MINISTERE DE L'ENSEIGNEMENT SUPERIEUR
ET DE LA RECHERCHE SCIENTIFIQUE
UNIVERSITE«Abbès LAGHROUR»DE KHENCHELA
FACULTE DES SCIENCES ET TECHNOLOGIE

Département des Sciences de la Matière

N° de série :……

Mémoire de fin d'études

Pour l'obtention du diplôme de Master (L.M.D)

Filière : Chimie

Spécialité : Chimie analytique et environnement

Intitulé :

Etude QSRR de la rétention chromatographique des hydrocarbures aromatiques polycycliques (HAP)

Réalisé par : - REMACHE Sirid
- REDAH Ouahiba

Dirigé par Dr. KERTIOU Nourddine

Membres de jury :
HAKKAR Farida MAA. Présidente
BOUAKKADIA Amel Dr. Examinatrice

Présenté le 25Juin2018

DEDICACES

Louange à dieu qui m'a donné l'effort et la volonté.

Par cette première page, Je dédie ce manuscrit à toutes les personnes qui m'ont aidé et soutenu pendant ces dernières années.

Je dédie ce modeste travail :

- A ma très chère mère et mon très cher père.
- A ma femme de m'avoir soutenu et encouragé.
- A mes filles : Selsabil, Anfal, Takoua.
- A mon fils Wassim
- A mes chères frères et sœurs.
- A tous les étudiants de la promotion 2018.

Mr : REMACHE Sirid

DEDICACES

Je remercie dieu qui m'a donné l'effort et la volonté.

Je dédie se modeste travail :

- **A ma mère que dieu y pitié de son âme.**
- **A mon époux djebaili Ahmed.**
- **A mes enfants abdelouaheb, farah, ritej.**
- **A tous les étudiants de la promotion 2018.**

Mme : REDAH Ouahiba

REMERCIEMENTS

Ce mémoire n'aurait pas vu le jour sans la confiance, la patience et la générosité du Mr. L'encadreur Dr. KERTIOU Noureddine que nous tenons à lui exprimer nos profondes gratitudes pour avoir dirigé ce mémoire, avec un encadrement sérieux et efficace.

Comme nous tenons à remercier la présidente de jury Mme HAKKAR Farida MAA et le membre de jury : Dr.Bouakkadia Amel Examinatrice et tous nos professeurs qui nous en enseignés durant les années 2016-2018.

Ainsi que nos amis de promotion qui nous ont aidés.

Mr. REMACHE Sirid

Mme.REDAH Ouahiba

Etude QSRR de la rétention chromatographique des HAP

Partie application

LISTE DES TABLEAUX

LISTE DES FIGURES

LISTE DES SYMBOLES ET ABREVIATIONS

ei : Résidu différence entre les valeurs observées (y_i) et estimée ($\hat{y}_i$).

F : Statistique de Fisher.

H : Matrice de projection, ou matrice chapeau.

hii : Eléments diagonaux de la matrice chapeau.

Ir : Indice de rétention.

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

MLR: Régression linéaire multiple.

n: Dimension de la population (échantillon).

n-p : Nombre de degrés de liberté.

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

PP : Méthode pas à pas.

PLS: Moindres carrés partiels.

PRESS: Somme des carrés des erreurs de prédiction.

OM : Orbit molucular.

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

QSAR : Quantitative Structure/ Activity Relationships.Relations Structure/ Activités Quantitatives).

QSPR : Quantitative Structure/ ProprietyRelationships.Relations Structure/ Propriétés Quantitatives).

QSRR : Quantitative Structure/ RetentionRelationships.Relations Structure/ Rétentions Quantitatives).

R^2 : Coefficient de détermination.

S : Erreur standard.

SCE : Somme des carrés des écarts.

SCT : Somme des carrés totale.

SDEC: Standard Deviation Error in Calculation: Déviation standard de l'erreur calculée.

SDEP : Standard Deviation Error of Prediction: Déviation standard de l'erreur de prédiction.

SDEPext : External Standard Deviation Error of Prediction: Déviation standard de

l'erreur de prédiction externe.

t : t de Student.

$\underset{\sim}{X}$: Matrice des valeurs observées des variables explicatives.

$\underset{\sim}{X}'$: Matrice transposée de $\underset{\sim}{X}$.

y : Vecteur de dimension n.

yi : Valeur observée.

$\hat{y}_i$: Valeur estimée.

α : Niveau de confiance.

INTRODUCTION GENERALE

INTRODUCTION GENERALE

Les Hydrocarbures Aromatiques Polycycliques (HAP) sont des molécules organiques issues de la combustion incomplète de matières carbonées, suite à des processus naturels (volcanisme), mineurs, et des processus anthropiques, majoritaires. Ils sont libérés dans tous les compartiments de l'environnement. Désormais largement répandus, la concentration des HAP dans l'environnement a fortement augmenté depuis les 150 dernières années [1].

Les HAP sont des molécules composées d'au moins deux noyaux aromatiques accolés. Ils diffèrent par le nombre de noyaux accolés ainsi que par leur agencement [2]. Les HAP sont particulièrement suivis par les procédures législatives en raison de leurs propriétés toxiques et carcinogènes à faibles concentrations [3].

Le couplage chromatographie gazeuse /spectrométrie de masse, facilite souvent les questions d'identification, mais peut être inefficace dans l'analyse des isomères ou des composés mineurs d'un mélange complexe .Les relations structure / rétention peuvent, dans ces conditions, aider à l'identification. Il s'agit de relier les réponses obtenues pour un ensemble d'évaluation à des propriétés physico-chimiques expérimentales ou théoriques, et/ou des descripteurs moléculaires de différents types fournis par divers logiciels spécialisés. Les techniques les plus courantes pour établir des modèles QSPR utilisent l'analyse de régression (régression multilinéaire : MLR ; projection des structures latentes par les moindres carrés partiels : PLS), les réseaux neuronaux RNA, et les méthodes de classification.

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

Des logiciels informatiques spécialisés permettent le calcul de plus de 3000 descripteurs moléculaires appartenant à différentes classes. Plutôt que de rechercher à expliquer la variable dépendante (propriété physique considérée) par tous les régresseurs (descripteurs moléculaires), on peut chercher seulement un ensemble réduit de régresseurs qui donne une reconstitution aussi satisfaisante de la variable à expliquer Parmi les

stratégies mises en œuvre pour la sélection d'un ensemble limité de variables explicatives, on peut citer : les méthodes de pas à pas, ainsi que les algorithmes évolutifs et génétiques.

Nous suggérons dans ce travail d'appliquer la méthodologie QSPR/QSRR pour la modélisation de : l'indice de rétention d'une série de 50 HAP. Nous avons choisi la méthode de régression multiple (MLR).

Notre mémoire comporte en plus de la bibliographie, une introduction et une conclusion générales, deux grandes parties :

Dans la partie théorique, nous avons exposé des généralités sur les HAP, leurs définition, leurs source et mécanisme de formation, aussi, nous avons exposé brièvement la chromatographie gazeuse et son appareillage, ainsi, développé tout ce qui a trait au prétraitement des molécules (introduction des molécules, optimisation de leurs géométrie) en vue du calcul des descripteurs moléculaires à l'aide de différents logiciels de modélisation moléculaire. Nous y avons également développé les connaissances théoriques de base utilisées tout au long de ce travail : régression multilinéaire et traitement statistique pour l'évaluation de modèle (robustesse de modèle, domaine d'application, test de randomisation et validation externe).

Dans la partie application, nous présentons et discutons le modèle traité.

PARTIE THEORIQUE

I- GENERALITE SUR LES HAP

Les HAP (Hydrocarbures Aromatiques Polycycliques) sont quotidiennement présents dans notre proche environnement. En raison de leur pression de vapeur, de nombreux HAP présents dans l'atmosphère existent simultanément sous forme gazeuse et particulaire (HAP adsorbés et/ou absorbés aux particules). Leur impact sanitaire et leur comportement dans l'environnement, diffèrent selon la forme considérée.

Ce sont des composés organiques dont la structure cyclique comprend au moins deux cycles aromatiques. Le nombre théorique de HAP susceptibles d'exister s'élève à plus de 1000. Seulement plus d'une centaine de HAP différents y ont été identifiés. Parmi ces HAP, 16 d'entre eux sont couramment analysés dans les différentes composantes de l'environnement, selon les recommandations de l'Agence Américaine de l'Environnement (USEPA) [4].

I.1- Chimie des HAP :

I.1.1- Sources et mécanisme de formation des HAP :

Les HAP proviennent essentiellement de phénomènes de pyrolyse-présynthèses de la matière organique (combustibles fossiles, bois …). Les mécanismes conduisant à la formation de HAP par pyrolyse-présynthèse ne sont pas encore totalement connus. Ils semblent toutefois impliquer la fragmentation des substances organiques en composés instables, sous l'effet de la température. Ces fragments, principalement des radicaux libres très réactifs, ont des temps de vie très courts. Une partie d'entre eux vont réagir avec l'oxygène présent pour former du CO_2 et de l'eau. Mais l'oxygène étant généralement insuffisant pour accomplir une oxydation totale, une partie de ces fragments vont réagir entre eux. La recombinaison de ces fragments va conduire lors du refroidissement à des composés organiques de plus en plus complexes.

Ces mécanismes autorisent la formation d'une grande variété de HAP de masse molaire comprise entre 78 (C_6H_6) et 1792 g.mol-1 ($C_{144}H_{64}$). Généralement, la nature et l'abondance des HAP formés va dépendre de paramètres tels que la composition du combustible de base (le rendement de formation des HAP augmente avec la concentration d'aromatiques, d'alcènes cycliques, d'alcènes, et d'alcanes), de la proportion d'oxygène, et de la température de combustion (plus la température est élevée, moins des HAP alkylés seront formés).

Un autre mode de formation des HAP provient de la formation géologique des combustibles fossiles tels que le pétrole ou le charbon lors de la dégradation des

substancesorganiques, à pression élevée et à température réduite (inférieure à 200 °C). En raison de la température relativement basse, les HAP sont formés plus lentement et la proportion de HAP alkylés augmente.

Dans notre environnement, il apparaît que les sources de HAP sont principalement anthropiques tableau(01) bien qu'épisodiquement, des processus de combustion naturelle (feux de forêt, volcans) puissent être à l'origine d'une grande production de HAP [4].

Tableau01 : Sources anthropiques des HAP

Sources stationnaires industrielles	Sources domestiques	Sources mobiles
Production d'aluminium	Chauffage (gaz naturel.GPL.bois.charbon	Voitures
Fabrication de pneu	Tabagisme	Avions
Créosotes et préservation du bois	Cuisson des aliments (barbecue. Friture)	Trains
Sidérurgie		Bateaux
Industrie du bitume et goudrons		
Cimenteries		
Moteurs à combustion		
Industries pétrochimiques et similaires		
Chauffage et électricité		
Incinérateurs de déchets ménagers et industriels		

Les sources nombreuses et variées des HAP sont à l'origine d'une présence assez importante dans l'environnement, à la fois dans les eaux (surtout dans les sédiments et les matières en suspension), dans les sols et dans l'air ambiant. On trouve beaucoup des HAP dans les goudrons issus de la houille et les produits qu'ils traitent (asphalte, plaques bitumées, colorants organiques…). Les HAP d'origine fossile rentrent également dans la composition des huiles de dilution, qui sont mélangées aux caoutchoucs utilisés dans la fabrication des pneus, par exemple. Ce sont cependant lesprocessus de combustion qui sont la source principale des HAP présents dans l'air ambiant. Ils sont alors liés aux particules

de suie, à partir de la combustion incomplète du charbon, des carburants, du bois, du tabac etc.

Les HAP font partie des Polluants Organiques Persistants (POP)[4],car ils se caractérisent par les quatre propriétés suivantes :

A) **Toxicité** : elles présentent un ou plusieurs impacts prouvés sur la santé humaine
B) **Persistance dans l'environnement :** ils sont généralement peu dégradés dans l'environnement naturel ou par les organismes vivant
C) **Bioaccumulation :** ces molécules s'accumulent dans les tissus vivants du fait de leur faible solubilité aqueuse et leur forte solubilité dans les lipides. De façon générale, lorsque la masse moléculaire de ces composés augmente, leur solubilité dans l'eau diminue alors que leur caractère lipophile augmente.
D) **Transport longue distance :** du fait de leurs propriétés de persistance et de bioaccumulation, ces composés semi-volatils peuvent se déplacer sur de longues distances et se déplacer loin des lieux d'émission, typiquement des milieux chauds (à forte activité humaine) vers les milieux froids.

I.1.2- Les propriétés physico-chimiques

Les propriétés physico-chimiques se révèlent très utiles pour évaluer l'impact potentiel des HAP dans l'environnement. Elles vont notamment permettre de mieux prévoir leur répartition, ainsi que leur comportement dans les différents compartiments de l'environnement (eau, sol, sédiments, atmosphère, végétaux, êtres vivants). Les HAP sont des composés non polaire, stables et ayant une faible volatilité. Il en résulte que les HAP sont hydrophobes, et persistants. Généralement, ces tendances s'accentuent lorsque la masse molaire augmente.

D'après les paramètres caractéristiques des HAP, une fois émis dans l'atmosphère, ces composés vont avoir tendance à s'accumuler dans les différents compartiments solides de l'environnement (sol, sédiment, matières en suspension). De plus, leur caractère lipophile leur permet d'être facilement transférés dans les différents compartiments de la chaîne alimentaire.

La plupart des HAP sont peu volatils, très peu solubles dans l'eau, peu mobiles dans le sol car facilement adsorbés. Ces substances sont stables (hydrolyse négligeable) mais leur biodégradabilité varie fortement selon les conditions du milieu. Etant

hydrophobes, liposolubles et généralement peu volatils, les HAP ont tendance à s'adsorber sur les matrices solides et notamment les matières organiques [4].

Les principaux paramètres couramment utilisés pour prédire la distribution des composés organiques, dans les différents compartiments environnementaux sont :

Tableau 02 : Les principaux paramètres physico-chimiques pour prédire la distribution des HAP, dans les différents compartiments environnementaux.

Propriétés physiques	**Tensiondevapeur**	Reflètelavolatilitéetdonclacapacitéd'uncomposéàresteren phasegazeuseouàsevolatiliser.
	Solubilité	Donneuneidéedela capacitéd'unemoléculeorganique àsedissoudredansl'eau. (Caractère Polaire) De l'ordre du µg/l: solubilité faible De l'ordre du mg/l solubilité moyenne De l'ordre du g/l: solubilité importante Engénéral,lesHAPont unefaiblesolubilité,comprise entre30mg/lpourlescomposéslégerset10-4mg/l pour les plus lourds.
	constante de Henry (K_H) : $C_{air\ (eq)}/C_{eau\ (eq)}$	Caractéristiquedel'équilibreentrelesphasesgazeuseetaque use
	Coefficient de partage du carbone organique (K_{OC})	Indique la propension des HAP à se lier à la matièreorganiquedusoloudusédiment
	Le coefficient de partage octanol-eau(K_{OW}) $K_{ow}=C_{octanol}/C_{eau}$	Affinitéd'uncomposé pourlamatière organique. Prévoirleurbioaccumulation. Estimerlamigrationdes HAPversdeslipides. Permet d'évaluer le caractère polaire des molécules logK_{OW}<1,5: substances non-bioaccumulables
	Facteurde Bioconcentration $BCF=C_{mat.vivante}/C_{eau}$	Donne la tendance qu'a une molécule à se bio-accumuler dans un organisme vivant donné.

Tableau02: suite

Propriétés chimiques *LesHAP peuventêtre classésen troisgroupes basés sur le nombre de cycles aromatiques qu'ils contiennent et leurs masses molaires moléculaires*	**HAP de faibles masses molairesmoléculaires**	De l'ordrede 152-178 g/mol, soit 2 à 3 cycles) naphtalène, acénaphtylène ,acénaphtène, fluorène, anthracène et phénanthrène– ***solubilitéetvolatilité la plus élevée***,
	HAP de masses molaires moléculairesintermédiaires	del'ordrede202g/mol,4 cycles) :fluoranthène, pyrène
	HAP à masses molaires moléculaires élevées	del'ordrede 228-278 g/mol,soit4à 6cycles: benzo(a)anthracène, chrysène,benzo(a)pyrène, benzo(b)fluoranthène, dibenzo(ah)anthracène, benzo(k)fluoranthène, benzo(ghi)pérylène, indéno(1,2,3, cd)pyrène –***sorption laplus forte***.

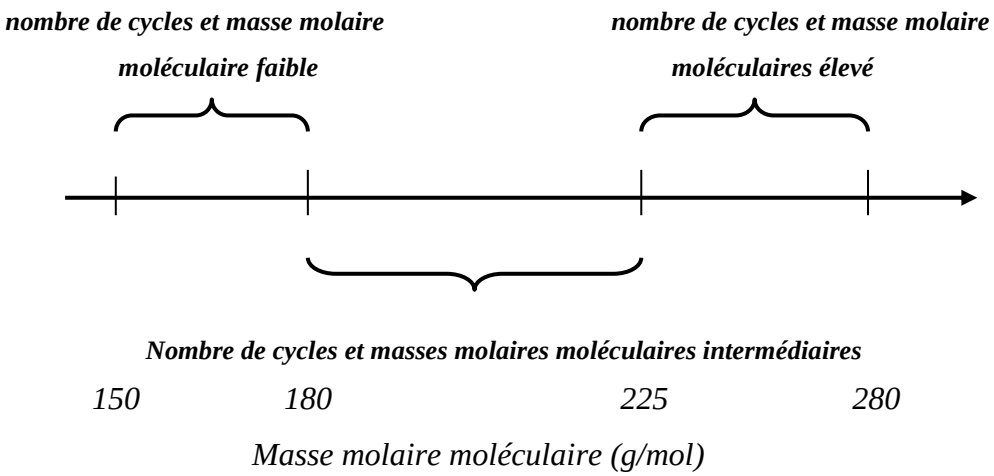


Figure 1 : valeurs réglementaires et recommandationde l'OMS [4].

