وزارة التعليم العالى و البحث العلمى

Université BADJI Mokhtar – Annaba
BADJI Mokhtar – Annaba University

جـامـعـة بـاجـي مـخـتـار – عـنـابـة

Faculté des Sciences
Département de Chimie

MÉMOIRE

Présenté pour l'obtention du diplôme de Magistère en chimie Analytique

Par : Karima BENNECIB

Option : Chimie de l'environnement

THÈME

Comportement de quelques phénols, polluants potentiels de l'environnement

Devant le jury :

Président :	**Mme. FERTIKH Nadia**	**M.C**	**U. B. M. Annaba**
Rapporteur :	**Mme. LARKEM Hamama**	**M.C**	**U. B. M. Annaba**
Examinateurs :	**Mr. MESSADI Djelloul**	**Pr**	**U. B. M. Annaba**
	Mme. ALI MOKHNACHE Salima	**Pr**	**U. B.M Annaba**
Invité	**Mr. SOUICI Mohammed Lotfi**	**M.C**	**U. B. M. Annaba**

Année 2009

SOMMAIRE

Pages

RESUMES *IV, V , VI*

LISTE DES TABLEAUX *VII*

LISTE DES FIGURES *IX*

SYMBOLES ET ABREVIATIONS *XI*

INTRODUCTION GENERALE ***2***

GENERALITES ***5***

CHAPITRE I APPROCHE EXTRATHERMODYNAMIQUE

I – RELATIONS EXTRATHERMODYNAMIQUES ***19***

I– 1 Généralités ***19***

I – 2 Relations de type extrathermodynamique ***20*** ***19***

I – 3 Application à la chromatographie

II – PHENOLS ***22***

III – LES MODELES ***23***

IV – MATERIEL ET METHODE ***24***

V – RESULTATS ET DISCUSSION ***25***

VI – CONCLUSION ***33***

CHAPITRE II APPROCHE QSAR

I-OPTIMISATION DE LA GEOMETRIE MOLECULAIRE ET GENERATION DES DESCRIPTEURS MOLECULAIRES ***35***

II- SELECTION D'UN SOUS-ENSEMBLE DE DESCRIPTEURS 35

SIGNIFICATIFS

II-1 Principe. ***35***

II – 2 Initialisation aléatoire du modèle ***36***

II – 3 Etape de croisement ***36***

II – 4 Etape de mutation ***36***

II – 5 Conditions d'arrêt ***37***

III– DEVELOPPEMENT DES MODELES ***37***

III– 1 La régression linéaire multiple (MLR)

38

III – 2 Les réseaux de neurones ***38***

III – 2 -1 Le neurone artificiel

38

III – 2 - 2. Propriétés des réseaux de neurones ***39***

III – 2 -.3.Les différents types de réseaux de neurones ***40***

III – 2 -.3-1 Les réseaux multicouches ou perceptrons multicouches (PMC) ***41***

III – 2 -.4. Apprentissage ***42***

III – 2 -.4.- 1 L'apprentissage de Widrow-Hof ***43***

III – 2 -.4.- 2 L'apprentissage par rétro propagation du gradient (Levenberg-Marquardt backpropagation) ***44***

