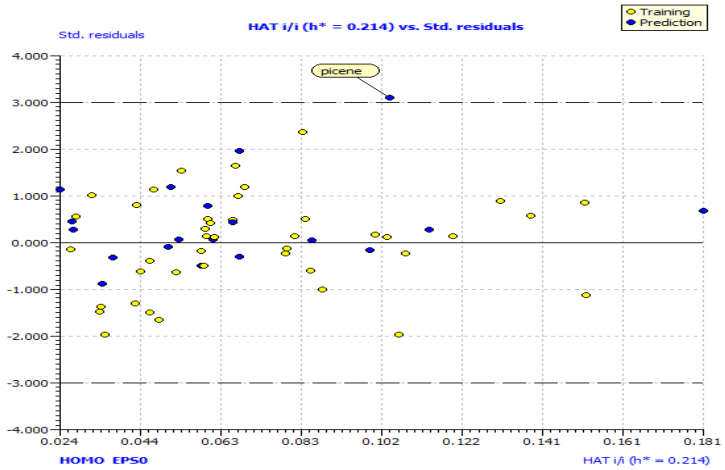


Figure 5. Diagramme de Willimas,

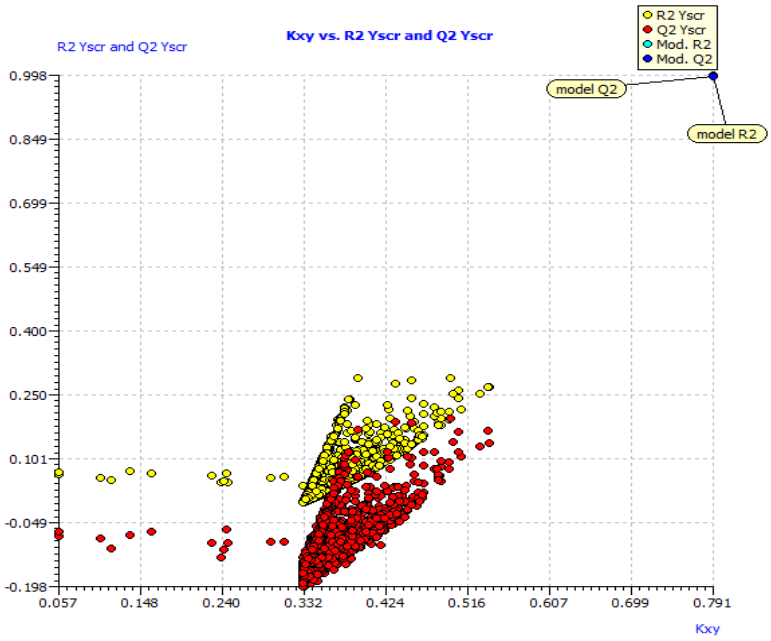


Figure 6. Test de randomisation

IV. Modèle Complet

Paramètres statistiques du modèle complet

Variable	Coeff,	Std, coeff,	Std, err,	(+/-) Co, int, 95%	p
Intercept	482,0875		26,4671	52,9796	0,000
HOMO	25,1756	0,0632	2,8879	5,7808	0,000
EPS0	34,3009	0,9578	0,2598	0,520	0,000

Critères de fitness

R2: 0,9982	R2adj: 0,9981	R2-R2adj: 0,0001	LOF: 24,9910
Kxx: 0,6348	Delta K: 0,1395	RMSE tr: 4,6713	MAE tr: 3,5059
RSS tr: 1331,0786	CCC tr: 0,9991	s: 4,7906	F: 15890,2957

Critère de validation interne

Q2loo: 0,9980	R2-Q2loo: 0,0002	RMSE cv: 4,9031
PRESS cv: 1466,4845	CCC cv: 0,9990	MAE cv: 3,6798

Tableau IV, Experimental and predicted values of boiling point (full model).