I.2-Toxicité des HAP

I.2.1- Propriétés cancérogènes et mutagènes

La toxicité des HAP peut être aiguë, faible ou modérée selon le composé considéré aux vues des concentrations auxquelles sont exposées les populations, les risques toxiques associés aux HAP sont généralement liés à une exposition chronique. Les risques les plus importants liés aux HAP sont leur effet mutagène et cancérigène. En effet, certains d'entre eux ont été classés comme cancérogènes probables ou possibles chez l'humain par le Centre international de recherche sur le cancer (CIRC), l'USEPA, et l'Union européenne.

Plusieurs mélanges de HAP en atmosphère de travail ont été également classés comme cancérogènes pour l'homme.Parmi les HAP, la toxicité du benzo(a)pyrène est la mieux documentée et la plus mesurée. Celui-ci a été classé comme cancérogène probable pour l'homme par le CIRC (groupe 2A), sa capacité à induire un cancer du poumon étant reconnue [4].

Tableau 3 : Potentiel cancérigène des HAP (IARC 1987-2002)

HAP	**Classement IARC**
Naphtalène	n.e
Acénaphtène	n.e
Acénaphtylène	n.e
Fluorène	3
Phénanthrène	3
Anthracène	3
Fluoranthène	3
Pyrène	3
Benz(a)anthracène	2A
Chrysène	3
Benzo(b) fluoranthène	2B
Benzo(k) fluoranthène	2B
Benzo(a) pyrène	2A
Dibenz(a,h) anthracène	2A
Benzo(ghi)pérylène	3
Indéno(1,2,3-cd)pyrène	2B

2A : probablementcancérigène pour l'homme ;2B peut êtrecancérigène pour l'homme. 3 : inclassable quant a la cancérogénicité pour l'homme (possible mais insuffisamment étudiée).n.e : non étudie.

A ce jour, 8 composés de la famille des HAP sont classés cancérogènes de catégorie 2 par l'Union européenne : le benzo[*a*]pyrène, le benzo[*k*]fluoranthène, le benzo[*j*]fluoranthène, le benzo[*a*]anthracène, le benzo[*b*]fluoranthène (ou benzo[*e*]acéphenanthrylène), le benzo[*e*]pyrène, le chrysène et le dibenzo[*a,h*]anthracène. Ces HAP sont majoritairement retrouvés sous forme particulaire. A noter que le naphtalène, HAP majoritairement retrouvé sous forme gazeuse, est classé cancérogène de catégorie 3.

I.2.2- Métabolisme

Les HAP présentent un caractère lipophile qui leur permet d'être transférés au sein des réserves lipidiques des organismes et dans les membranes cellulaires (essentiellement constituées de phospholipides). La présence de telles molécules entraîne rapidement la réaction des systèmes biochimiques de détoxication dont le rôle est de rendre hydrosolubles ces composés dangereux, afin de faciliter leur excrétion par voie rénale, biliaire ou branchiale. [5]

Dans l'organisme, certains tissus cellulaires, en particulier les tissus pulmonaires, hépatiques et cutanés, contiennent donc des enzymes chargées de catalyser une série de réactions permettant de détoxiquer les composés nocifs présents. En ce qui concerne les HAP cancérigènes, les réactions de détoxication est catalysée par le CYP1A1, qui appartient à une famille d'enzymes appelées cytochromes P450. Le benzo(a)pyrène se liait préférentiellement au gène suppresseur humain p53 dans les cellules épithéliales bronchiques. Ce gène suppresseur est alors muté et s'exprime anormalement. Il perd alors son rôle protecteur contre la prolifération de cellules malignes. Il est important de noter que les propriétés cancérigènes des HAP dépendent essentiellement de la structure du composé considéré. Plusieurs facteurs favorisent le caractère cancérigène : le nombre de cycles (ce sont les HAP à 4 noyaux aromatiques ou plus qui sont généralement cancérigènes), l'arrangement stérique de la molécule (les molécules planes sont moins toxiques). De plus, pour être cancérigène, le HAP doit posséder une région «baie» et être dissymétriqueFigure 2.

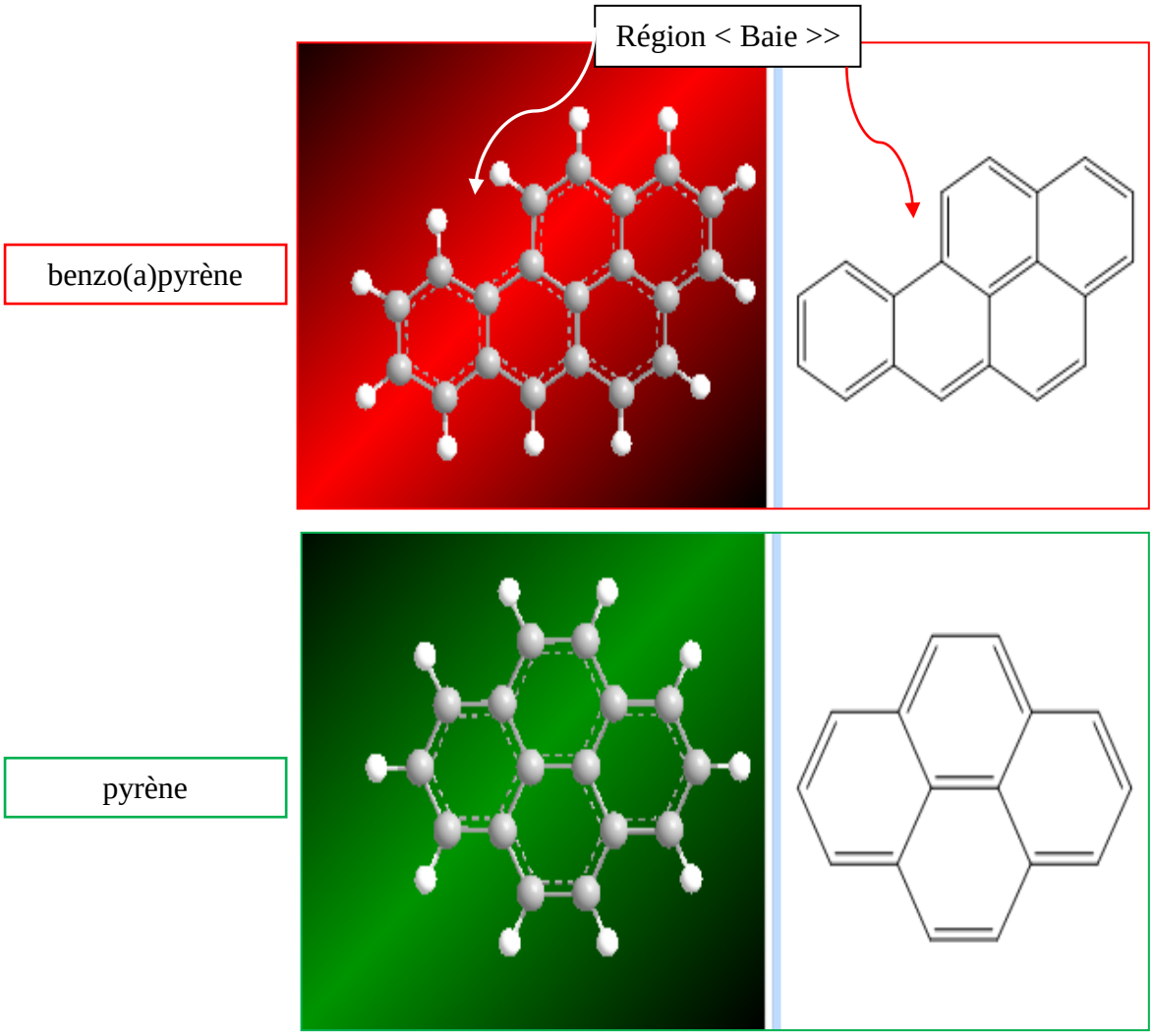


Figure2 : Structures chimiques du benzo(a)pyrène (cancérigène) et du pyrène (non cancérigène).

II- Chromatographie en phase gazeuse

II.1- Introduction

La chromatographie en phase gazeuse (CPG), comme toutes les techniques de chromatographie, permet de séparer les molécules d'un mélange éventuellement très complexe, de natures et de volatilités très diverses. Elle s'applique principalement aux composés gazeux ou susceptibles d'être vaporisés par chauffage sans décomposition.

Le mélange à analyser est vaporisé à l'entrée d'une colonne, qui renferme une substance active solide ou liquide appelée *phase stationnaire*, puis il est transporté à travers celle-ci à l'aide d'un *gaz porteur*. Les différentes molécules du mélange vont se séparer et sortir de la colonne les uns après les autres après un certain laps de temps qui est on fonction de l'affinité de la phase stationnaire pour ces molécules [6].

II.2- Appareillage

Les appareils de chromatographie gazeuse sont appelés **chromatographes**. Ils sont principalement composés [7] :

• d'un **four** (type chaleur tournante) qui permet une programmation de température ajustable de 20°C (-100°C pour certains systèmes) à 450°C et qui est également équipé d'un système de refroidissement rapide ;

• d'un **système d'injection**, qui va permettre d'introduire et de rendre volatil l'échantillon à analyser. L'injection peut se faire d'une manière manuelle ou automatique à l'aide d'un échantillonneur ;

• d'une **colonne** (*capillaire ou à garnissage)*, sur laquelle les différentes molécules de l'échantillon injecté vont se séparer suivant leurs affinités avec la phase stationnaire ;

• d'un **système de détection**, qui va permettre de mesurer le signal émis par les différentes molécules et de pouvoir les identifier. Pour l'enregistrement du signal émis par le détecteur, des logiciels sur PC remplacent avantageusement les enregistreurs analogiques sur papier ;

• d'un **système de détendeur-régulateur** pour les gaz utilisés (hélium, hydrogène, azote et air comprimé).

Sur les chromatographes modernes on trouve des systèmes électroniques pour la régulation des gaz qui sont également purifiés par des cartouches filtrantes. Pour faciliter l'identification, les chromatographes sont souvent couplés à d'autres instruments analytiques, notamment les spectromètres de masse et infra-rouge.

Le gaz porteur doit être chimiquement inerte, tels: He, N2, O2, CO2 . Son choix dépend du détecteur utilise.

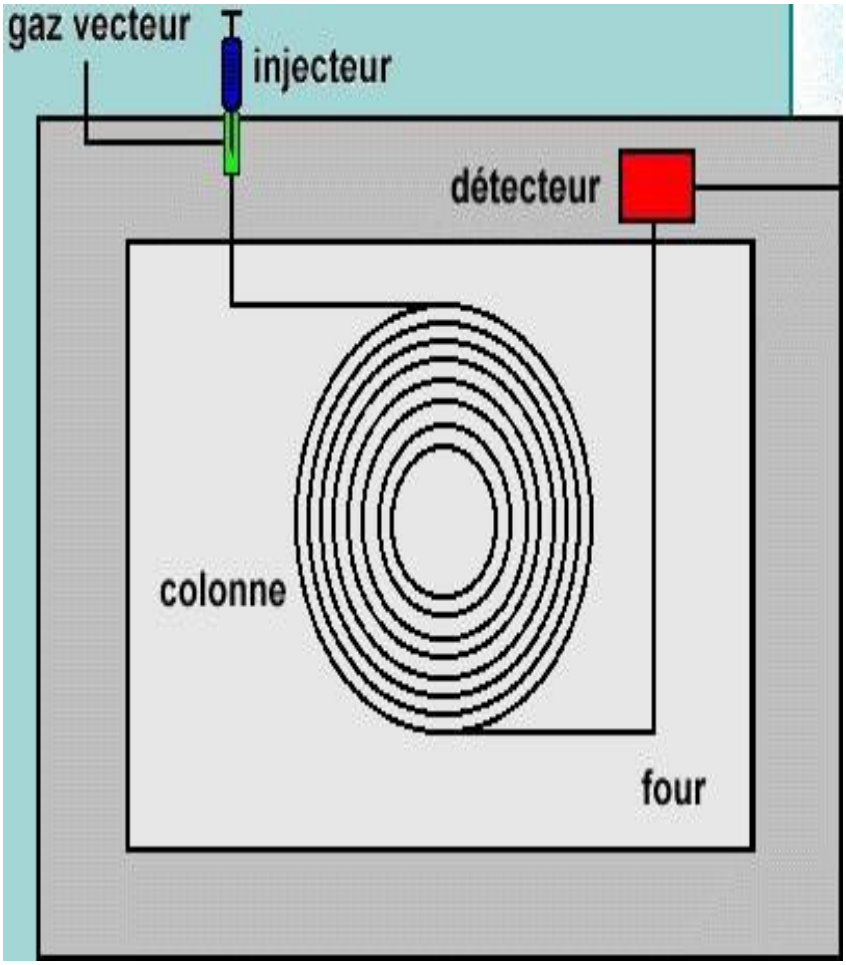


Figure 3 : Schéma d'un appareil de chromatographie gazeuse

Le couplage chromatographie gazeuse /spectrométrie de masse, s'il facilite souvent les questions d'identification peut être inefficace dans l'analyse des isomères ou des composés mineurs d'un mélange complexe .Les relations structure / rétention peuvent, dans ces conditions, aider à l'identification. Il s'agit de relier les réponses obtenues pour un ensemble d'évaluation à des propriétés physico-chimiques expérimentales ou théoriques, et/ou des descripteurs moléculaires de différents types fournis par divers logiciels spécialisés [08].

III - Les relations structure-propriété/activité quantitatives

Les relations quantitatives structure-activité/propriété (QSPR/QSAR) sont de plus en plus utilisées, du fait de la croissance des moyens de calculs. Très récemment, la mise en place du nouveau règlement européen REACH, qui recommande leur utilisation pour limiter le recours à l'expérience, donne un nouvel essor au développement de tels modèles prédictifs. Dans les dernières années, l'utilisation des méthodes QSPR n'a cessé de progresser. Elle est même devenue indispensable en chimie analytique et pharmaceutique. Leur développement dans une gamme plus large d'applications, leur ouvre d'ailleurs de grandes perspectives (ex : temps de rétention, températures critiques, pressions critiques, densité…etc.).

Il s'agit de présenter ici le principe des modèles QSPR ainsi que ceux des différents outils employés pour leur mise en place et leur évaluation : bases de données expérimentales, descripteurs, outils d'analyse de données **[**09**]**.

III.1- Définition de la méthode QSPR

Le QSPR (Quantitative Structure-Property Relationship) est le procédé par lequel des liens quantitatifs sont établis entre la structure moléculaire d'un ensemble de composés avec une propriété physicochimique. Les grandes phases de développement d'un modèle QSPR peuvent être décrites comme suit :

- ✓ Choisir des descripteurs adaptés au problème structure-propriété.
- ✓ Exploiter les valeurs des descripteurs comme variables, afin de définir une relation qui les corrèle à la propriété en question, à l'aide de machines d'apprentissage. C'est la fouille de données.
- ✓ Établir des critères de performance et de validation qui aideront au choix des meilleurs modèles pour le problème posé et estimer des incertitudes de prédiction [10].

III.2- Principe de la méthode QSPR

Le principe des méthodes QSPR est, comme leur nom l'indique, 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 physicochimique, etc.), pour une série de composés chimiques similaires à l'aide de méthodes d'analyses de données. Aussi la forme générale de modèle est :

Propriété = f(Descripteurs) **(1)**

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 multilinéaires (MLR), régressions aux moindres carrés partiels (PLS), arbres de décision, réseaux de neurones.

Une fois cette relation mise en place et validée sur un jeu validation, elle peut alors être employée pour la prédiction de la propriété de nouvelles molécules, pour lesquelles la valeur expérimentale n'est pas disponible, voire pour des molécules encore non synthétisées. De tels modèles peuvent également, dans certains cas, être utilisés pour mieux appréhender les phénomènes moléculaires mis en jeu dans la propriété d'intérêt [11].

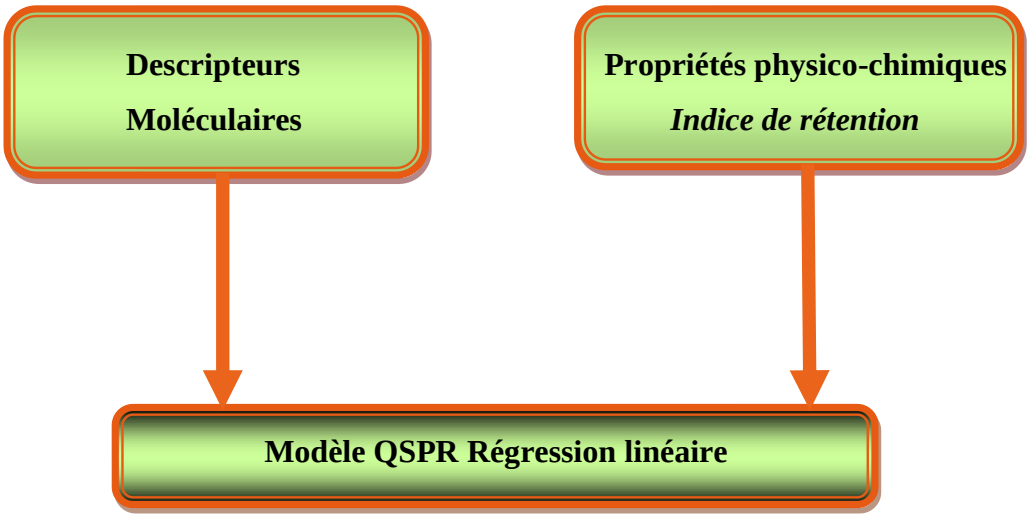


Figure 4 : Principe de la méthode QSPR

III.3- Les méthodes mathématiques utilisés par le model QSPR

- Régression linéaire
- Régression linéaire multiple (RLM)
- Régression en composantes principales(PCR)
- Régression moindre carré (PLS)

III.4- La relation structure-rétention quantitative (QSRR) :

Les relations structure – rétention chromatographique sont étudiées pour troisraisons principales : explication des mécanismes des séparations chromatographiques,prédiction des grandeurs de rétention, et caractérisation des propriétés physiques des solutés. Celles-ci sont particulièrement importantes pour estimer la réactivité et la bio-activité des contaminants et / ou des polluants, ce qui peut faciliter le suivi de leur évolution dans la nature.