III – 2 –5 Critères d'arrêt **45**

III – 2 –6 Construction d'un modèle ***45***

46

III – 2 –6- 1 Construction de la base de données ***46***

III – 2 –6- 2 Définition de la structure du réseau ***47***

III – 2 – 6 - 3 Nombre de couches et de neurones cachés ***47***

III – 2 – 6 - 4 Présentation de l'environnement utilisé ***48***

IV **– PARAMETRES D'EVALUATION DE LA QUALITE DE L'AJUSTEMENT** ***49***

IV – 1 Robustesse du modèle ***49***

IV – 2 Détection des observations aberrantes

50

IV – 3 Test de randomisation ***50***

IV – 4 Validation externe ***51***

PARTIE EXPERIMENTALE

FACTEUR DE CAPACITE K' ***53***

I-1- Modèle hybride algorithme génétique

/ réseaux de neurones artificiels ***53***

I - 1- 1 choix des paramètres statistiques ***54***

I - 1- 2 Choix du nombre de couches cachées ***54***

I - 1- 3 Choix du nombre de neurones dans la couche cachée *54*

I - 1- 4 Choix de la fonction de transfert *54*

I - 1- 5 Choix des paramètres d'apprentissage *54*

II - 1- 6 Résultats et discussion *55*

I - 1- 6 – 1 Evaluation de la qualité de l'ajustement *55*

I - 1- 6 – 2 Vitrification de la qualité de l'ajustement *55*

I-2 Validation externe *56*

V – CONCLUSION GENERALE *59*

ANNEXE *62*

REFERENCES BIBLIOGRAPHIQUES

REMERCIEMENTS

Ce mémoire n'aurait pas vu le jour sans la confiance, la patience et la générosité du responsable de la P.G. Monsieur le Professeur D. MESSADI que je remercie vivement. Je voudrais aussi le remercier pour le temps et la patience qu'il ma accordés tout au long de ces années et pour avoir accepté d'examiner ce travail.

Je tiens également à remercier :

Madame LARKEM.H, pour la direction de ce travail ;
Madame. N.FERTIKH, pour avoir acceptée la présidence de ce jury;
Madame. S. ALI-MOKHNACHE, pour avoir acceptée de participer à ce jury;.
Monsieur M. L.SOUICI pour avoir accepté l'invitation.

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

Dédicaces

Je dédie ce travail

A la personne qui mérite le diplôme de magistère grâce à son: soutien moral, sa patience, ses encouragements infinis, ainsi que son sourire au mal et au bon .Que le Bon dieu prenne soin d'elle pour moi, mes sœurs et mes frères.

A **Maman** *je le dédie*

Karima-B

ملخص:

لقد استخدم تقربين مختلقين للربط معامل الاحتفاظ للفينولات غير المجانسة و التي تم فصلها و باختلاف الظروف التجريبية (درجة الحرارة، الاختبار و تكوين PI-HPLC بطريقة الطور المتحرك).

إن التقرب بعلاقات extra thermodynamiqueالمطبقة و المحددة مسبقا لحساب حجم الفوائد.

ABSTRACT

Two different approaches have been used to relate the retention factor of non congeneric phenols separated by reversed phase HPLC for different experimental conditions (temperature testing, mobile phase composition).

The approach of extrathermodynamique type is complicated and limited in advance to calculate the end time.

The hybrid approach QSRR: Genetic algorithms; imposing temperature experiments and the composition of the methanol mobile phase hydro-organic products as descriptors / neural networks leads to a robust model with good predictive ability.

Key words : : Phenols; HPLC-PI; Model LEFR ; QSRR; genetic algorithm / RNA

Résumé:

Deux approches différentes ont été utilisées pour relier le facteur de rétention de phénols non congénères séparés par CLHP-PI pour différentes conditions expérimentales (température d'expérimentation ; composition de la phase mobile)

L'approche extrathermodynamique est compliquée et limitée dans le calcul à l'avance de la grandeur d'intérêt.

L'approche QSRR hybride:Algorithmes génétiques ;en imposant la température des expérimentations et la composition en méthanol de la phase mobile hydro-organique comme descripteurs expérimentaux /réseaux de neurones conduit à un modèle robuste avec une bonne capacité prédictive.

Mots clés : Phénols ; CLHP-PI ; Modèle LEFR ; QSRR ; Algorithme génétique/RNA

LISTE DES TABLEAUX

	Titre	Page(s)
Tableau I	*Valeurs de k' mesurées en fonction de la température (t) et de la fraction volumique (x) du méthanol*	26
Tableau II	*Jeux de constantes A, B, C (éq. (9)) pour les 10 phénols*	26
Tableau III	*Pente* $\left(b = - \frac{\Delta H}{R} \right)$ *de la droite d'équation (6) et coefficients, r, de Bravais – Pearson*	28
Tableau IV	*Coefficients* A_i *(i = 1 à 3) de l'équation (9) pour l'ensemble d'estimation des phénols*	31
Tableau V	*Descripteurs moléculaires intervenant dans la modélisation de facteurs de capacité*	53
Tableau VI	*Structure optimale du réseau de neurones*	55
Tableau VII	*Les valeurs K' observées, prédites et les erreurs pour l'ensemble de validation externe par RNA*	57/58