ID	Name	St	Teb Exp	Pred, model	erreur	ID	Name	Stat	Teb Exp	Pred, model	Residual
1	2-Methylnaphthalene	Tr	514	520,382	6,382	**32**	benzo[b]fluoranthene	Tr	754	762,547	8,547
2	1-Methylnaphthalene	Tr	518	522,241	4,241	**33**	benzo[k]fluoranthene	Tr	754	760,653	6,653
3	2,6-diMethylnaphthalene	Tr	535	536,357	1,357	**34**	benzo[e]pyrene	Tr	769	762,547	-6,453
4	2,7-diMethylnaphthalene	Tr	535	536,357	1,357	**35**	picene	Tr	792	811,187	19,187
5	1,7-diMethylnaphthalene	Tr	536	538,252	2,252	**36**	Benzo[a]pyrene	Tr	769	760,653	-8,347
6	1,3-diMethylnaphthalene	Tr	538	538,252	0,252	**37**	Perylene	Tr	770	762,547	-7,453
7	1,6-diMethylnaphthalene	Tr	539	538,252	-0,749	**38**	Dibenz[a,h	Tr	808	809,329	1,329
8	2,3-diMethylnaphthalene	Tr	541	538,252	-2,749	**39**	Benzo[ghi]perylene	Tr	815	814,261	-0,740
9	1,4-diMethylnaphthalene	Tr	541	540,110	-0,890	**40**	Coronene	Tr	863	866,010	3,010
10	1,5-diMethylnaphthalene	Tr	542	540,110	-1,890	**41**	Naphtho[1,2,3,4-def	Tr	865	864,795	-0,205
11	1,2-diMethylnaphthalene	Tr	544	540,110	-3,890	**42**	Benzo[rst	Tr	867	862,937	-4,064
12	2,3,5-	Tr	558	556,121	-1,879	**43**	Dibenzo[def,p	Tr	868	864,795	-3,205
13	2,3,6-	Tr	559	554,227	-4,773	**44**	Dibenzo[b,def	Tr	869	862,937	-6,064
14	2-Methylfluorene	Tr	591	597,363	6,363	**45**	Benzo[c]fluorene	Tr	679	683,672	4,672
15	3-Methylphenanthrene	Tr	625	622,630	-2,370	**46**	Dibenzo[def,mno]	Tr	820	812,402	-7,598
16	2-Methylphenanthrene	Tr	628	622,630	-5,370	**47**	Dibenz[a,j	Tr	804	809,329	5,329
17	9-Methylphenanthrene	Tr	628	624,524	-3,476	**48**	Indeno[1,2,3-cd]	Tr	804	816,155	12,155
18	2-Methylanthracene	Tr	632	620,772	-	**49**	Indeno[1,2,3-cd]	Tr	807	814,261	7,261
19	1-Methylphenanthrene	Tr	632	624,524	-7,476	**50**	Benzo[b]triphenylene	Tr	808	811,187	3,187
20	1-Methylanthracene	Tr	636	622,630	-	**51**	Benzo[j	Tr	753	762,547	9,547
21	3,6-	Tr	636	638,641	2,641	**52**	Cyclopenta[cd	Tr	712	710,118	-1,882
22	4-Methylpyrene	Tr	683	676,238	-6,762	**53**	naphthalene	Tr	491	504,371	13,371
23	2-Methylpyrene	Tr	683	674,380	-8,620	**54**	acenaphthylene	Tr	543	556,121	13,121
24	1-Methylpyrene	Tr	683	676,238	-6,762	**55**	acenaphthene	Tr	552	556,121	4,121
25	benzo[b]fluorene	Tr	675	681,778	6,778	**56**	fluorene	Tr	567	581,388	14,388
26	benzo[a]fluorene	Tr	680	683,672	3,672	**57**	phenanthrene	Tr	611	606,655	-4,345
27	benzo[ghi]fluoranthene	Tr	705	711,977	6,977	**58**	anthracene	Tr	613	604,761	-8,239
28	benz[a]anthracene	Tr	708	707,045	-0,955	**59**	4H-cyclopenta[def]	Tr	632	633,102	1,102
29	triphenylene	Tr	712	710,797	-1,203	**60**	fluoranthene	Tr	656	660,263	4,263
30	chrysene	Tr	714	708,939	-5,061	**61**	pyrene	Tr	666	658,369	-7,631
31	naphthacene	Tr	723	705,186	-						

Tableau V, Valeurs expérimentales et prédites du point d'ébullition pour les produits chimiques avec des données inconnues