La relation structure-rétention quantitative (QSRR) est une technique capable d'améliorer l'identification des analytes en prédisant leur temps de rétention sur une colonne de chromatographie gazeuse (CG) et / ou leurs propriétés. Cette approche est particulièrement utile lorsque(CG) est couplé à une plateforme de spectrométrie de masse à haute résolution (HRMS).

IV- COLLECTE DES DONNEES

La propriété considérée (Indice de rétention) est l'un des grandeurs physiquesimportant dans la chromatographie.Les données utilisées dans ce travail, concernent 50 HAP, [12] sont réunies dans le tableau (4) :

Tableau 4 :Nomenclature et valeurs de propriété des HAP étudiés.

N°	Composé	Ir
01	**Naphthalene**	1150
02	**Azulene**	1241
03	**Biphenyl**	1373
04	**Acetylnaphthalene**	1437
05	**Acenaphthene**	1458
06	**Fluorene**	1555
07	**2-Methylfluorene**	1673
08	**1-Methylfluorene**	1677
09	**trans-Stilbene**	1686
10	**Phenanthrene**	1741
11	**Anthracene**	1750
12	**2-Methylanthracene**	1870
13	**4,5-Dimethylenephenanthrene**	1875
14	**1-Methylphenanthrene**	1890
15	**9-Methylanthracene**	1920
16	**Fluoranthrene**	2020
17	**pyrene**	2070
18	**1,2-Benzofluorene**	2179

Tableau 4 : suite

N°	Composé	Ir
19	2,3-Benzofluorene	2195
20	3,4-Benzofluorene	2195
21	1-Methylpyrene	2215
22	3-Methylpyrene	2220
23	3,4-Benzophenanthrene	2332
24	1,2-Benzanthracene	2389
25	Chrysene	2395
26	Triphenylene	2395
27	Naphthacene	2425
28	Benzo[g,h,i]fluoranthene	2431
29	7-Methyl-1,2-Benzanthracene	2575
30	2,3-Benzofluoranthene	2700
31	4,5-Benzofluoranthene	2700
32	8,9-Benzofluoranthene	2706
33	7,12-Dimethyl-1,2-benzanthracene	2713
34	1,2-Benzopyrene	2760
35	Benzo(a)pyrene	2773
36	Perylene	2800
37	3-Methylcholanthrene	2906
38	12-Methylcholanthrene	2906
39	1,2,7,8-Dibenzanthracene	3078
40	1,2,3,4-Dibenzanthracene	3114
41	1,2,5,6-Dibenzanthracene	3114
42	Dibenzo[cd,jk]pyrene	3136
43	Picene	3150
44	dibenzo(a,L)pyrene	3423
45	dibenzo(a,i)pyrene	3477
46	dibenzo(a,e)pyrene	3507
47	Coronene	3544
48	1,2,4,5-Dibenzopyrene	3567
49	2,3,7,8-Dibenzopyrene	3600
50	2,3,6,7-Dibenzopyrene	3600

- On a choisi aléatoirement 10 HAP [pour la validation externe] ;

- Les 40 HAP [pour calibration qui sert à la construction du modèle].

V- OPTIMISATION DE LA GEOMETRIE MOLECULAIRE ET GENERATION DES DESCRIPTEURS MOLECULAIRES

V.1- Préparation de base des données

V.1.1- Logiciels « ChemDraw 3D pro 15»

ChemDraw 3D Pro [13] est le progiciel de dessin de structures chimiques le plusutilisé au monde grâce à ses fonctions d'édition avancées. Il permet d'interroger des bases de données chimiques et de dessiner la structure plane d'un composé conforme à lanomenclature de l'IUPAC.

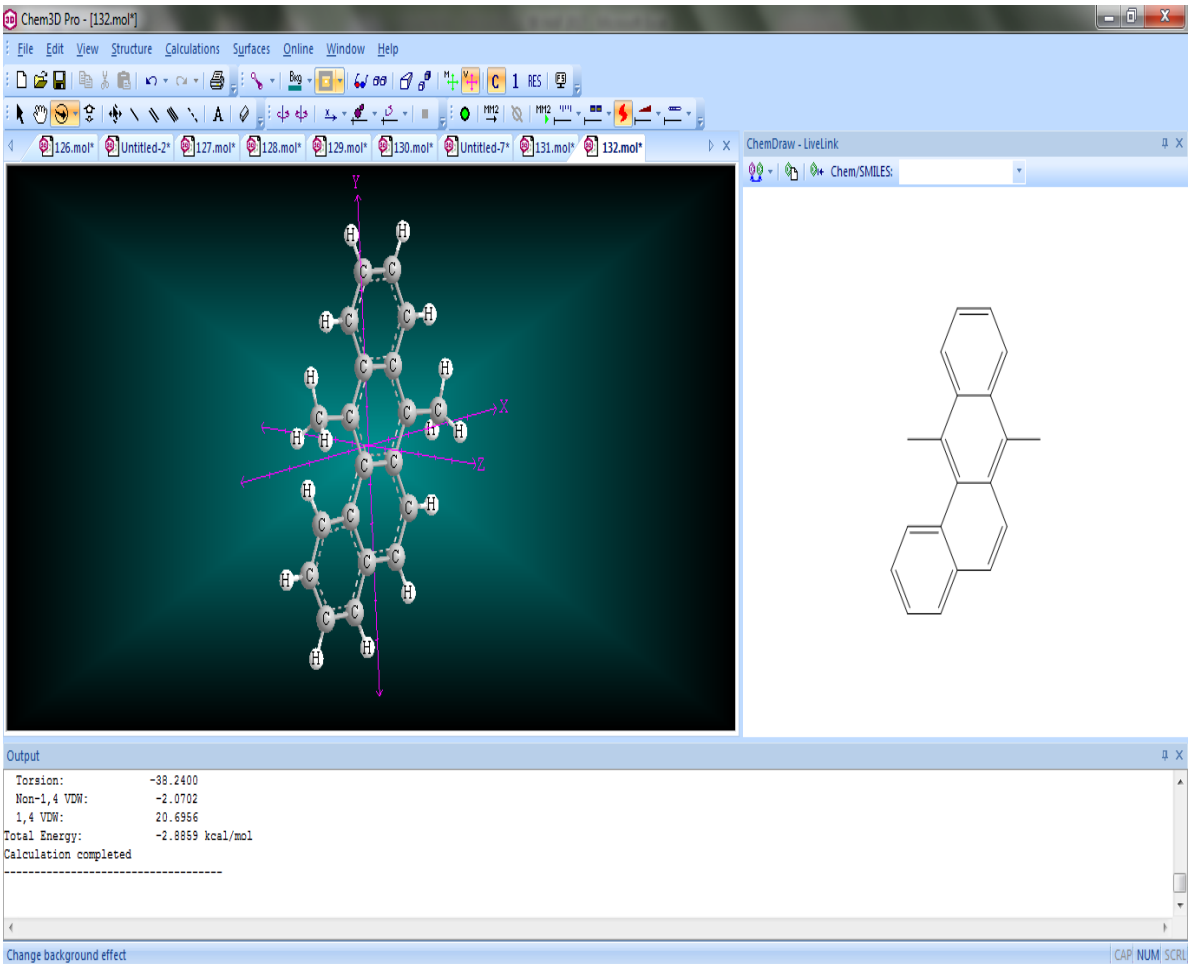


Figure 5 : Logiciel ChemDraw 3D pro

V.1.2- Logiciels « HyperChem 8 »

HyperChem [14] le dans une même interface un ensemble d'outils dédiés à la modélisation moléculaire, qui est connu pour sa qualité, flexibilité, et facilité d'usage.

❖ *Fonctionnalités :*

HyperChem est le logiciel qui vous permet de faire réellement de la modélisation : il possède plus de méthodes de calculs (mécaniquemoléculaire, semi-empirique et ab- initio) pour que vous puissiez calculer plus de propriétés.

HyperChem est utilisé dans cette étude pour construire et optimiser les molécules, chaque molécule est enregistrée comme un fichier nommé "Hin" après l'optimisation.

Nous avons utilisé la méthode semi empirique MM+ pour l'optimisation. On a 50 molécules donc on obtient 50 fichiers Hin, en suite on calcule les descripteurs moléculaires à partir de ces fichiers par le logiciel Dragon [15].

V.2- Récupération et stabilisation des molécules de fichier Hin

V.2.1- Stabilisation de la structure des molécules (minimisation de l'énergie)

Pour stabiliserla forme de structure de chaque molécule ou minimisation de l'énergie onutilise le HyperChem, qui pouvez effectuer une minimisation de l'énergie (ou de la géométried'optimisation) d'une molécule en utilisant une variété de méthodes de calcul. Les deuxmécanismes moléculaires et les méthodes semi-empiriques sont disponibles.Minimisation de l'énergie modifie la géométrie ou la forme d'une molécule d'abaisser l'énergie potentielle de la molécule et pour donner une conformation plus stable.

V.2.2- Mécanique Moléculaire

Champs de force mécanique moléculaire utiliser les équations de la mécaniqueclassique pour décrire les surfaces d'énergie potentielle et des propriétés physiques desmolécules. Une molécule est décrite comme une collection d'atomes qui interagissent lesuns avec les autres par de simples fonctions analytiques. Cette description est appelée unchamp de force. Une composante d'un champ de force est l'énergie résultant de lacompression et l'étirement d'un cautionnement [16]. HyperChem comprend quatre hautmécaniquemoléculaire des champs de force : de nouvelles implémentations detechniques élaborées et publiées par des groupes de recherche respectés Mais nous dansce travail on concerte à méthode MM+ de caractéristiques Field Force présente dans figure suivante :

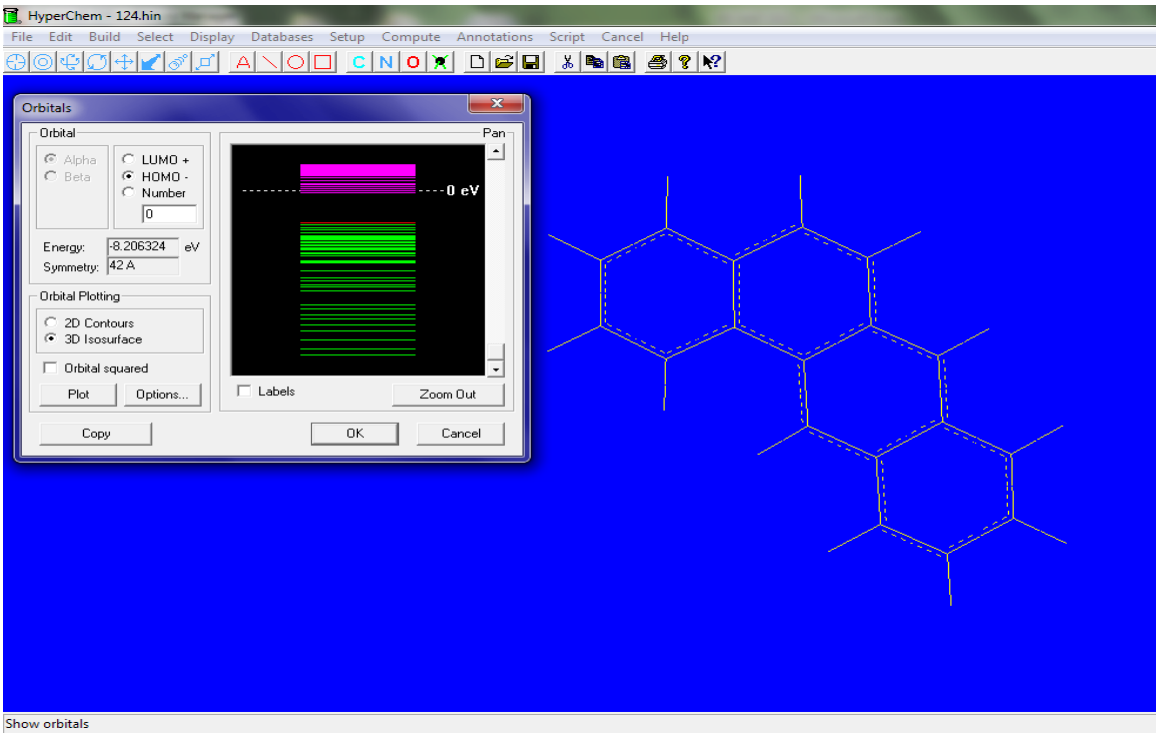


Figure 6 : Logiciel HyperChem

V.3- Récupération des fichiers HyperChem HIN

Une fois que vous avez construit une structure en HyperChem, vous pouvez l'enregistrer pour une utilisation ultérieure. C'est une bonne idée puisque cela vous fait gagner du temps si vous voulez revoir votre structure à une date ultérieure. Pourquoi construire deux fois ?! Vous pouvez le faire en allant dans Fichier et Enregistrer. Vous devez lui donner un nom comme Hin. Le fichier peut être rappelé pour la visualisation et la manipulation à une date ultérieure. Enregistrez-le dans le dossier public dans le répertoire approprié. Dans le calcul de descripteurs moléculaires, les structures chimiques des composés optimisés sont nécessaires. Les structures chimiques de tous les 50 composés dans notre jeu de données, ont été établies dans le logiciel HyperChem, et pré-optimisés en utilisant le champ avant MM+ mécanique.

V.4- Calcul des 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. Dans la prochaine étape pour tous les 50composés, les descripteurs moléculaires ont été calculés par le logicieldragon qui peut calculer plus de 1600 descripteurs moléculaires pour chaquestructure dans notre jeu des données.

V.4.1- Le Logiciel « Dragon »

Dragon est une application pour le calcul des descripteurs moléculaires. Cesdescripteurs peuvent être utilisés pour évaluer l'influence de la structure moléculaire oules relations propriétés-structure, aussi pour l'analyse de symétrie et la projection desbases de données des molécules Figure 7.

V.4.2- Descripteurs moléculaires

V.4.2-1- Définition d'un descripteur

Le descripteur moléculaire est le résultat final d'une procédure logique et mathématique qui transforme l'information chimique chiffrée dans une représentation symbolique d'une molécule dans un nombre utile ou le résultat de quelque expérience standard.

Les descripteurs moléculaires sont les traits communs les plus considérables de structure moléculaire qui peut être utilisée pour développer la « Relation Structure - Propriété ». Dans notre cas, la propriété est l'indice de rétention IR, des descripteurs moléculaires ont été proposés dérivé de théories différentes et approches avec le but de prédire biologique et propriétés physico-chimiques des molécules [17].

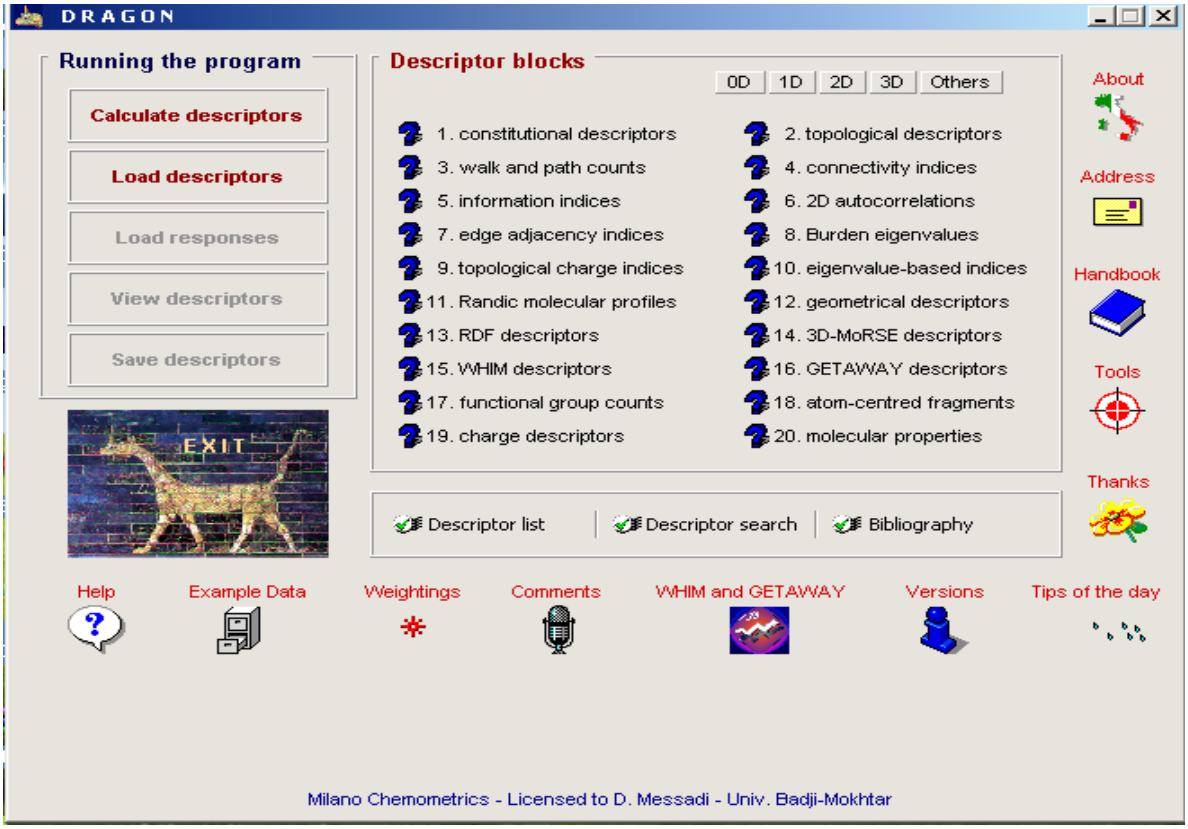


Figure 7 : Logiciel « Dragon »

V.4.2.2- Groupe des descripteurs moléculaires

En effet, les descripteurs moléculaires sont basés sur plusieurs théories différentes,tel que quantum chimie, théorie de l'information, chimie organique, théorie du graphique,et ainsi de suite, et est utilisé pour modeler des plusieurs propriétés différentes de produitschimiques dans les champs du scientifique tel que toxicologie, chimie analytique, chimiephysique…etc. [18].