LISTE DES FIGURES

	Titre	Page(s)
Figure 1	*Noms et structures des phénols étudiés*	25
Figure 2	*Corrélation entre les k' mesurés et calculés par le modèle de Kowalska*	27
Figure 3	*Variation de Ln k' à 305 K en fonction de* $b = -\frac{\Delta H}{R}$ ***a/*** *pour les phénols étudiés ;* ***b/*** *après élimination du p- nitrophénol et de l'acide salicylique*	29
Figure 4	*k' des 59 données d'estimation calculées par le modèle de Horváth en fonction des k' mesurés*	31
Figure 5	*k' des 12 données d'essai calculés d'après le modèle de Horváth en fonction des k' mesurés*	32
Figure 6	*le neurone artificiel générique*	39
Figure 7	*Fonctions d'activation*	39
Figure 8	*Structure générale du perceptron multicouches*	41
Figure 9	*Apprentissage par un algorithme de rétro-propagation*	44
Figure 10	*Illustration de l'arrêt précoce*	45
Figure 11	*Variation du FIT en fonction du nombre de descripteurs pour K'*	53
Figure 12	*Choix du nombre de neurones de la couche cachée*	54
Figure 13	*Graphe des valeurs prédites* $\hat{K}'$ *en fonction des valeurs observées K'*	56
Figure 14	*Graphe des* $\hat{K}'$ *prédites en fonction des* K' *observées pour validation*	57

SYMBOLES ET ABREVIATIONS

AM1 : Austin Model 1.

DFITS : Statistique permettant de mesurer l'influence d'une observation i sur la valeur ajustée.

Di : Distance de COOK.

d : Statistique de Durbin-Watson.

di : Résidu standardisé.

EQM: Erreur quadratique moyenne.

EQMP: Erreur quadratique moyenne sur l'ensemble de prédiction.

$EQMP_{ext}$: Erreur quadratique moyenne sur l'ensemble de prédiction externe.

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

F : Statistique de Fisher.

FIT: Fonction de KUBINYI.

GA: Algorithme génétique (Genetic Algorithm).

$\underset{\sim}{\mathbf{H}}$: Matrice de projection, ou matrice châpeau.

hii : Eléments diagonaux de la matrice châpeau.

IR: indice de rétention.

IW : Poids entrée-couche cachée.

k': facteur de capacité.

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

LW : Poids couche cachée-sortie.

MLR: Régression linéaire multiple.

MCP: Moindres carrés partiels.

n: Dimension de la population (échantillon).

n-p : Nombre de degrés de liberté.

PMC: Réseaux multicouches.

PRESS : Somme des carrés des erreurs de prédiction.

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

p_c : Probabilité de croisement.

p_M: Probabilité de mutation.

QSRR : Quantitative Structure/ Retentions Relationships. (Relations Structure/ Rétention Quantitatives).

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

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

RNA: Réseaux de neurones artificiels.

R^2 : Coefficient de détermination.

r_i : Résidu studentisé interne.

S : Erreur standard.

SCE : Somme des carrés des écarts.

SCT : Somme des carrés totale.

t : t de Student.

t_i : Résidu studentisé externe.

w_k : Poids à l'instant k.

w_{k+1} : Poids à l'instant k-1.

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

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

$\underset{-}{x_j}$: Variable explicative.

x_j : $j^{ième}$ valeur de x .

x_{max} : Valeur maximale.

$x_{\min}$: Valeur minimale.

x_{norm} **:** Valeur normalisée.

$\underset{-}{y}$: Vecteur de dimension n.

y_i : Valeur observée.

$\hat{y}_i$: Valeur estimée.

α : Niveau de confiance; Facteur d'apprentissage.

σ^2 **:** Variance.

δ_k : Différence entre la sortie attendue et la sortie effective à l'instant k.

INTRODUCTION GENERALE

Introduction générale

Les phénols sont des composés importants biologiquement et du point de vue de l'environnement .Non seulement ils possèdent d'importantes fonctions physiologiques et certaines activités pharmaceutiques,mais ils peuvent encore influencer la saveur de certaines boissons .D'autres composés phénoliques sont utilisés pour le tannage ,en cosmétique,dans l'industrie organique (fabrication de matières plastiques,produits pharmaceutiques,explosifs...) ,ainsi que pour le développement photo, ce qui en fait d'importants polluants potentiels de l'environnement

Les relations structure/Retention quantitatives, désignées par l'abréviation QSRR (Quantitave Structure/Retention Relationships), très utilisées depuis une vingtaine d'années, constituent des modèles mathématiques pour l'approximation des relations, souvent complexes, entre la structure caractérisée par des descripteurs moléculaires, et les propriétés physico-chimiques des composés.