ID	Name	Status	Teb Exp	Pred, par eq, model,	H i/i (h*=0,1475)
1	2-methylnaphthalene	Training	514	519,209	0,0492
2	1-methylnaphthalene	Training	518	522,3773	0,0477
3	2,6-dimethylnaphthalene	Training	535	537,0843	0,0424
4	2,7-dimethylnaphthalene	Training	535	535,9514	0,0422
5	1,7-dimethylnaphthalene	Training	536	539,4813	0,0422
6	1,3-dimethylnaphthalene	Training	538	539,7582	0,0427
7	1,6-dimethylnaphthalene	Training	539	539,3554	0,0421
8	2,3-dimethylnaphthalene	Training	541	538,2477	0,0414
9	1,4-dimethylnaphthalene	Training	541	543,2538	0,0463
10	1,5-dimethylnaphthalene	Training	542	542,1461	0,0431
11	1,2-dimethylnaphthalene	Training	544	541,3405	0,0415
12	2,3,5-trimethylnaphthalene	Training	558	558,0919	0,0374
13	2,3,6-trimethylnaphthalene	Training	559	555,695	0,037
14	2-methylfluorene	Training	591	594,1003	0,0304
15	3-methylphenanthrene	Training	625	619,9875	0,0239
16	2-methylphenanthrene	Training	628	620,0882	0,0236
17	9-methylphenanthrene	Training	628	619,8417	0,0322
18	2-methylanthracene	Training	632	629,5329	0,0642
19	1-methylphenanthrene	Training	632	621,6544	0,0244
20	1-methylanthracene	Training	636	631,6438	0,0665
21	3,6-dimethylphenanthrene	Training	636	637,5194	0,0187
22	4-methylpyrene	Training	683	682,5917	0,0396
23	2-methylpyrene	Training	683	679,9521	0,0342
24	1-methylpyrene	Training	683	683,7246	0,0485
25	benzo[b]fluorene	Training	675	679,0465	0,0208
26	benzo[a]fluorene	Training	680	681,8967	0,0184
27	benzo[ghi]fluoranthene	Training	705	702,3411	0,0717
28	benz[a]anthracene	Training	708	708,8611	0,0199
29	triphenylene	Training	712	701,2596	0,0705
30	chrysene	Training	714	706,4244	0,0218
31	naphthacene	Training	723	718,5827	0,1207
32	benzo[b]fluoranthene	Training	754	753,6715	0,0726
33	benzo[k]fluoranthene	Training	754	758,4999	0,0296
34	benzo[e]pyrene	Training	769	761,9291	0,0276
35	picene	Training	792	805,0377	0,0643
36	aq Benzo[a]pyrene	Training	769	767,5128	0,0539
37	Perylene	Training	770	770,6902	0,0654
38	Dibenz[a,h]anthracene	Training	808	805,7968	0,0491
39	Benzo[ghi]perylene	Training	815	816,497	0,0467
40	Coronene	Training	863	862,3884	0,0751
41	Naphtho[1,2,3,4-def]chrysene	Training	865	864,9733	0,067
42	Benzo[rst]pentaphene	Training	867	867,0415	0,0756
43	Dibenzo[def,p]chrysene	Training	868	866,509	0,0686
44	Dibenzo[b,def]chrysene	Training	869	871,5732	0,1087
45	Benzo[c]fluorene	Training	679	683,9863	0,0166
46	Dibenzo[def,mno]chrysene	Training	820	824,2045	0,123
47	Dibenz[a,j]anthracene	Training	804	805,1926	0,0517
48	Indeno[1,2,3-cd] fluoranthene	Training	804	805,1229	0,1143