Actuellement, il est possible de calculer plus de 1600 descripteurs moléculaires quisont représentés dans le tableau suivant qui peuvent être classés en 20 classes (blocs) logiques.

Tableau 5 : Quelques blocks des descripteurs calculés par le logiciel dragon

Classe	**Sous- classe**
Descripteurs géométriques	- Descripteurs liés à l'aire de la surface. - Descripteurs du champ stérique moléculaire. - Descripteurs liés au volume. - Descripteurs liés à la distance.
Descripteurs liés aux orbitales Moléculaires	- Energie des OM frontières - Ordres de liaison - Indices de réactivité de Fukui.
Descripteurs topologiques	- Indices topologiques (connectivité). - Descripteurs théoriques d'information. - Descripteurs topo-chimiques.
Descripteurs liés à la distribution de charge	- Charges atomiques partielles. - Moments électriques moléculaires - Polarisabilités moléculaires.
Descripteurs constitutionnels	- Dénombrement des atomes ou des liaisons. - Descripteurs basés sur les masses atomiques. - Descripteurs électro-topologiques.
Descripteurs température Dépendants	- Descripteurs facteurs de Boltzmann pondérés. - Fonctions thermodynamiques.
Descripteurs mixtes	- Descripteurs topographiques. - Descripteurs électro-topologiques. - Descripteurs de la charge partielle de l'aire de la surface

Suite tableau 5 :

Classe	Sous- classe
Descripteurs de solvatation	- Energie de dispersion de solvatation. - Energie électrostatique de solvatation. - Entropie de solvatation. - Descripteurs d'énergie de solvatation linéaire - Descripteurs de liaison hydrogène. - Enthalpie libre de formation de cavité.

V.4.2.3- Importance des descripteurs

Les descripteurs moléculaires jouent un rôle fondamental en chimie, sciences pharmaceutiques, la protection de l'environnement, recherche de la santé et contrôle dequalité, ils peuvent être obtenus quand les molécules sont transformées dans une représentationmoléculaire qui autorise quelque traitement mathématique. Les descripteurs moléculairessont très importants pour :

- ❖ Indiquons la description de la configuration de la molécule à étudie.
- ❖ Décrivons tous les paramètres descriptifs de la molécule.

Les descripteurs moléculaires sont utilisés pour, une connaissance de statistiques, chimio-métriques, et les principes des approches QSAR/QSPR sont nécessaires en plus de la connaissance spécifique du problème [19].

V.4.3- Diagramme montre les étapes de prédiction

Les premiers essais de modélisation QSPR développé par Wiener, depuis, l'essor de nouvelles techniques demodélisation par apprentissagelinéaires d'abordpuis nonlinéairesont permis la miseen place de nombreuses méthodes ; elles reposent pour la plupart sur la recherche d'unerelation entre un ensemble de nombres réels, descripteurs de la molécule, et la propriétéou l'activité que l'on souhaite prédire [20].

La prédiction des propriétés s'effectué par plusieurs étapes intéressantes illustrantdans le diagramme suivant :

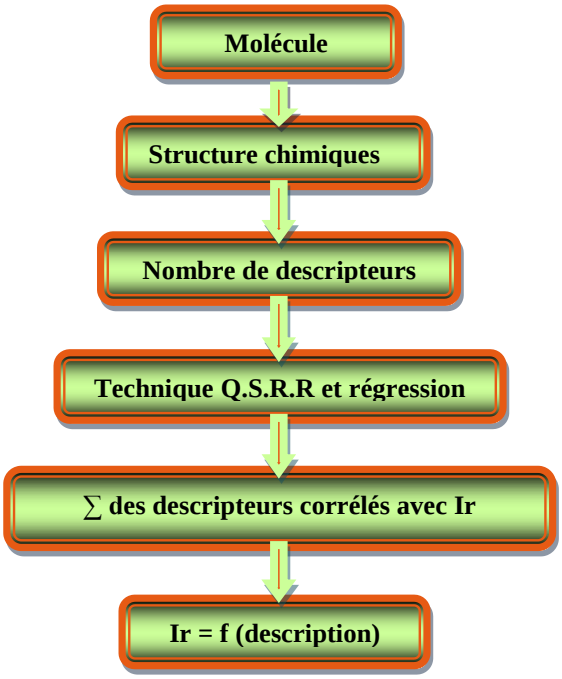


Figure 8 : Diagramme des étapes de prédiction

V.4.4- L'objectif de la prédiction

L'objectif principal est d'établir un modèle de prédiction de propriété en sebasant sur la structure moléculaire et plus précisément les descripteursmoléculaires. Ce modèle permet, en suite la classification des composées subi àla prédiction.

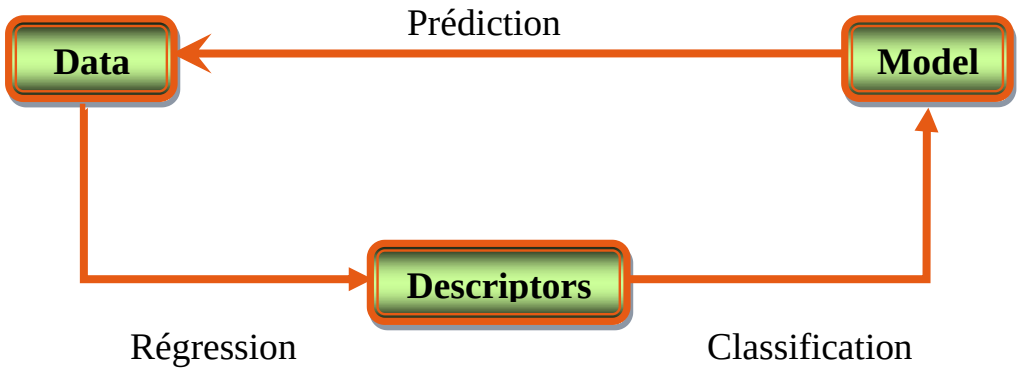


Figure 9 : Le cycle de prédiction

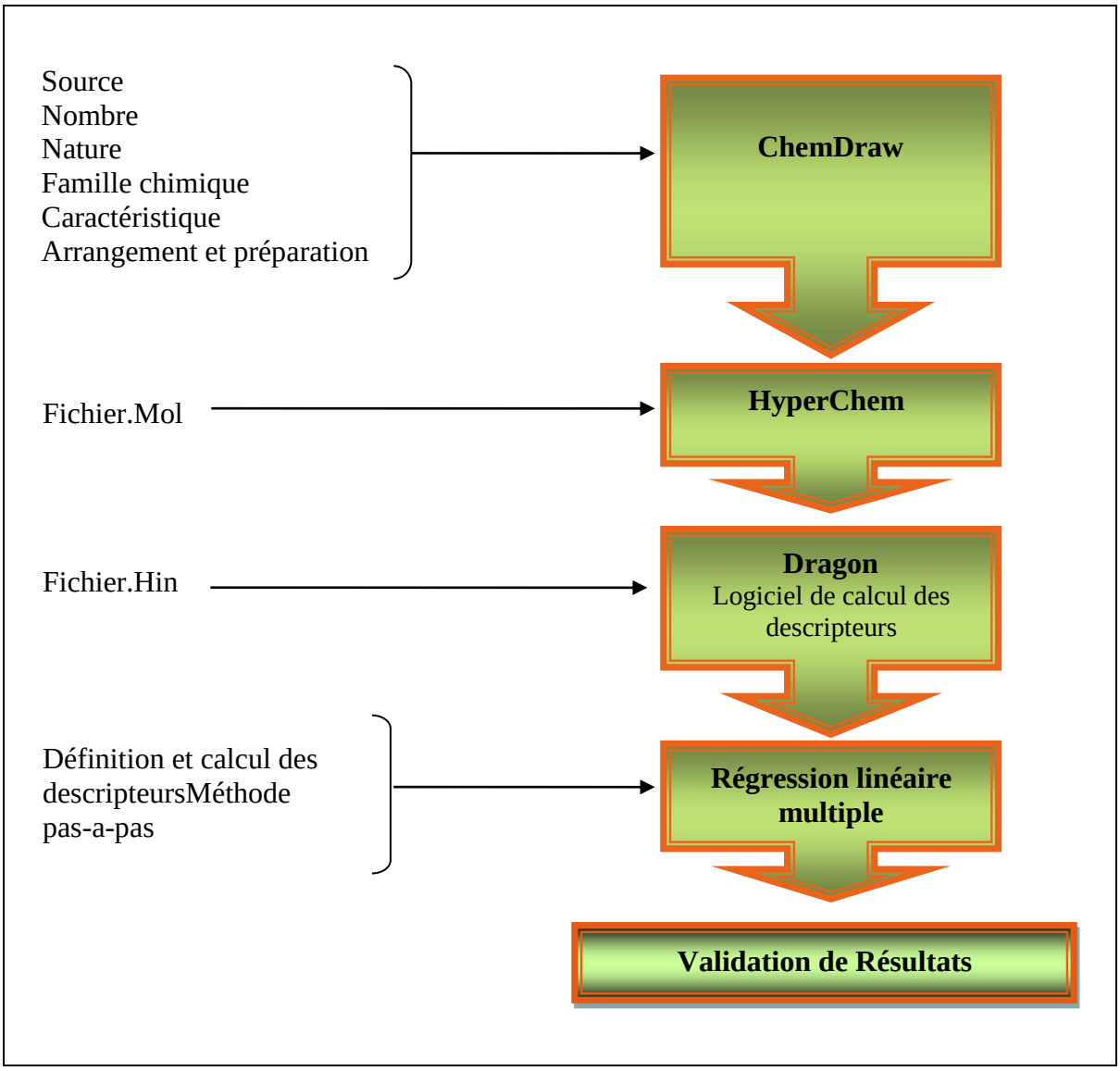


Figure 10 : Diagramme de notre travail

Le diagramme précédant illustré les étapes de travail et les techniques qu'ona utilisé dans la procédure de prédiction de la propriété étudiée (Ir).

VI- METHODES UTILISEES POUR LE DEVELOPPEMENT DE MODELES QSAR/QSPR

VI.1- Introduction

L'application pratique des gammes des descripteurs moléculaires dans le développement de modèles QSAR/QSPR n'est pas une tâche facile [21], tout d'abord, un très grand nombre (>1600) de descripteurs moléculaires, de différentes complexités et de conceptions diverses ont été imaginés et proposés au cours des (60 dernières) années. Ensuite, pendant ce temps, aucune règle stricte n'a été établie, ni même proposée, pour la sélection de descripteurs adaptés parmi la myriade de descripteurs disponibles. Ce choix a souvent été basé sur l'intuition chimique des chercheurs, ou en se pliant à la tradition. Une autre difficulté dans la sélection des descripteurs QSAR/QSPR découle du non-standardisation des gammes de descripteurs. Les gammes empiriques des constantes d'induction, de résonance et d'effet stérique des constituants, ou les échelles empiriques d'effets de solvant comportent des erreurs intrinsèques liées aux erreurs respectives des mesures expérimentales. Par ailleurs, les méthodes quanto- mécaniques appliquées aux calculs des descripteurs moléculaires et aux distributions de charges liés aux OM sont souvent basées sur différents paramètres semi- empiriques, ou l'utilisation de différents ensembles de base dans les calculs ab- initio.

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

Dans ce qui suit nous passerons en revue diverses approches utilisées pour le développement des " meilleures" équations QSPR dans de grands espaces de descripteurs. En dernier ressort, les modèles QSAR/QSPR peuvent être développés selon des modèles mathématiques différents, généralement en relation avec l'analyse statistique multi-variée.

Le premier modèle, et le plus largement utilisé, consiste en une équation (multi) linéaire obtenue par régression des données expérimentales en fonction d'un ensemble de descripteurs pré- sélectionnés (ou d'un seul), en utilisant la méthode des moindres carrés ordinaires (MCO). Dans quelques cas, les modèles physiques ou chimiques connus du phénomène étudié laissent prévoir certaines formes mathématiques non linéaires(exponentielles ou logarithmiques) de la dépendance entre les données

expérimentales et les descripteurs moléculaires. Les modèles QSAR/QSPR peuvent alors être établis à l'aide de la technique de régression par les moindres carrés non linéaires. D'autres modèles ont été développés en utilisant l'analyse factorielle ou l'analyse en composantes principales.

L'intérêt de ces méthodes est qu'elles évacuent le problème de multi-colinéarité inhérent aux méthodes de régression linéaires, cependant, l'interprétation des équations QSAR/QSRR est alors entravée par la nature formelle des facteurs ou des composantes principaux. Une alternative aux méthodes très classiques de régression linéaire multiple(MLR) et d'analyse en composantes principales (ACP) est la technique de régression parles moindres carrés partiels (MCP ou PLS) [22] [26].

VI.2- Méthodes de régressions linéaires et multilinéaires

VI.2-1 Aperçu général

Comme signalé auparavant, l'investigateur choisit dans chaque cas un ou plusieurs descripteurs supposé(s) refléter les interactions physiques ou chimiques à la base de la propriété moléculaire ou de la caractéristique du phénomène étudié. Ce choix, encore une fois, est habituellement fondé sur l'intuition chimique, la tradition, ou simplement la disponibilité du descripteur. Néanmoins, cinq principes peuvent aider à la sélection de descripteurs moléculaires convenables pour l'établissement de modèles QSAR/QSPR. Ce sont :

a) Un nombre maximal de données expérimentales (de préférence toutes) doivent être caractérisées par des valeurs de descripteurs originaux complémentaires.

b) Les valeurs des descripteurs doivent être obtenues de la même source et, de préférence, mesurées selon le même protocole expérimental ou calculées en utilisant le même logiciel.

c) Le nombre de descripteurs dans les modèles de régression multiples doit être minimisé, sans perte d'information, ce que mettent en évidence les critères statistiques (valeurs des tests t et F…).

d) Dans les modèles MLR, les descripteurs utilisés doivent être statistiquement orthogonaux.

e) Pourvu que les autres critères soient similaires, la nature physique ou chimique du descripteur sélectionné doit être la plus proche de la propriété ou du phénomène étudié.

En réalité, il est difficile de se conformer pratiquement aux 5 principes énoncés.

Cependant, la négligence de plusieurs d'entre eux peut conduire à des équations inutiles sans aucun pouvoir prédictif sinon très limité.

VI.2.2- Evaluation préliminaire des données

Avant d'entamer le développement effectif des équations de régression 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).

On distingue habituellement dans un tel pré- traitement des données les analyses uni-variées des analyses bi-variées [27,32].

Dans l'analyse uni-variée, il est recommandé de vérifier la conformité des données à la distribution normale. Une précaution particulière doit être prise lors de la procédure de régression subséquente si les valeurs de la propriété étudiée, ou d'un descripteur, ne suivent pas la loi de Laplace-Gauss.

Pour un ensemble de descripteurs différents, il est nécessaire d'effectuer une analyse des données bi-variée, c'est-à-dire de calculer le coefficient de corrélation linéaire R entre chacune des paires de l'ensemble des descripteurs. Si R est statistiquement significatif(R>0,9), ces deux descripteurs ne peuvent être utilisés simultanément lors de l'analyse par MLR.

VI.2.3- Régression linéaire multiple

Un modèle de régression linéaire multiple entre une variable expliquée Y et p variables explicatives $X_1,\ldots,X_P$, s'écrit pour tout i=1,…, n :

$$\boldsymbol{y_i = \beta_0 \sum_{j=1}^{p} \beta_j X_{ij} + \xi} \qquad \textbf{(02)}$$

Ou les y_i, xi1, xi2…,x_{ip} : sont des données respectivement relatives aux variables Y, X_1..., X_p.

Les estimateurs βj sont calculés en utilisant la méthode des moindres carrés ordinaires.

Les variables aléatoires ξi représentent les termes d'erreur non observables du modèle.

On peut estimer ces erreurs par les résidus ordinaires ei, différence entre les valeurs observées yi et les valeurs estimées ŷi.

Pour construire le modèle et admettre que les coefficients de la régression sont sans biais et convergents, on montre qu'il faut poser comme hypothèses :

a) Les résidus (E) ont une espérance mathématique nulle :

$$E(e) = 0 \qquad (03)$$

b) Le modèle choisi est correct (aucune variable explicative n'a été omise).

c) Les résidus sont indépendants entre eux :

$$E(ei, ej) = 0 \quad \text{si } i \neq j \qquad (04)$$

Leurs covariances sont nulles.

d) Les résidus ont tous même variance σ^2(propriété d'homoscédasticité).

Par ailleurs, l'emploi de tests statistiques pour analyser la variation expliquée par la régression conduit à admettre que :

Les résidus suivent une distribution normale (de Laplace-Gauss).L'analyse des résidus présente un intérêt à plusieurs égards. Elle permet en effet de vérifier, a posteriori, la validité du modèle utilisé, en ce qui concerne, d'une part la forme de celui- ci (linéarité ou non linéarité de la relation, par exemple) et d'autre part, certaines hypothèses plus spécifiques, telles que l'égalité des variances résiduelles, la normalité des résidus ou l'absence d'auto- corrélation.

Pour minimiser l'influence des erreurs de détermination des valeurs explicatives (ou régresseurs) sur la précision des résultats de la régression 5 données (variables dépendantes, ou encore observations) doivent, à la limite, être associées à chaque variable explicative. Le nombre de degrés de liberté final (n-p-1) doit être tel que [33] :

$$n - p - 1 \geq 10 \qquad (05)$$

n, étant la dimension de l'échantillon, et p le nombre de variables explicatives entrant dans la construction du modèle.