Les techniques les plus courantes pour établir des modèles QSRR 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.

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 avons appliqué des méthodes hybrides: algorithme génétique/réseaux de neurones artificiels (GA/RNA) pour modéliser, le facteur de capacité de 10 phénols.

Notre mémoire comporte en plus de la bibliographie,nous avons présenté des généralités sur les phénols (Définitions Origine/fabrication;les phénols et l'environnement dangers pour la santé publique……), chromatographie, et deux grandes parties :

Dans une première partie nous avons utilisé une approche extrathermodynamique pour modéliser les facteurs de rétention (ou de capacité) de 10 phénols séparés par CLHP – PI, à

différentes températures, et pour différentes compositions d'une phase mobile méthanol – eau.

Dans la deuxième partie, nous avons développé tout ce qui a trait au pré-traitement des molécules (introduction des molécules, optimisation de leur 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: algorithmes génétiques, réseaux de neurones artificiels, traitement statistique pour l'évaluation de la qualité de l'ajustement (robustesse des modèles; détection des observations aberrantes; test de randomisation; validation externe).

Ensuite, nous présentons et discutons les modèles calculés.

PARTIE GENERALITES

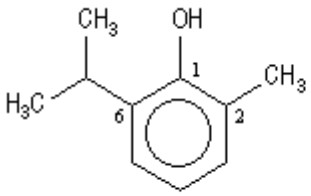
Phénols

Définitions

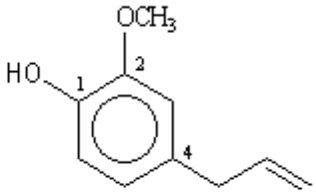
On appelle phénols les dérivés hydroxylés du benzène et des hydrocarbures aromatiques, dans lesquels le groupe OH est lié à un atome de carbone du cycle benzénique. Les dérivés polyhydroxylés sont appelés polyphénols. Rappelons que chez les alcools le groupe OH est lié à un atome.

Nomenclature

Ils sont nommés comme des phénols substitués. Le numéro le plus petit est affecté à l'atome de carbone porteur du groupe -OH.



6-Isopropyl-2-méthylphénol



2-Méthoxy-4-(prop-2-ényle)-phénol

Origine/fabrication

Ce sont des alcools aromatiques qui proviennent des végétaux. Les phénols simples, déchets du métabolisme végétal, sont assemblés en polyphénols comme la lignine. Les composés phénoliques définissent un ensemble de substances que l'on a appelées pendant longtemps " matières tannoïques " d'une façon générale et imprécise parce qu'on ne connaissait pas, avec suffisamment de précision, la nature de ces substances. Il y a quatre principales familles de composés phénoliques : les acides phénols, les flavones, les anthocyanes les tanins

Dans le groupe des phénols, les crésols et la substance mère elle-même sont le principal composé, mais il convient également de mentionner le thymol, les naphtols, la phénophtaléine, le trichlorophénol et le pentachlorophénol. Les composés naturels tels que la pyrocatéchine, le gayacol et leurs dérivés n'ont pas d'effets toxicologiques significatifs. L'un des dérivés de la pyrocatéchine, l'adrénaline, est bien connu. Le phénol est présent à l'état naturel dans le bois et les aiguilles de pin, dans l'urine des herbivores (sulfate phénolique) et dans le goudron de houille. Les phénols monovalents forment dans la nature de nombreuses substances odorantes (par exemple vanille, thymol dans le thym, carvacrol, zingivérone dans le gingembre, aldéhyde salicylique). Parmi les phénols polyvalents de synthèse,

l'hexachlorophène est particulièrement toxique. Le phénol est obtenu par distillation du goudron de houille (sel. RÖMPP 1983, 1 t de houille permet d'obtenir env. 0,25 kg de phénol). Cependant, la méthode de synthèse qui prévaut à l'heure actuelle est la dissociation de l'hydroperoxyde de cumène, avec comme sous-produit l'acétone. Du phénol est aussi parfois produit à partir du benzène en passant par l'acide sulfonique de benzène ou par le chlorobenzène.