49	Indeno[1,2,3,cd]pyrene	Training	807	814,0046	0,0438
50	Benzo[b]triphenylene	Training	808	806,9762	0,0528
51	Benzo[j]fluoranthene	Training	753	759,3108	0,0334
52	Cyclopenta[cd]pyrene	Training	712	710,6781	0,0185
53	naphthalene	Training	491	501,5512	0,06
54	acenaphthylene	Training	543	545,6804	0,0977
55	acenaphthene	Training	552	557,387	0,0361
56	fluorene	Training	567	575,2935	0,0492
57	phenanthrene	Training	611	602,1122	0,0343
58	anthracene	Training	613	612,6807	0,0588
59	4H-cyclopenta[def]phenanthrene	Training	632	629,6097	0,0256
60	fluoranthene	Training	656	653,9413	0,0395
61	pyrene	Training	666	664,1069	0,0356
62	1,8-dimethylnaphthalene	Unknown		543,1783	0,0461
63	2-ethylbiphenyl	Unknown		596,7933	0,0517
64	1-methylacenaphthylene	Unknown		568,2688	0,0486
65	9-ethylfluorene	Unknown		624,3464	0,0541
66	1,2,3,4,5,6,7,8-octahydroanthracene	Unknown		596,5179	0,0618
67	1-methylfluorene	Unknown		593,2496	0,0446
68	1,2,3,4,5,6,7,8-octahydrophenanthrene	Unknown		596,6239	0,0801
69	1,2,3,4-tetrahydrophenanthrene	Unknown		605,2592	0,0236
70	1,2,3,10b-tetrahydrofluoranthene	Unknown		652,4307	0,0518
71	9-n-propylfluorene	Unknown		648,4553	0,0637
72	9-n-butylfluorene	Unknown		672,706	0,0747
73	4,5,9,10-tetrahydropyrene	Unknown		653,6842	0,0293
74	4,5-dihydropyrene	Unknown		653,9863	0,0277
75	2-phenylnaphthalene	Unknown		651,7253	0,0243
76	9-ethylphenanthrene	Unknown		651,3703	0,0272
77	2-ethylphenanthrene	Unknown		647,8496	0,0369
78	2,7-dimethylphenanthrene	Unknown		633,9948	0,0304
79	1,2,3,6,7,8-hexahydropyrene	Unknown		661,4635	0,0222
80	9-isopropylphenanthrene	Unknown		663,257	0,0276
81	1,8-dimethylphenanthrene	Unknown		640,6679	0,0194
82	9-n-hexylfluorene	Unknown		721,132	0,1065
83	9-n-propylphenanthrene	Unknown		675,5959	0,0325
84	9,10-dimethylanthracene	Unknown		656,7689	0,1074
85	11-methylbenzo[a]fluorene	Unknown		700,856	0,0214
86	4,5,6-trihydrobenz[de]anthracene	Unknown		681,1162	0,0203
87	5,12-dihydronaphthacene	Unknown		696,7555	0,0586
88	1-ethylpyrene	Unknown		714,4315	0,0466
89	2,7-dimethylpyrene	Unknown		695,7721	0,0334
90	1,2,3,4,5,6,7,8,9,10,11,12-dodecahydrotriphenylene	Unknown		700,0008	0,0852
91	benzo[c]phenanthrene	Unknown		717,5017	0,0602
92	1,2'-binaphthyl	Unknown		754,7192	0,0309
93	11-methylbenz[a]anthracene	Unknown		727,7235	0,0246
94	2-methylbenz[a]anthracene	Unknown		725,5874	0,0235
95	1-methylbenz[a]anthracene	Unknown		726,6913	0,0219
96	1-n-butylpyrene	Unknown		762,8827	0,0405
97	1-methyltriphenylene	Unknown		720,7514	0,0566
98	9-methylbenz[a]anthracene	Unknown		725,235	0,0226
99	3-methylbenz[a]anthracene	Unknown		725,0084	0,022

100	8-methylbenz[a]anthracene	Unknown		727,3711	0,0236
101	6-methylbenz[a]anthracene	Unknown		727,6732	0,0245
102	3-methylchrysene	Unknown		722,9653	0,0223
103	5-methylbenz[a]anthracene	Unknown		727,3962	0,0236
104	2-methylchrysene	Unknown		722,8143	0,0226
105	12-methylbenz[a]anthracene	Unknown		730,5485	0,0284
106	4-methylbenz[a]anthracene	Unknown		726,6913	0,0219
107	5-methylchrysene	Unknown		725,7651	0,021
108	6-methylchrysene	Unknown		726,319	0,0205
109	4-methylchrysene	Unknown		725,7651	0,021
110	1-methylchrysene	Unknown		725,3623	0,0216
111	7-methylbenz[a]anthracene	Unknown		730,5737	0,0285
112	2,2'-binaphthyl	Unknown		751,4754	0,0361
113	1,3-dimethyltriphenylene	Unknown		737,7798	0,0507
114	1,12-dimethylbenz[a]anthracene	Unknown		748,0766	0,0301
115	7,12-dimethylbenz[a]anthracene	Unknown		752,1512	0,0419
116	1,6,11-trimethyltriphenylene	Unknown		756,36	0,0375
117	dibenz[a,c]anthracene	Unknown		806,9762	0,0528
118	benzo[b]chrysene	Unknown		810,8823	0,0433

ANNEXE II

Température de fusion

Tableau.1 Valeurs expérimentales et prédites des composés de test utilisés lors la construction du modèle RNA.

Num	Composés	Tfus exp	Tfus Préd. par RNA
3	2,3,5-trimethylnaphthalene	298	297.0634
6	4-methylphenanthrene	323	323,1604
8	1-methylfluorene	360	363,5471
9	9-methylphenanthrene	364	315,1914
12	azulene	373	356,9995
13	2-phenylnaphthalene	377	418,9557
17	fluorene	390	389,1962
19	4-methylpyrene	421	402,5794
21	pyrene	429	424,528
22	1-methylchrysene	434	476,6675
25	benzo[j]fluoranthene	439	423,6491
31	benzo[a]fluorene	463	456,1848
34	2-methylanthracene	482	464,6519
35	benzo[b]fluorene	482	475,4772
36	anthracene	489	504,4739
39	naphthacene	530	542,1575
42	dibenz[a,h]anthracene	543	541,9206
44	perylene	551	513,3119
46	benzo[b]chrysene	567	551,9998
48	2,6-dimethylanthracene	523	518,5576
51	coronene	633	636,1986
52	indene	271	275,1573