VI.3- Méthodes de sélection des descripteurs

- **Régression pas à pas**

Une technique est parfois employée pour sélectionner un nombre réduit de variables qui explique pourtant une quantité raisonnable de variation. Il existe plusieurs variantes de cette régression dite "pas à pas" (stepwise regression) en anglais.

A - Méthode rétrograde (backward selection)

Cette méthode consiste à construire un modèle de régression complet (intégrant toutes les variables explicatives), et à en retirer une par une les variables dont le t partiel est non-significatif (en commençant parcelle qui explique le moins de variation).

Inconvénient : une fois qu'une variable a été retirée, elle ne peut plus être réintroduite dans le modèle, même si, à la suite du retrait d'autres variables, elle redevenait significative. Cette approche est néanmoins assez libérale (elle a tendance à garder un nombre plus élevé de variables dans le modèle final que les autres approches ci-dessous).

B - Méthode progressive (forward selection)

Approche inverse de la précédente : elle sélectionne d'abord la variable explicative la plus corrélée à la variable dépendante. Ensuite, elle sélectionne, parmi celles qui restent, la variable explicative dont la corrélation partielle est la plus élevée (en gardant constantes là où les variables déjà retenues), et ainsi de suite tant qu'il reste des variables candidates dont le coefficient de corrélation partiel est significatif. Inconvénient :lorsqu'une variable est entrée dans le modèle, aucune procédure ne contrôle si sa corrélation partielle reste significative après l'ajout d'une ou de plusieurs autres variables. Cette technique est en général plus conservatrice que la précédente, ayant tendance à sélectionner un modèle plus restreint (moins de variables explicatives) que la sélection rétrograde.

Des simulations récentes [33] montrent que même la sélection progressive, la plus conservatrice des trois variantes, est trop libérale, c'est-à-dire qu'elle laisse souvent entrer au moins une variable non significative dans le modèle. C'est la raison pour laquelle nous proposons désormais d'appliquer un double critère d'arrêt à la sélection pas à pas (plus spécifiquement à la sélection progressive) :

* Le niveau a habituel.

* Le R^2_{aj} du modèle comprenant toutes les variables candidates.

Pour ce deuxième critère, on calcule tout d'abord le R^2aj global d'une régression multiple comprenant toutes les variables explicatives candidates. Ensuite, durant la

procédure de sélection, on arrête la sélection lorsque le niveau a présélectionné ou le R^2ajglobal est atteint.

Cette procédure garantit une erreur de type I correcte et réduit fortement le nombre de variables explicatives introduites à tort dans le modèle. C'est la méthode de sélection que nous préconisons. Dans ce travail les sélections des descripteurs ont été réalisées par pas-à-pas en utilisant le logiciel de calcul statistique MINITAB version 16.2.0 [34]

VI.4- Paramètres d'évaluation de la qualité de l'ajustement

Deux paramètres sont couramment utilisés :

- **Le coefficient de détermination multiple R^2**

Afin de se faire une idée sur la qualité de l'ajustement ainsi réalisé nous avons calculé le coefficient de détermination, R^2, qui exprime la part de la variation dépendante Y=indice de rétention) ″expliquée ″ ou ″justifiée″ par la régression. Ce paramètre correspond au carré du coefficient de corrélation, il est compris entre 0 et1 et s'exprime toujours en pour cent.

Si la valeur de R^2 est proche de 1 ou 100% ; nous avons donc un ajustement d'excellente qualité ; par contre si la valeur de R^2 est faible et proche de 0 ou 0% l'ajustement est mauvais.

$$R^2 = 1 - \frac{\mathrm{SCE}}{\mathrm{SCT}} = 1 - \frac{\sum_1^n (y_i - \hat{y}_i)^2}{\sum_1^n (y_i - \bar{y}_i)^2} \qquad (6)$$

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

- **La racine de l'erreur quadratique moyenne de prédiction (**désignée également par SDEP) :

$$\mathrm{SDEP} = \sqrt{\frac{1}{n}\sum_1^n (y_i - \hat{y}_{(i)})^2} = \sqrt{\frac{\mathrm{PRESS}}{n}} \qquad (07)$$

VI.5- Facteur d'inflation de la variance [FIV]

Le facteur d'inflation de la variance sert à détecter si descripteur présente une association linéaire forte avec les prédicteurs restants (présence de multi colinéarité parmi les prédicateurs). Le facteur d'inflation de la variance donne une mesure de l'accroissement de la variance d'un coefficient de régression estimé s'il existe une corrélation entre prédicteurs (multi colinéarité).FIV =1 indique qu'il n'y a pas de relations, si non FIV est supérieur à 1 le facteur FIV le plus grand parmi tous les prédicteurs sert souvent d'indicateur de multi colinéarité importance, Si le FIV > 5-10 la qualité de l'estimation des coefficients de régression est faible.

VI.6- Test de randomisation

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

VI.7- Validation externe

Il est intéressant, pour juger de la qualité du modèle, de considérer la racine de l'écart quadratique moyen (RMSE, pour Root Mean Squared Error), calculée sur différents ensembles :

- Ensemble d'estimation (appelée SDEC)
- Ensemble de validation croisée (appelée également SDEP)
- Ensemble de prédiction externe (désignée par SDEPext).

Ces valeurs RMSE sont mieux adaptées, pour juger de la qualité d'un modèle que les valeurs de R2 et Q2 seules, qui constituent de bons tests uniquement pour des données réparties régulièrement.

$$SDEC = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}} \qquad (08)$$

$$SDEP_{ext} = \sqrt{\frac{\sum_{i=1}^{n_{ext}}(y_i - \hat{y}_i)^2}{n_{ext}}} \qquad (09)$$

La validation croisée par « leave – one - out » (LOO) [35] consiste à recalculer le modèle sur (n-1) observations, et à utiliser le modèle ainsi obtenu pour calculer la grandeur d'intérêt du composé écarté, notée ŷ(i). On répète le procédé pour chacune des grandeurs d'intérêt. La somme des carrés des erreurs de prédiction, désignée par le symbole PRESS (équation. (7)), est une mesure de la dispersion des estimations. On l'utilise pour définir le coefficient de prédiction :

$$Q^2_{LOO} = \frac{SCT - PRESS}{SCT} \quad (10)$$

Contrairement à R^2 au coefficient qui augmente avec le nombre de paramètres du modèle, le facteur Q^2_{LOO} affiche une courbe avec maximum (ou avec palier) obtenu pour un certain nombre de descripteurs, puis décroît de façon monotone. Ce fait confère une grande importance au coefficient Q^2_{LOO} .Une valeur $Q^2_{LOO} > 0.5$ est considérée comme satisfaisante, et une valeur supérieure à 0,9 est excellente [36]. Si de petites valeurs de Q^2_{LOO} indiquent des modèles peu robustes, caractérisés par de faibles capacités prédictives internes, le contraire n'est pas nécessairement vrai. En fait, si une forte valeur de Q^2_{LOO} est une condition nécessaire de robustesse et d'une possible capacité prédictive élevée d'un modèle, cette condition seule n'est pas suffisante, et peut conduire à une surestimation de la capacité prédictive du modèle lorsqu'il est appliqué à des composés réellement externes.

Evidemment, on peut être amené à écarter 2, 3 ou un plus grand nombre d'éléments à la fois, ce qui conduit aux procédures LMO (leave – many- out). Cependant, ces procédures ne sont que rarement rapportées avec les résultats QSPR courants, et n'ont pas été pleinement exploitées dans le présent travail.

Dans le cas où on a suffisamment de données qui n'ont pas servi dans la création du modèle ou après collecte de nouvelles, on peut ou on doit procéder à la validation de ce dernier, c'est la validation externe. La statistique se rapportant à ce procédé, notéeQ^2_{ext}, est calculée comme suit :

$$Q^2_{ext} = 1 - \frac{\sum_{i=1}^{n_{ext}} (y_i - \hat{y}_i)^2 / n_{ext}}{\sum_{i=1}^{n} (y_i - \bar{y}\ \ )^2 / n} \quad (11)$$

Pour une grande valeur de Q^2_{LOO}, une valeur élevée de Q^2_{ext} permet de présager d'une bonne capacité prédictive du modèle. Dans notre cas : $\boldsymbol{n}$=40 et $\boldsymbol{n}_{ext}$ =10.

PARTIE APPLIQUATION

Modélisation de l'indice de rétention Ir

1- Sélection des descripteurs

Plusieurs classes de descripteurs moléculaires ont été calculées pour la molécule entière : Descripteurs relatifs à la construction moléculaire, et descripteurs topologiques. Nous rapporterons quelques résultats obtenus pour **50**HAP dans le tableau 06**.**

Tableau 06 : Valeurs des descripteurs moléculaires sélectionnés

N°	*Composé*	Ir	VRZ1	RDF055m	C-024
1	**Naphthalene**	1150	36.189	0.028	8
2	**Azulene**	1241	36.053	0.215	8
3	**Biphenyl***	1373	47.226	0.078	10
4	**Acetylnaphthalene**	1437	54.725	2.220	7
5	**Acenaphthene**	1458	50.776	0.667	6
6	**Fluorene**	1555	56.720	0.598	8
7	**2-Methylfluorene***	1673	63.797	0.715	7
8	**1-Methylfluorene**	1677	63.776	0.671	7
9	**trans-Stilbene**	1686	58.472	1.452	10
10	**Phenanthrene**	1741	63.140	0.159	10
11	**Anthracene**	1750	62.837	0.086	10
12	**2-Methylanthracene***	1870	70.182	0.097	9
13	**4,5-Dimethylenephenanthrene**	1875	76.994	0.877	2
14	**1-Methylphenanthrene**	1890	70.510	0.146	9
15	**9-Methylanthracene**	1920	70.319	0.134	9
16	**Fluoranthrene**	2020	80.479	2.566	10
17	**Pyrene**	2070	79.740	0.271	10
18	**1,2-Benzofluorene***	2179	87.045	0.862	10
19	**2,3-Benzofluorene**	2195	86.816	0.842	10
20	**3,4-Benzofluorene**	2195	87.166	1.922	10
21	**1-Methylpyrene**	2215	87.873	0.252	9
22	**3-Methylpyrene**	2220	87.873	0.252	9
23	**3,4-Benzophenanthrene**	2332	94.552	4.910	12
24	**1,2-Benzanthracene**	2389	94.137	0.245	12
25	**Chrysene**	2395	94.400	0.182	12
26	**Triphenylene**	2395	94.958	0.286	12
27	**Naphthacene**	2425	93.786	0.194	12
28	**Benzo[g,h,i]fluoranthene***	2431	98.100	1.625	10
29	**7-Methyl-1,2-Benzanthracene***	2575	102.43	0.202	11
30	**2,3-Benzofluoranthene**	2700	114.07	3.028	12
31	**4,5-Benzofluoranthene**	2700	114.07	3.028	12
32	**8,9-Benzofluoranthene**	2706	113.68	2.748	12
33	**7,12-Dimethyl-1,2-benzanthracene**	2713	110.95	0.370	10
34	**1,2-Benzopyrene**	2760	113.43	0.373	12
35	**Benzo(a)pyrene**	2773	113.62	0.366	12
36	**Perylene**	2800	113.35	0.296	12
37	**3-Methylcholanthrene**	2906	122.72	1.448	9

Tableau06 : suite

N°	*Composé*	Ir	VRZ1	RDF055m	C-024
38	**12-Methylcholanthrene**	2906	122.64	2.044	9
39	**1,2,7,8-Dibenzanthracene***	3078	128.97	0.567	14
40	**1,2,3,4-Dibenzanthracene**	3114	129.97	0.298	14
41	**1,2,5,6-Dibenzanthracene**	3114	129.14	0.357	14
42	**Dibenzo[cd,jk]pyrene**	3136	129.40	0.708	14
43	**Picene**	3150	129.45	0.311	14
44	**dibenzo(a,L)pyrene**	3423	151.24	6.435	14
45	**dibenzo(a,i)pyrene**	3477	150.99	0.547	14
46	**dibenzo(a,e)pyrene***	3507	151.16	0.39	14
47	**Coronene**	3544	154.16	0.403	12
48	**1,2,4,5-Dibenzopyrene**	3567	151.16	0.39	14
49	**2,3,7,8-Dibenzopyrene***	3600	150.99	0.542	14
50	**2,3,6,7-Dibenzopyrene**	3600	150.99	0.548	14

(*) Composés de validation

Les modèles ont été calculés par les méthodes de sélection Pas-à-pas "stepwise" du logiciel Minitab [30].Le niveau de signification, a été fixé à 0.05 tant pour l'introduction que pour l'expulsion des variables.

Nous traiterons dans ce travail la propriété physique envisagée (Ir), en considérant les mêmes ensembles d'estimation et de validation. Rappelant que les **50** composées ont été éclatées aléatoirement en deux sous ensemble comportant **40** pour la calibration et **10** pour la validation.

Dans les études QSPR, cinq composés, à la limite, doivent être associées à chaque variable explicative.

Le nombre de degrés de liberté final doit être au moins égal à (10) soit, en désignant par k le nombre de descripteurs :

$$n - k - 1 \geq 10 \qquad (12)$$

Cette condition est bien vérifiée pour le modèle à 3 variables.

Une indépendance globale acceptable des descripteurs sera vérifiée lorsque les facteurs d'inflation de la variance (FIV) calculés pour chacun d'eux sont inférieurs à 5.

2- Evaluation préliminaire des données

Avant de procéder au développement effectif des équations de régression QSPR, nous avons vérifié la qualité statistique des données initiales (variable dépendante et descripteurs).

3 - Calcul des corrélations entre les différents descripteurs

Le coefficient de corrélation, r, de Bravais-Pearson a servi pour mettre en évidence les relations possibles entre les différents descripteurs des 40 composés, la matrice de corrélation obtenue à l'aide de la commande "corrélation" du logiciel MINITAB, montre que les descripteurs sont entre eux plus ou moins corrélés. Les couples des descripteurs qui présentent des valeurs de r > 0.90, sont très fortement corrélés et apportent la même information, ce qui fait qu'ils ne peuvent apparaître dans une même équation de régression.

Tableau 07 : Corrélations Ir ;Mv ;Me ;Ms ;ARR ;SPI ;TI2 ; Rww; ...

N°		Ir	Mv	Me	Ms	ARR	SPI	TI2	Rww	DELS	S3K
1	Mv	0.781									
2	Me	0.425	0.603								
3	Ms	0.386	0.371	0.307							
4	ARR	0.407	0.556	0.255	-0.265						
5	SPI	0.032	0.402	0.095	0.344	0.522					
6	TI2	0.05	-0.24	0.103	0.191	0.116	0.009				
7	Rww	0.402	0.624	0.274	0.564	0.476	0.446	0.562			
8	DELS	0.755	0.543	0.663	0.191	0.06	0.34	-0.039	-0.146		
9	S3K	0.392	0.032	0.057	0.143	0.048	0.072	0.867	0.494	0.263	
10	PJI2	0.109	0.237	0.071	0.205	0.072	0.252	0.49	0.194	0.219	0.449

4- Calcul des équations de régression

Après avoir éliminé les descripteurs qui n'obéissent pas à la condition précédente, nous avons utilisés la méthode de sélection de variables significatives suivante :

- **Méthode pas-à-pas (PP)**

Le logiciel Minitab assure le traitement des données par cette méthode, les résultats sont illustrés dans le tableau suivant :

Tableau 08 : Sélection des modèles par la méthode pas-à-pas (stepwise)

Stepwise Regression : Ir versus VRZ1;RDF055m;C-024...