Des émissions sont produites par la combustion incomplète d'essence et de goudron de houille, dans les effluents **Pathologie/toxicologie**

Le phénol est utilisé pour la fabrication de produits tels que résines synthétiques, colorants, produits pharmaceutiqu**e.**

Propriétés des phénols

1-Propriétés physiques

1-1Structure de la molécule de phénol

L'énergie de résonance évaluée grâce à la réaction d'hydrogénation vaut 167 $kJ.mol^{-1}$. Elle est donc plus élevée que pour le benzène (150 $kJ.mol^{-1}$). On interprète ce résultat par la participation d'un doublet non liant de l'atome d'oxygène à la résonance. Les mesures aux rayons X montrent que la molécule est plane ce qui autorise une délocalisation maximale. Cette participation à la délocalisation électronique se traduit aussi par le raccourcissement de la longueur de la liaison C-O et par l'augmentation de l'énergie de cette liaison par rapport à celle d'un alcool comme le cyclohexanol.

2-1Constantes physiques

Les températures de changement d'état des phénols sont plus élevées que celle des hydrocarbures de même masse molaire. On l'interprète par le fait que ces composés sont associés par liaison hydrogène intermoléculaire. Le phénol lui même est un solide à la température ordinaire.

TF (°C)	**TE** (°C)	**s/H_2O** ($g.L^{-1}$) (20 °C)	**μ** (D)
41	181	93	1,59 (Ph vers OH)

La miscibilité avec l'eau dépend beaucoup de la température. Elle est totale si T > 63 °C.

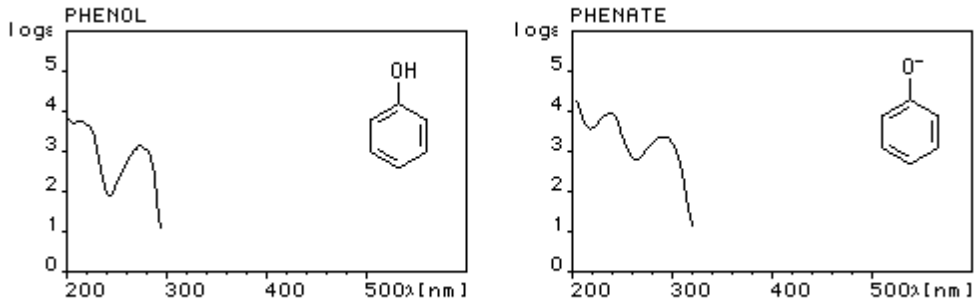
3-1Spectroscopie infrarouge

- En solution dans un solvant apolaire, on observe un pic fin à 3610 cm^{-1}. Il s'agit de la vibration d'élongation de la liaison O-H libre.
- Pour le composé pur, on observe une bande large 3200 $cm^{-1} < \sigma < 3400\ cm^{-1}$. Il s'agit des liaisons O-H associées par liaison hydrogène intermoléculaire.

Certains composés polyfonctionnels comme le 2-hydroxybenzaldéhyde (salicylaldéhyde) possèdent une liaison hydrogène intramoléculaire (σ # 3480 cm^{-1}). Ce type de liaison se distingue facilement d'une liaison intermoléculaire. En effet, un tel pic n'est pas affecté lors de la dilution du composé dans un solvant inerte comme CCl_4.

4-1.Spectroscopie UV-Visible

Le phénol absorbe dans l'ultraviolet. Ses solutions sont incolores. La déprotonation et le passage à l'ion phénolate provoquent un effet *bathochrome* (déplacement de la bande d'absorption vers les grandes longueurs d'onde) et *hyperchrome* (renforcement de l'intensité de l'absorption).



Le nitrophénol possède une bande d'absorption centrée à 270 nm. Il absorbe dans l'ultraviolet et il est incolore. Par addition de soude il est transformé quantitativement en sa base conjuguée : l'ion nitrophénolate qui absorbe dans le visible λ_m # 400 nm et qui possède une couleur jaune. De ce fait, le système nitrophénol - ion nitrophénolate est utilisé comme indicateur coloré acido-basique.