Alpha-to-Enter: 0.05 Alpha-to-Remove: 0.05

Response is Ir on 190 predictors, with N = 40

Step	1	2	3	4	5	6
Constant	421.6	425.9	342.6	328.3	349.2	173.1
VRZ1	20.52	20.74	19.69	20.46	20.81	20.40
T-Value	78.70	96.29	87.29	69.69	72.59	65.24
P-Value	0.000	0.000	0.000	0.000	0.000	0.000
RDF055m		-23.8	-23.0	-27.6	-29.7	-30.9
T-Value		-4.63	-6.36	-8.10	-9.49	-10.49
P-Value		0.000	0.000	0.000	0.000	0.000
C-024			17.3	22.6	22.0	22.9
T-Value			6.23	7.94	8.59	9.52
P-Value			0.000	0.000	0.000	0.000
Mor14m				123	147	163
T-Value				3.53	4.56	5.34
P-Value				0.001	0.000	0.000
JGI6					-1694	-2312
T-Value					-3.08	-4.09
P-Value					0.004	0.000
EEig02r						63
T-Value						2.54
P-Value						0.016
S	53.9	43.5	30.6	26.6	23.9	22.2
R-Sq	**99.39**	**99.61**	**99.81**	**99.86**	**99.89**	**99.91**
R-Sq(adj)	99.37	99.59	99.80	99.85	99.88	99.89
PRESS	124519	83550.5	42122.7	32675.5	27449.4	24574.8
R-Sq(pred)	99.31	99.54	99.77	99.82	99.85	99.86

En traçant le graphe de R^2 en fonction de nombre de descripteurs ; on voit que le bon modèle qui possède le moins de descripteurs est celui à 3, parce qu'après le modèle à 3 les valeurs de R^2 augmentent d'un pas faible **figure 11.**

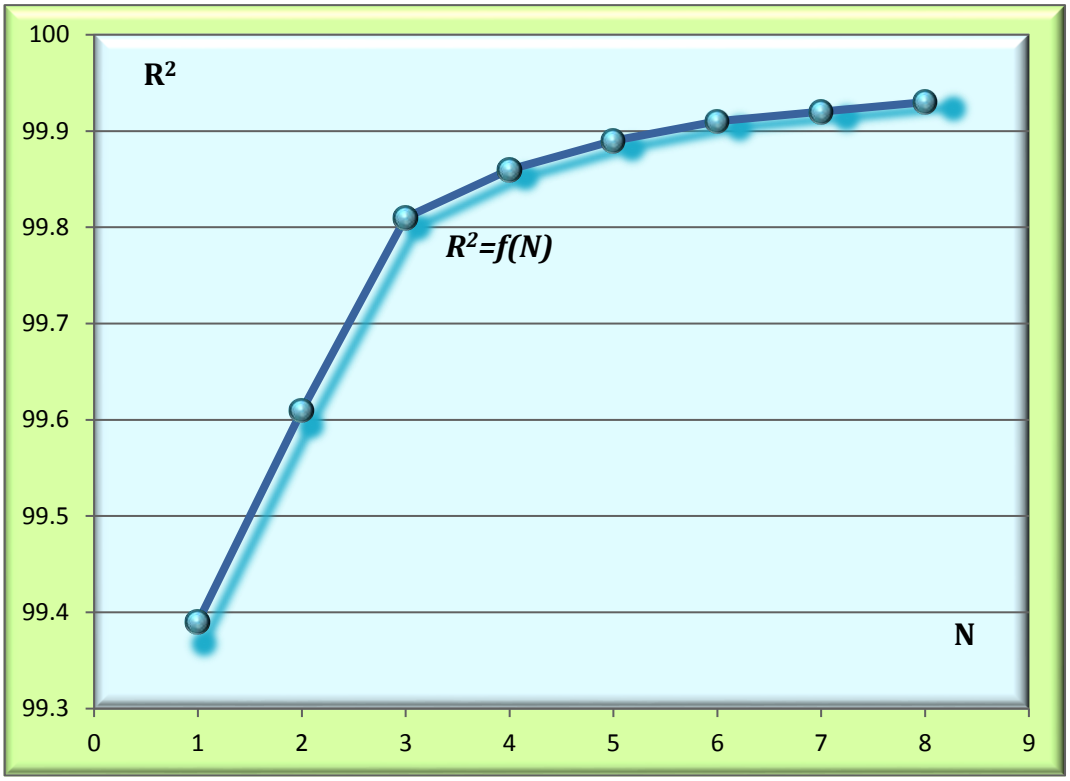


Figure 11 : Variation de R^2 en fonction du nombre de descripteur N

Parmi les modèles optimaux générés, celui qui fournit la valeur maximale pour les paramètres statistiques R^2, Q^2 tout en vérifiant la condition : FIV > 5comporte les 3 descripteurs calculés par le logiciel DRAGON, dont le symbole, la classe et la signification sont réunis dans **tableau 09.**

Tableau 09 : Classes et significations des descripteurs

Descripteur	Classe	Signification
VRZ1	Randic molecular profiles	molecular profile no. 01 and molecular profile no. 15 Molecular profiles are sequences of molecular descriptors proposed by Randic and derived from the interatomic geometric distances of a molecule.
RDF055m	Topological descriptors	The polarity number is calculated on the distance matrix as the number of pairs of vertices at a topological distance equal to three
C-024	Walk and pathcounts	Molecular path counts are descriptors obtained from a H-depleted molecular graph, based on the graph path, which is a walk without any repeated vertices or edges. The molecular path count MPCk of order k is the total number of paths of length k in the graph.

5- Matrice de corrélation :

Pour vérifier que les descripteurs ne sont pas fortement corrélés on a utilisé la matrice de corrélation.

Tableau 10 : Corrélation entre Ir et les descripteurs **VRZ1 ; RDF055m ; C-024**

Ir ; VRZ1 ; RDF055m ; C-024			
	Ir	**VRZ1**	**RDF055m**
VRZ1	0.996		
	0.000		
RDF055m	0.182	0.228	
	0.261	0.157	
C-024	0.762	0.732	0.136
	0.000	0.000	0.403

D'après ce tableau de corrélation on voit que les descripteurs ne sont pas fortement corrélés à l'exception entre VRZ1 et C-024 la corrélation est un peu remarquable.

6- Equation et paramètres de régressions :

L'équation et les paramètres de régression sont illustrés dans le tableau 11 :

Tableau 11 : Equation et paramètres de régression

Regression Analysis :**Ir** versus **VRZ1; RDF055m; C-024**

The regression equation is
Ir = 343 + 19.7 VRZ1 - 23.0 RDF055m + 17.3 C-024

Predictor	Coef	SECoef	T	P	VIF
Constant	342.61	20.21	16.95	0.000	
VRZ1	19.6930	0.2256	87.29	0.000	2.329
RDF055m	-22.988	3.614	-6.36	0.000	1.051
C-024	17.324	2.780	6.23	0.000	2.263

S = 30.5709 R-Sq = 99.8% R-Sq(adj) = 99.8%
PRESS = 42122.7 R-Sq(pred) = 99.77%

7- Analyse de régression

Les statistiques calculées établissent la pertinence du modèle. En effet, la valeur de R^2 signifie que 99,88 % de la variabilité de Ir est expliquée par les 3 descripteurs sélectionnés.

Selon les valeurs du test t (| t |), on peut classer les descripteurs sélectionnés dans ce modèle d'après leur contribution qui se présente dans l'ordre :VRZ1>RDF055m> C-024. Les valeurs des VIF (< 5) suggèrent que ces descripteurs sont faiblement corrélés les uns avec les autres. Ainsi, le modèle peut être considéré comme une équation de régression optimale.

Tableau 12 : Valeurs des Ir expérimentales, calculées, ei, eistd et hii pour l'ensemble de calibration

N°	composés	$Ir_{(exp)}$	$Ir_{(pre)}$	e_i	eistd	h_{ii}
1	Naphthalene	1150	1193.6	-37.35758	-1.27455	0.10041
2	Azulene	1241	1186.7	60.54262	2.06406	0.09909
3	Acetylnaphthalene	1437	1491.1	-51.04728	-1.73504	0.09357
4	Acenaphthene	1458	1431.7	29.69781	1.00614	0.08770
5	Fluorene	1555	1585.0	-25.62474	-0.85156	0.05181
6	1-Methylfluorene	1677	1705.0	-24.94906	-0.83380	0.06247
7	trans-Stilbene	1686	1634.5	56.75035	1.91371	0.07914
8	Phenanthrene	1741	1756.2	-9.69272	-0.32349	0.05988
9	Anthracene	1750	1751.9	3.65705	0.12216	0.06161
10	4,5-Dimethylphenanthrene	1875	1874.2	-0.76721	-0.03345	0.44905
11	1-Methylphenanthrene	1890	1884.3	9.83589	0.32467	0.03895
12	9-Methylanthracene	1920	1880.9	43.34367	1.43098	0.03930
13	Fluoranthrene	2020	2042.4	-19.46422	-0.65191	0.06653
14	pyrene	2070	2080.6	-6.51672	-0.21423	0.03105
15	2,3-Benzofluorene	2195	2206.9	-8.62343	-0.28222	0.02231
16	3,4-Benzofluorene	2195	2188.9	8.76882	0.28905	0.03627
17	1-Methylpyrene	2215	2224.0	-6.21524	-0.20497	0.03717
18	3-Methylpyrene	2220	2224.0	-1.21524	-0.04008	0.03717
19	3,4-Benzophenanthrene	2332	2300.3	33.74927	1.26286	0.25213
20	1,2-Benzanthracene	2389	2399.4	-6.07028	-0.20077	0.04277
21	Chrysene	2395	2406.0	-6.69077	-0.22138	0.04347
22	Triphenylene	2395	2414.6	-15.38646	-0.50844	0.04104
23	Naphthacene	2425	2393.7	35.72423	1.18238	0.04410
24	2,3-Benzofluoranthene	2700	2728.2	-26.84367	-0.90405	0.07678
25	4,5-Benzofluoranthene	2700	2728.2	-26.84367	-0.90405	0.07678
26	8,9-Benzofluoranthene	2706	2726.9	-19.37413	-0.64781	0.06339
27	7,12-Dimethyl-1,2-	2713	2693.3	21.22467	0.70356	0.04701
28	4,5-Benzopyrene	2773	2780.5	-4.73676	-0.15576	0.03156
29	Perylene	2800	2776.9	25.88935	0.85184	0.03276
30	3-Methylcholanthrene	2906	2883.0	22.34411	0.76572	0.10836

Tableau 12 : suite

N°	composés	$Ir_{(exp)}$	$Ir_{(pre)}$	e_i	eistd	h_{ii}
31	12-Methylcholanthrene	2906	2867.7	37.40423	1.28680	0.11524
32	1,2,3,4-Dibenzanthracene	3114	3138.7	-21.88575	-0.73123	0.06197
33	1,2,5,6-Dibenzanthracene	3114	3121.1	-4.15992	-0.13889	0.06070
34	Dibenzo[cd,jk]pyrene	3136	3118.2	20.52064	0.68326	0.05548
35	Picene	3150	3128.2	24.69246	0.82486	0.06164
36	dibenzo(a,L)pyrene	3423	3416.6	4.58544	0.19394	0.41464
37	dibenzo(a,i)pyrene	3477	3547.1	-69.01272	-2.32792	0.07970
38	Coronene	3544	3578.4	-34.93728	-1.20847	0.12479
39	1,2,4,5-Dibenzopyrene	3567	3554.1	14.05043	0.47494	0.08355
40	2,3,6,7-Dibenzopyrene	3600	3547.1	54.00979	1.82182	0.07968

L'analyse des résidus fait ressortir, des erreurs absolues moyennes égales à 23,862. La colonne 07 (tableau 12) donne les valeurs de hii, $i^{ème}$ terme diagonal de la matrice de projection : $H=X\,(X'X)^{-1}\,X'$ ou : X est la matrice des valeurs observées des variables explicatives et X' sa transposée, ces valeurs sont utiles pour le calcul des résidus caractéristiques. La valeur critique pour déterminer les points leviers correspond à $h^* = \frac{3p}{n} = \frac{3X4}{40} = 0.225$. On constate que tous les hii sont inférieures à cette valeur critique à l'exception des composés10 ,19 et 36.

Toutes les valeurs des paramètres statistiques de calibration sont regroupées dans le tableau 13.

Tableau 13 : Valeurs des paramètres statistiques pour l'ensemble de calibration

N	40
R^2	99.8
Q^2	99.77
F	6449.09
S	30.57
SDEC	29.00
SDEP	32.45

Pour la robustesse du modèle est assurés par la valeur de Q^2_{LOO}> 99 alors que les valeurs de l'erreur quadratique moyenne de prédiction et de calcul sont petites et proches ; en plus ce modèle est significatif avec une grande valeur du paramètre de Fisher :
(F = 6449,09).

Le domaine d'application a été discuté à l'aide du diagramme de Williams (figure 12) qui représente les résidus de prédiction standardisés en fonction des valeurs des leviers (h_{ii}).

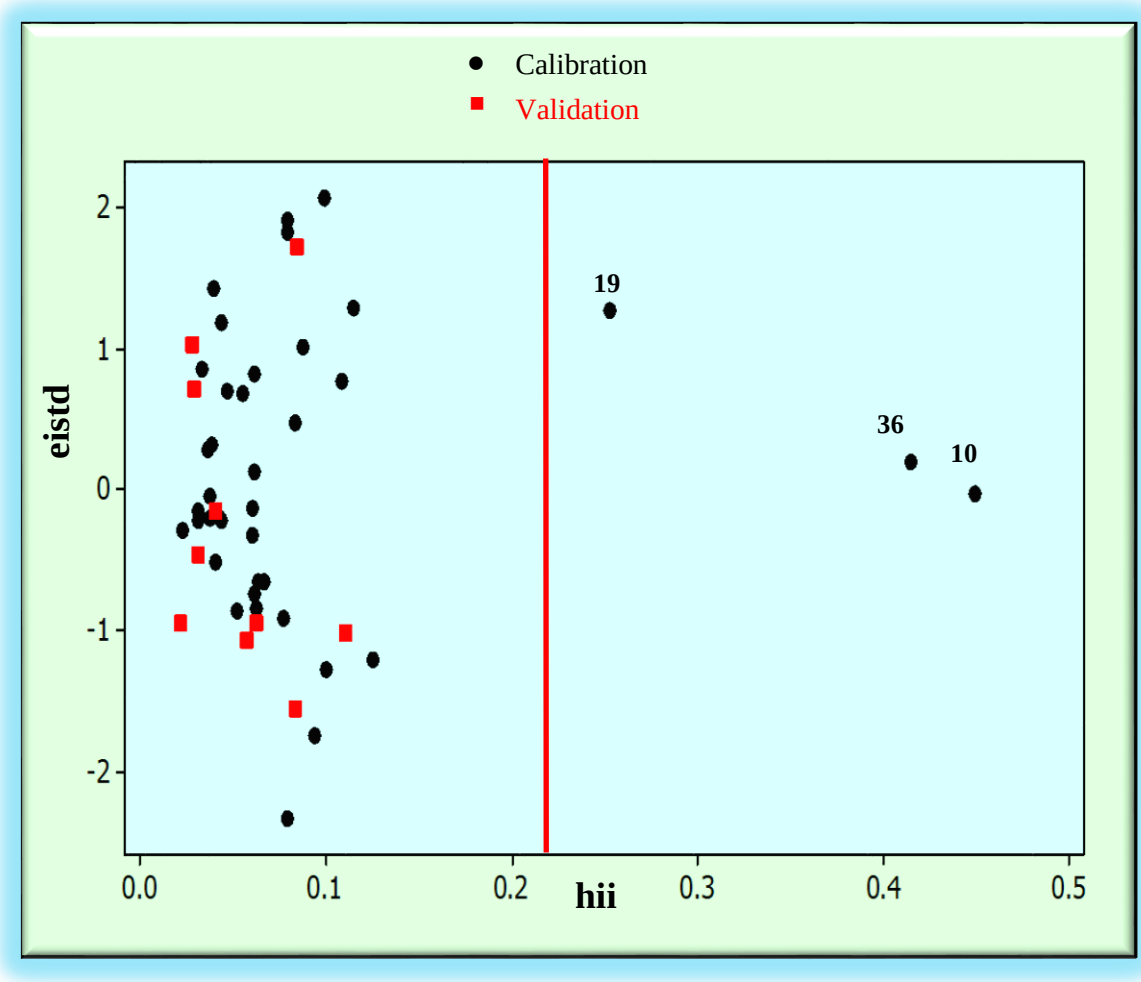


Figure 12 : Diagramme de Williams

Le diagramme de Williams fait ressortir trois points influents ce sont les composés (4,5-Dimethylenephenanthrene) figure 13, (3,4-Benzophenanthrene) figure 14 et (dibenzo(a,L)pyrene) figure 15, le point commun entre ces composés est le réarrangement des cycles avec la présence d'un autre cycle presque fermé qui n'est pas le cas dans l'ensemble des composés.

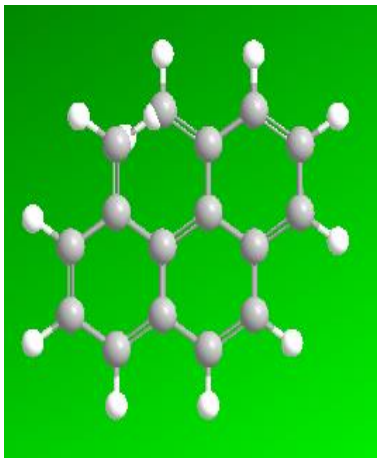


figure 13

4,5-Dimethylenephenanthrene

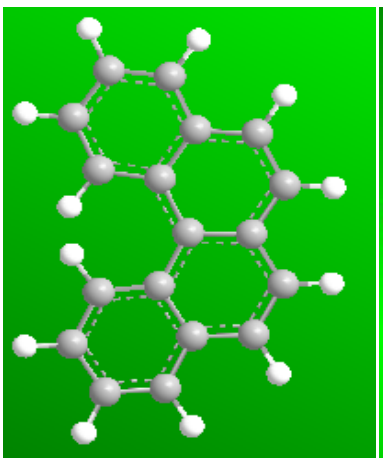


figure 14

3,4-Benzophenanthrene

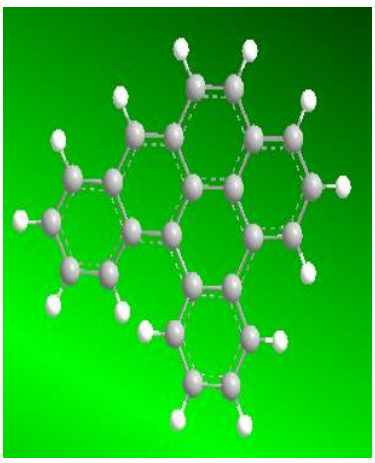


figure 15

dibenzo(a,L)pyrene

8 - Vérification de la qualité de l'ajustement :

La qualité de l'ajustement à été vérifiée en représentant les valeurs calculées de Ir avec notre modèle (colonne 4tableau 12) en fonction des celles observées ou expérimentales (colonne 3tableau 12). La figue 16 montre un bon ajustement traduit par une faible dispersion autour de la droite de régression.