Propriétés chimiques :

Les molécules aromatiques ont toutes un potentiel anti-microbien et anti-toxique. Elles normalisent le terrain humoral et sont vitalisantes. Elles favorisent le drainage ; toutefois, elles ont un rôle faible en tant que draineur. Il vaut mieux dans ces cas-là utiliser les plantes sous une autre forme : en plante fraîche, en teinture-mère ou en extrait fluide.

Les phénols ont les propriétés suivantes : toniques et stimulantes : ces molécules aromatiques soutiennent le travail du corps, et lui donnent les moyens d'être plus opérationnel ; elles le nourrissent.

Elles sont anti-infectieuses à large champ d'action, elles sont capables de tuer tous les germes. Immuno-stimulantes, elles relèvent les globulines. Hyperthermisantes, elles font monter la chaleur du corps.

Elles sont aussi antispasmodiques, ainsi l'eugénol que l'on trouve dans le clou de girofle.

Oxydation

L'oxydation du phénol peut avoir lieu sous l'action de très nombreux oxydants : Fe^{3+}, O_2, etc. symbolisés par [O]. Elle conduit à la formation de radicaux phénoxyles relativement stables, qui évoluent pour donner par couplage des produits complexes souvent colorés, dont la structure est mal définie. C'est la raison pour laquelle les récipients contenant du phénol doivent être soigneusement conservés à l'abri de l'air.

Propriétés acido-basiques

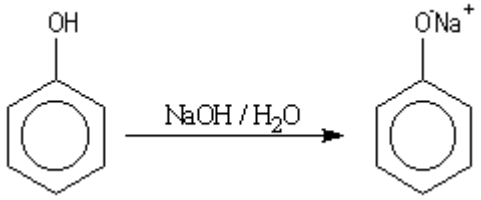
Les phénols sont plus acides que les alcools. En effet, un ion phénolate est stabilisé par résonance et est plus stable qu'un ion alcoolate. En effet, lors de la prise du proton groupement hydroxyle, le doublet électronique est partagé sur quatre carbones; ainsi, la charge est délocalisée sur autant de carbones et l'ion est beaucoup plus stable que sur un alcool où la charge négative serait trop importante et s'approprierait le proton laissé immédiatement après. Cet acide est toutefois un acide relativement faible; en conséquence, sa base conjuguée, l'ion phénolate, est une base très forte.

Acidité

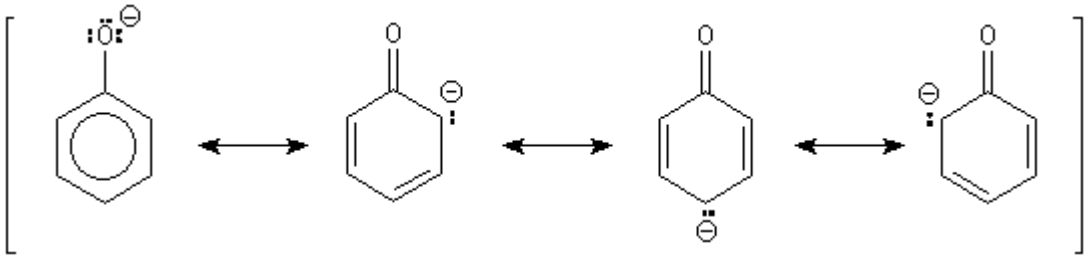
Comparons les pK_a des couples phénol/phénolate et cyclohexanol/cyclohexanolate :

Couple	**$PhOH/PhO^-$**	**$CyOH/CyO^-$**
pK_a	9,9	18

Le phénol est donc environ cent million de fois plus acide que le cyclohexanol. Il est déprotoné de façon quantitative par la soude (pK OH^-/H_2O = 14) pour donner une solution de phénolate de sodium.