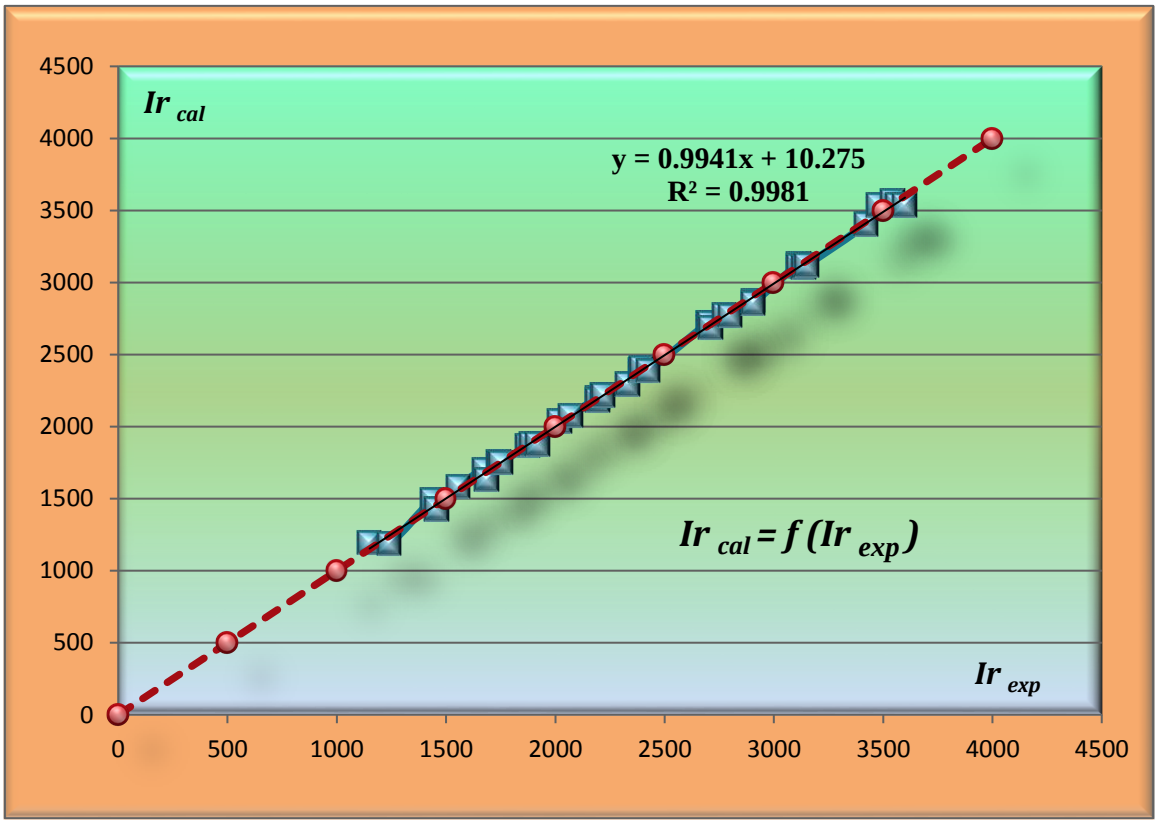


Figure 16 : Graphe des valeurs Ir calculées en fonction des valeurs expérimentales.

9 –Test de randomisation :

Dans le but d'établir que le modèle obtenu n'est pas dû au hasard ou à une sur spécification, nous avons appliqué le test de randomisation. Ainsi 100 nouveaux vecteurs de l'indice de rétention ont été générés par permutation des positions des composantes du vecteur réel.

La figure 16 qui représente le graphe des coefficients statistiques Q^2 et R^2permet de comparer les résultats obtenus pour les modèles randomisés (carrés bleus)au modèle réel de départ (carré rouge brique).

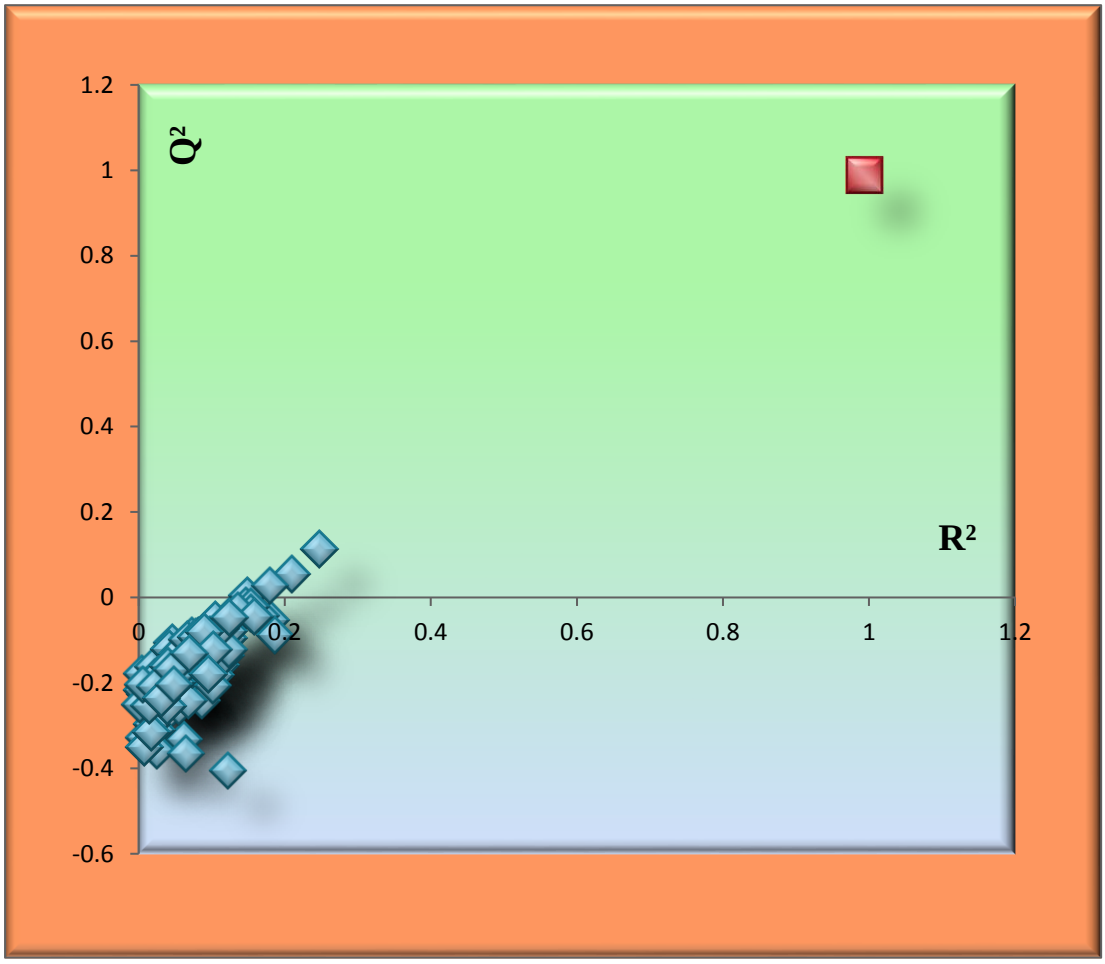


Figure 17 : Test de randomisation

Les carrés bleus représentent l'indice de rétention ordonnée de façon aléatoire, et le carré rouge brique correspond au modèle réel.

Il est clair que les statistiques obtenues pour les vecteurs modifiés sont plus petites que celles du modèle QSPR réel, et pour la majeure partie on obtient même un $Q^2 < 0$. Ceci permet d'assurer qu'une relation structure/rétention réelle a été très bien établie.

10 -Validation externe

Pour généraliser le modèle choisi on procède à une validation externe sur les 10 composés choisis aléatoirement et qui ne font pas partie de l'ensemble d'essai. Les résultats obtenus (tableau 14) montrent que les valeurs prédites sont très proches des valeurs observées, ce qui confirme que le modèle choisi décrit de façon excellente la relation indice de rétention prédite et indice de rétention observée (figure 18).

Tableau 14 : Valeurs des Ir expérimentales, calculées, ei, eistd et hii pour l'ensemble de validation

N°	Composé	$Ir_{(exp)}$	$Ir_{(cal)}$	e_i	eistd	h_{ii}
01	Biphenyl	1408	1444.56	-36.55820	-1.02060	0.110865
02	2-Methylfluorene	1673	1704.46	-31.45590	-0.94820	0.062350
03	2-Methylanthracene	1870	1879.05	-9.05440	-0.15786	0.040040
04	1,2-Benzofluorene	2179	2210.96	-31.96050	-0.93930	0.022211
05	Benzo[g,h,i]fluoranthene	2431	2411.19	19.80500	0.71622	0.029055
06	7-Methyl-1,2-Benzanthracene	2575	2546.60	28.39620	1.03237	0.028391
07	Benzo(a)pyrene	2760	2776.75	-16.74960	-0.45898	0.031363
08	1,2,7,8-Dibenzanthracene	3078	3113.03	-35.02560	-1.07206	0.057195
09	dibenzo(a,e)pyrene	3507	3554.12	-47.12140	-1.55320	0.083545
10	2,3,7,8-Dibenzopyrene	3600	3550.17	49.83110	1.72571	0.083973

Toutes les valeurs des paramètres statistiques de validation sont regroupées dans le tableau ci-dessous :

Tableau 15 : *Valeurs des* paramètres statistiques pour l'ensemble de validation

n	10
Q^2ext	99.77
$SDEP_{ext}$	32.92

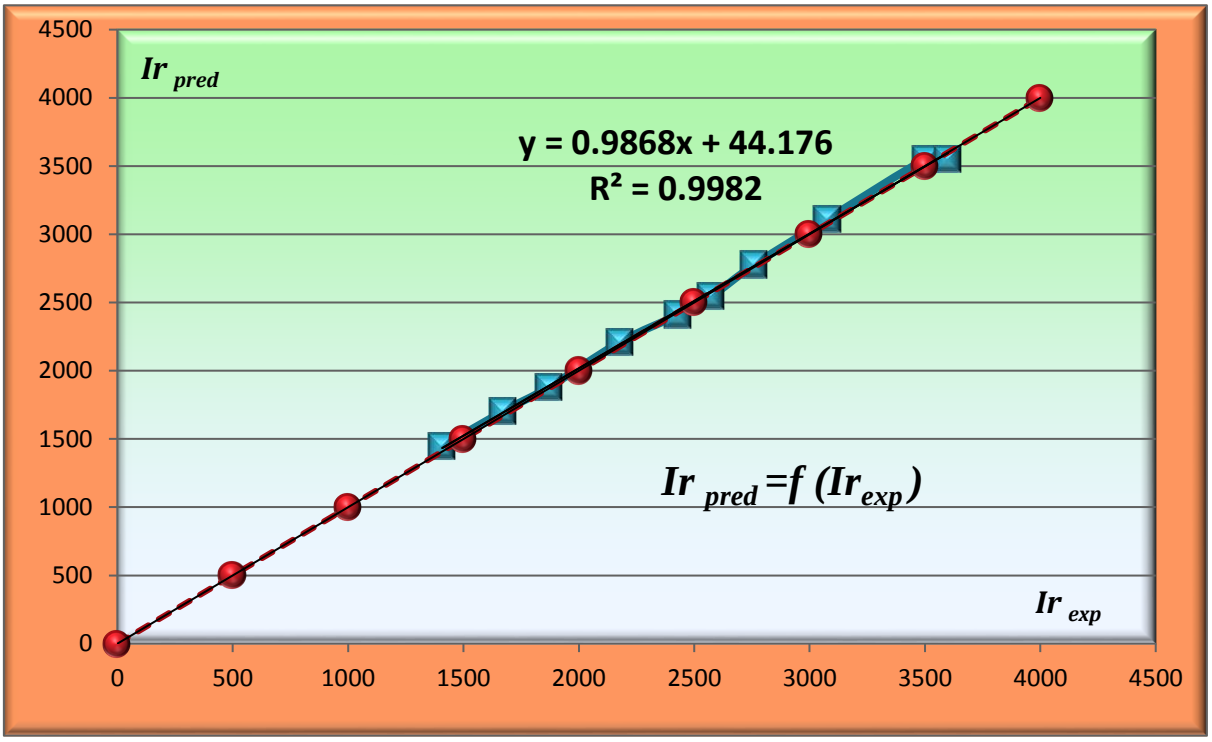


Figure 18 : Graphe des valeurs Ir prédites en fonction des valeurs expérimentales

CONCLUSION GENERALE

CONCLUSION GENERALE

Nous avons appliqué la méthodologie QSRR pour relier la propriété (Indice de rétention) d'un mélange hétérogène d'HAP, comportant un ou plusieurs cycles benzéniques ; à des descripteurs moléculaires théoriques reflétant certaines particularités.

Le modèle QSRR a été établis en utilisant l'analyse de régression multilinéaire.

Les 50 données de base ont été éclatées aléatoirement en deux ensembles disjoints :

- un ensemble principal de 40 composés utilisés pour le calcul et, éventuellement, les essais du modèle ;
- un ensemble de 10 composés pour la prédiction externe.

La taille du modèle (p=3) est fixée par la valeur optimale de R^2 en fonction de nombre de descripteurs. La sélection des variables explicatives a été réalisée par la méthode de pas-à-pas.

Les statistiques ($R^2 > 99$ % ; RMSE < 33 ; $Q^2 > 99$ %) calculées établissent la pertinence du modèle MLR développé.

L'analyse des résidus a permis de détecter trois observations influentes et l'absence des observations abérantes.

La qualité de l'ajustement a été vérifiée en traçant le graphe des valeurs calculées (prédites) en fonction de celles observées, un bon ajustement était observé cela traduit par une faible dispersion.

Les valeurs RMSE sont proches les unes des autres, ce qui permet, tout à la fois, de s'assurer de la bonne capacité prédictive et de la possibilité d'extension suffisante du modèle obtenu.

Ainsi, l'Indice de Rétention (Ir) peut être prédit à partir de la structure moléculaire des HAP en utilisant la modélisation linéaire MLR.

Enfin, d'autres méthodes (RNA, SVM ….) qui peuvent s'avérer plus avantageuses en ce qui concerne la précision et l'interprétation des modèles, et du point de vue de la capacité de généralisation, peuvent être tester.

RÉFERENCES BIBLIOGRAPHIQUES

REFERENCES BIBLIOGRAPHIQUES

[1]. S. Gabet, 2004. Remobilisation d'Hydrocarbures Aromatiques Polycycliques (HAP) présents dans les sols contaminés à l'aide d'un tensioactif d'origine biologique. Thèse de doctorat, université de Limoges, Science. Techno. N°12-2004 : 17p

[2]. S. Kuony, 2005. Caracterisation D'arene Dioxygenases Impliquees Dans La Biodegradation Des Hydrocarbures Aromatiques Polycycliques Chez Mycobacterium Sp. 6py1. Universite Joseph Fourier – Grenoble I Sciences Et Geographie. Cnrs Umr5092. 153p.

[3]. A. Martinez- bernal, 2005. Elimination Des Hydrocarbures Aromatiques Polycycliques Présents Dans Les Boues D'épuration Par Couplage Ozonation – Digestion Anaérobie. Universite Montpellier Ii Sciences Et Techniques Du Languedoc. Thèse Docteur. Inra. 223p.

[4]. _http://uved.univ nantes.fr/GRCPB/sequence2/html/ressources/matrice/hap/fiche_hap _benzoapyrene.pdf 1/2/2018

[5]. http://www.seine-aval.fr/wp-content/uploads/2017/01/Contamination-HAP.pdf 5/2/2018

[6]. F. Lawrence, 1973. Chromatography of Environmental Hazards, Elsevier Scientific Publishing Company Amsterdam / London / New York. Vol II.

[7]. http://www.rocler.qc.ca/pdubreui/chromatographie/CG/chroma2.html 10/02/2018

[8]. K. Heberger, 2007. Quantitative Structure-(Chromatographic) Retention Relationship. Journal of Chromatography A. Vol 1158. Springer, Dordrecht.255p.

[9]. A.R. Katritzky, D.C. Fara., R.O. Petrukhin, 2002.Top. Med. Chemm 1333-1356.

[10]. C. Hansch, 1969, a quantitative approach to biochemical structure activity relationships. Accounts of chemical research, 2, 232-239.

[11]. R. Leardi , 2001 Chemometr. 15, 559-569.

[12]. L. Fisbbein, 1973. Chromatography of Environmental Hazards, Vol. II. Elsevier, Amsterdam

[13]. ChemDraw Release 2002. Chemical Structure Drawing Standard, 7.0.1

[14]. HyperchemTM Release 2000. for windows, Molecular Modelling system, 7.5

[15]. R. Todeschini, V. Consonni, M. Pawan, 2005. DRAGON, Software for the calculation of Molecular Descriptors. Release 5.3 for Windows, Milano.

[16]. T. Thomas-Danguin, 1997.Intensité olfactive des composés purs et de mélanges: application au masquage des odeurs, Université Claude Bernard, Lyon, p224.

[17]. Dragon_ Aide blocs des descripteurs.

[18]. P. Le Cloirec, 2002. Introduction au traitement de l'air, Les techniques de l'ingénieur

traité environnement (G 1700) : 1-8.

[19]. Saadi Khaled. 2009. Contribution l'étude de la Relation structure chimique- odeur Utilisation de la technique Random Forest (Application à la famille des pyrazines), Mémoire de Magister. UNIVERSITE KASDI Merbah Ourgla.p30.

[20]. AI ACCESS, 91940, Les Ulis, France.

[21]. M. karelson,2000. Molecular descriptors in QSAR/QSPR. Wiley- Interscience, p. 385

[22]. B. Kowalski, R. Gerlach, H.Wold,1982. Systems under Indirect Observation (K. Jorenskog et H. Wold, eds.), North Holland, Amsterdam, 191-206 .

[23]. L. Erikson, E. Johannson, N. Kettaneh- Wold,2001.Multi and Megavariate Data Analysis- Principles and Applications. Umetrics Academy, Ume.

[24]. S. Wold, A. Ruhe, H.Wold, W. Dunn,1984. SIAMJ. Sci. Stat. Comput., 5, 735.

[25]. S. Wold,1984. Chemometrics: Mathematics and Statistics in Chemistry. Reidel, Dordrecht, The Netherlands.

[26]. P. Gelada, B. R. Kowalski, Anal. 1986. Chim. Acta, 185, 1.

[27]. A. R. Katritzky, V. S. Lobanov, M. Karelson,1994. CODESSA Reference Manual. University of Florida, Gainesville.

[28]. V. Y. Nalimov, 1962. The Application of Mathematical Statistics to Chemical Analysis, Addison- Wesley, Reading, MA.

[29]. R. Calcutt, R. Body,1983. Statistics for Analytical Chemists. Champman & Hall, New York .

[30]. J. C. Miller, J. N. Miller,1988. Statistics for Analytical Chemistry. Ellis Horwood, New York .

[31]. P. C. Meier, R. E. Zund,1993. Statistical Methods in Analytical Chemistry. Wiley, New York .

[32]. P. Dagnélie,1998. Statistique Théorique Et Appliquée. Tomes 1 et 2. De Boeck & Larcier s. a.

[33]. R. Tomassone, E. Lesquoy, C. Miller,1983. La régression : nouveaux regards sur une ancienne méthode statistique. Masson, INRA .variables. Ecology 89(9): 2623-2632.

[34]. MINITAB Release 16.2.0.0 for Microsoft language pack 2.Variable Subset Selection by Genetic Algorithm.Release for Windows. Milano Srl.

[35]. N.R Draper, H. Smith, 1998. Applied Regression Analysis, Third Edition, Wiley series in Probability and Statistics, New York.

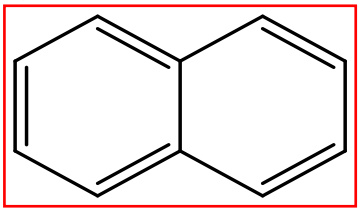
[36]. L. Eriksson, J. Jaworska, A.P. Worth,2003. M.T.D. Perspective, 111(10), 1361-1375.

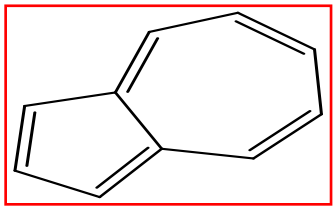
ANNEXE

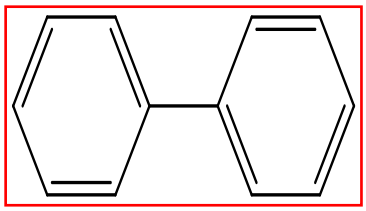
ANNEXE

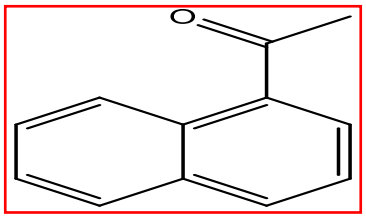
La structure chimique des 50 HAP dessinés par le logiciel Chemdraw

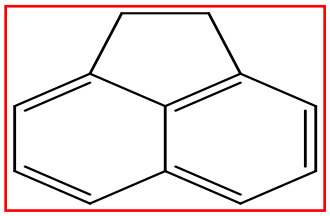
(*) Composés de validation

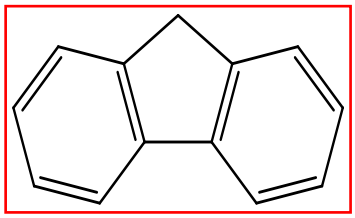

Naphthalene

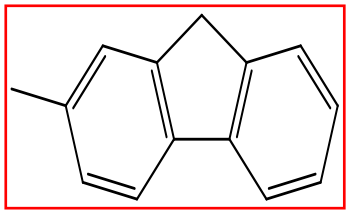

Azulene

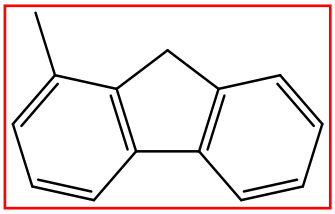

Biphenyl*

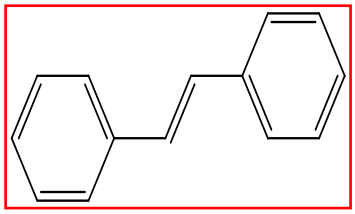

Acetylnaphthalene

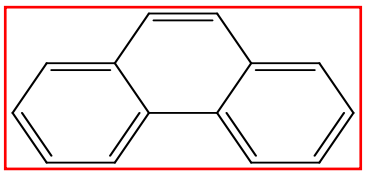

Acenaphthene

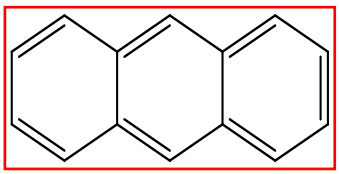

Fluorene

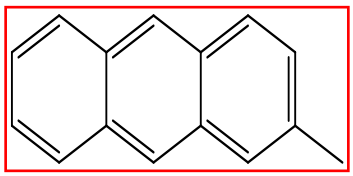

2-Methylfluorene*

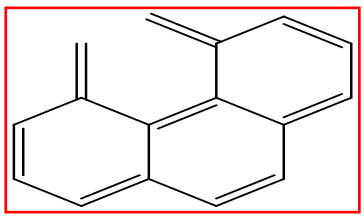

1-Methylfluorene

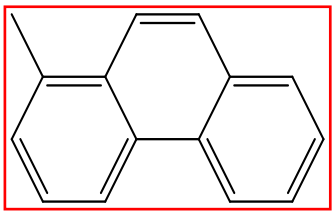

trans-Stilbene

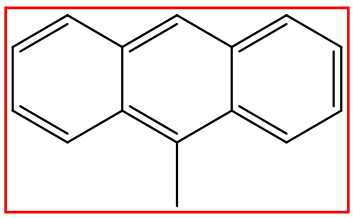

Phenanthrene

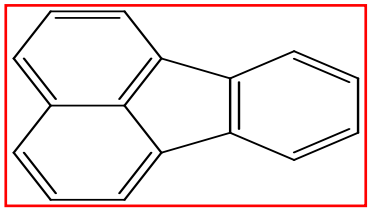

Anthracene

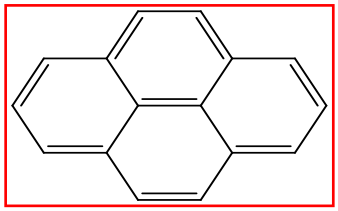

2-Methylanthracene*

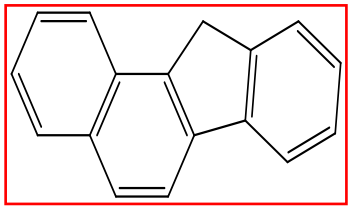

4,5-Dimethylephenanthrene

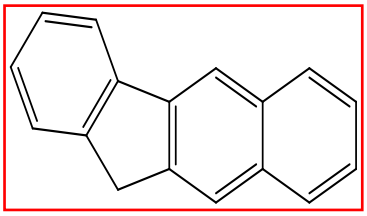

1-Methylphenanthrene


9-Methylanthracene

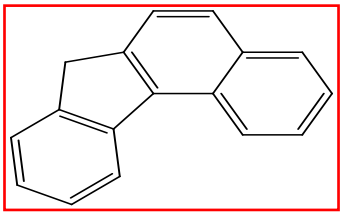

Fluoranthrene


pyrene

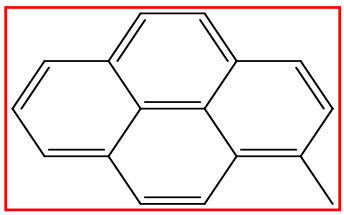

1,2-Benzofluorene*



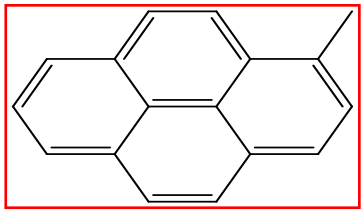
2,3-Benzofluorene



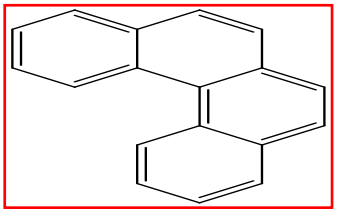
3,4-Benzofluorene



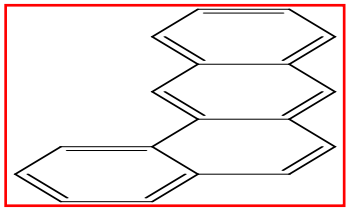
1-Methylpyrene



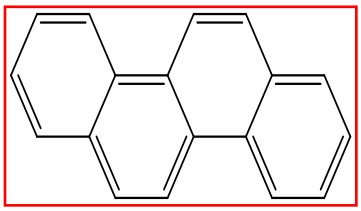
3-Methylpyrene



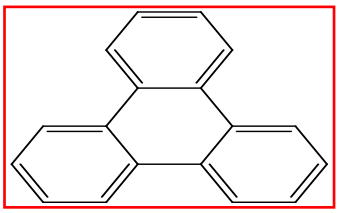
3,4-Benzophenanthrene



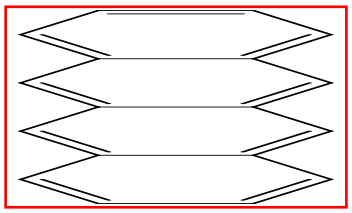
1,2-Benzanthracene



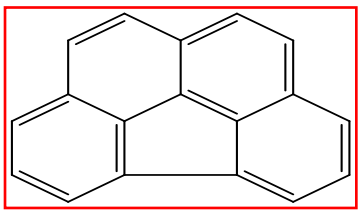
Chrysene



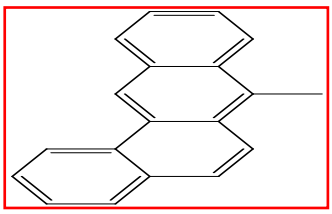
Triphenylene



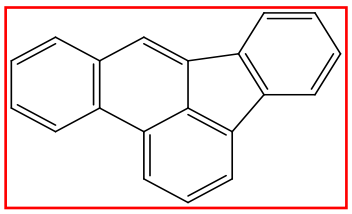
Naphthacene



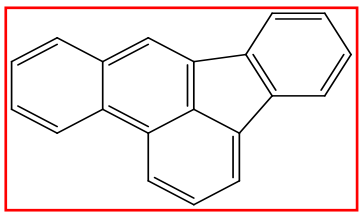
Benzo[g,h,i]fluoranthene*



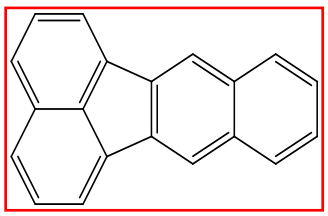
7-Methyl-1,2-Benzanthracene*



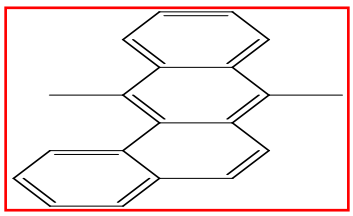
2,3-Benzofluoranthene



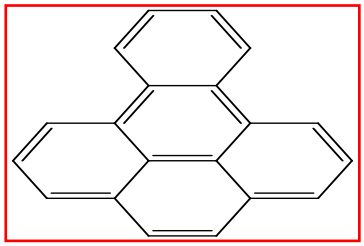
4,5-Benzofluoranthene



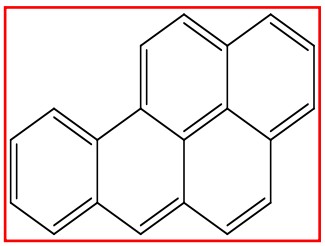
8,9-Benzofluoranthene



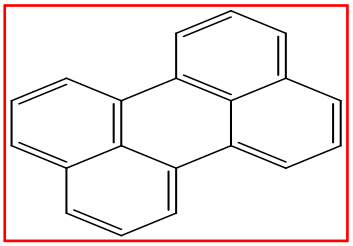
7,12-Dimethyl-1,2-benzanthracene



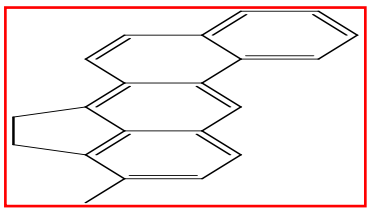
1,2-Benzopyrene



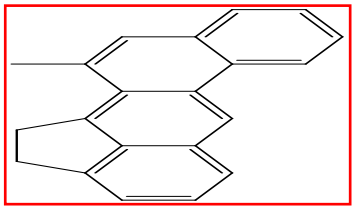
Benzo(a)pyrene*



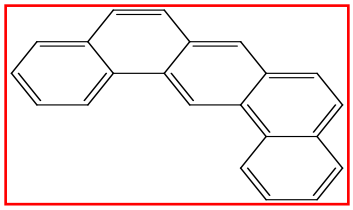
Perylene



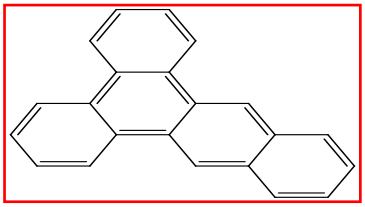
3-Methylcholanthrene



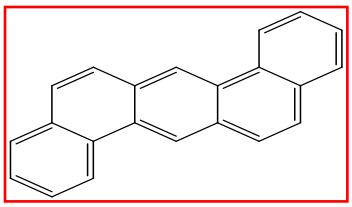
12-Methylcholanthrene



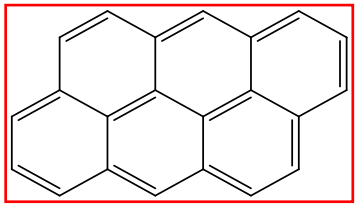
1,2,7,8-Dibenzanthracene*



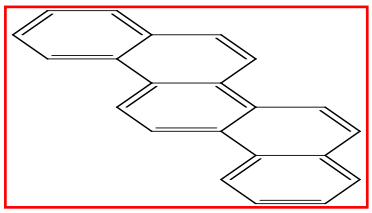
1,2,3,4-Dibenzanthracene



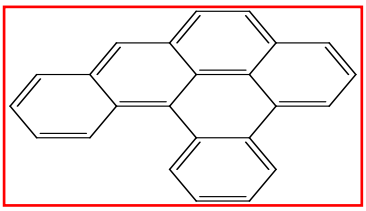
1,2,5,6-Dibenzanthracene



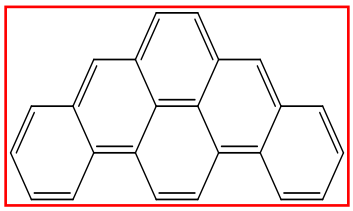
Dibenzo[cd,jk]pyrene



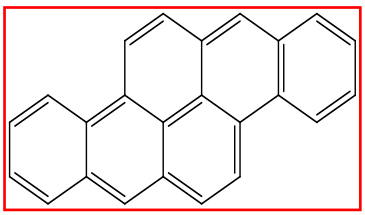
Picene



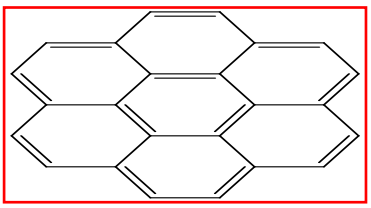
dibenzo(a,L)pyrene



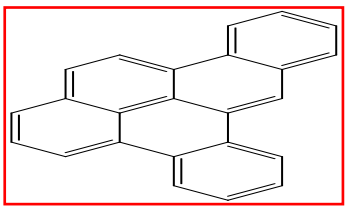
dibenzo(a,i)pyrene



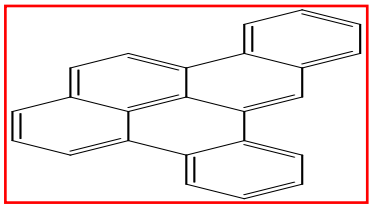
dibenzo(a,e)pyrene*



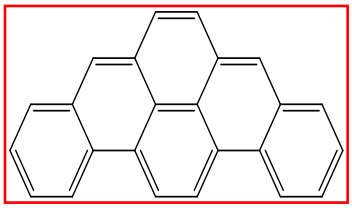
Coronene



1,2,4,5-Dibenzopyrene



2,3,7,8-Dibenzopyrene*



2,3,6,7-Dibenzopyrene

ملخص

دراسة الاحتباس الكروماطوغرافي للمركبات العطرية متعددة الحلقات بواسطة QSRR

تحليل ودراسة الهيدروكربونات العطرية متعددة الحلقات تتطلب معرفة الخصائص العامة والفيزيائية الكيميائية المتعلقة بأثرها على البيئة. في هذا العمل نقترح ربط خطيا مؤشر الاحتفاظ (في كروماتوغرافيا الغاز) من مئات الهيدروكربونات العطرية متعددة الحلقات مع الوصفات الجزيئية النظرية من خلال نهج QSRR، وتحسب هذه الواصفات باستخدام برامج النمذجة دراغون.

وسيتم تأكيد الإحصاءات المختلفة التي تم تحديدها لمجموعات المعايرة والتحقق (معامل التحديد والتنبؤ: R^2، Q^2 والانحراف المعياري s).

الكلمات الدالة :

الهيدروكربونات العطرية متعددة الحلقات - مؤشر الاحتفاظ - كروماتوجرافيا الغاز - التمثيل العددي للهيكل الكيميائي نموذج QSRR.

ABSTRACT

QSRR study of the chromatographic retention of PAH

The analysis and study of polycyclic aromatic hydrocarbons (PAHs) require knowledge of the general and physicochemical properties related to their impact on the environment. In this work we propose to linearly correlate the retention index (in gas chromatography) of hundreds HAP with theoretical molecular descriptors by a QSRR approach, these descriptors are calculated using the DRAGON modeling software.

The different statistics established for the calibration and validation sets (coefficient of determination and prediction: R^2, Q^2 and standard deviation s) will be confirmed.

Keywords:

Polycyclic Aromatic Hydrocarbons - Retention index - Gas chromatography - Numerical representation of the chemical structure - Model QSRR.

RESUME

Etude QSRR de la rétention chromatographique des HAP

L'analyse et l'étude des hydrocarbures aromatiques polycycliques (HAP) requièrent la connaissance des propriétés générale et physico-chimiques liées à leur impact sur l'environnement. Dans ce travail nous proposons de corréler linéairement l'indice de rétention (en chromatographie en phase gazeuse) de centaines HAP avec des descripteurs moléculaires théoriques par une approche QSRR, ces descripteurs sont calculés en utilisant le logiciel de modélisation DRAGON.

Les différentes statistiques établies pour les ensembles de calibration et de validation (coefficient de détermination et de prédiction :R^2 ; Q^2 et écart type s) seront confirmées.

Mot clés:

Hydrocarbures Aromatiques Polycycliques – Indice de rétention – Chromatographie en phase gazeuse - Représentation numérique de la structure chimique – Modèle QSRR.