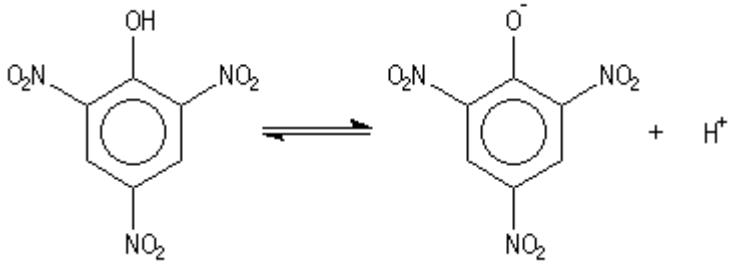


D'une façon générale, les phénols sont beaucoup plus acides que les alcools. La charge négative dispersée dans le cycle, est mieux supportée par la structure et la stabilisation qui en résulte est à l'origine de la diminution de la basicité. On peut rendre compte de cette propriété en écrivant les formes mésomères suivantes :



Les charges négatives apparaissent sur les atomes de carbone en position ortho et para. L'acidité est accrue par la présence de groupes attracteurs sur le cycle.

Le 2, 4, 6-trinitrophénol est un acide quasiment fort pour lequel le pK_a du couple vaut 0,8. Son nom d'acide picrique témoigne de cette propriété.



Classiquement, on interprète l'accroissement de stabilité de la base conjuguée par la résonance du doublet non liant de l'oxygène avec le cycle aromatique substitué par les groupes nitro attracteurs inductifs et mésomères.

Puisqu'il s'agit d'acidité relative à un solvant donné en l'occurrence l'eau et non d'acidité en phase gazeuse, il faut faire attention que la solvatation joue ici un rôle très important et il faut tenir compte de la solvatation différente de l'acide et de sa base conjuguée.

Basicité

Les phénols sont des bases beaucoup plus faibles que les alcools : pK_a ($PhO^+H_2/PhOH$) = -7
On peut l'interpréter par une protonation de l'oxygène beaucoup plus difficile que chez les alcools du fait de la délocalisation du doublet.

Estérification

mais il faut Chlorure ou Anhydride d'acide.

$$Ar\text{-}OH + R\text{-}COCl \longrightarrow R\text{-}CO\text{-}OAr + HCl$$

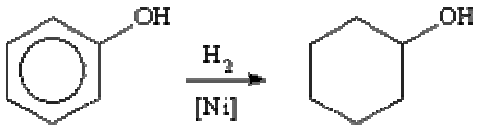
Déshydratation

ne donne que des éthers à haute température sur catalyseur

$$2\ Ar\text{-}OH + ThO_2 \text{ à } 400°C \longrightarrow Ar\text{-}O\text{-}Ar + H_2O$$

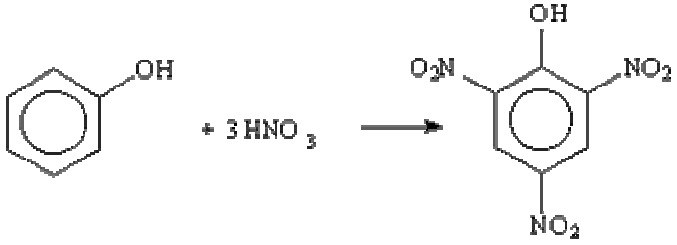
Réactions du cycle.

Addition H_2 sur catalyseur donne un cyclohexanol



Substitutions aromatiques électophiles

OH oriente en o et p et active le cycle acide picrique ou trinitrophénol, réaction pratiquement à froid.



Les phénols sont des aromatiques activés qui réagissent même avec des électrophiles faibles, notament les diazonium .

Polymérisation avec le Formol (méthanal).Conduit à la BAKELITE un des premiers "plastique" qui fut très largement utilisé comme isolant électrique.

Quelques phénols importants

Carvacrol : Eucalyptus à fleurs multiples à cryptone (Encalyptus polybractea cryptonifera)
Chavicol : Cannelier de Chine (Cinnamomum cassia)
Eugénol : Cannelle de Ceylan (Cinnamomum verum)
Thymol : Ciste ladanifère à pinène (Cistus ladaniferus pineniferum

Catéchol

Le catéchol (pyrocatéchol) est le 1,2-dihydroxybenzène. Il peut être obtenu à partir de l'orthochlorophénol par la réaction de fusion alcaline. On l'utilise comme antioxygène car il inhibe les réactions en chaîne d'oxydation en captant les radicaux. De même, il empêche la polymérisation spontanée de certains composés éthyléniques comme le styrène. On l'élimine de ce dernier en ajoutant de la soude au mélange. Le catéchol est déprotoné en milieu basique et les ions passent en phase aqueuse. On sépare le styrène et la phase aqueuse par décantation.

Résorcinol

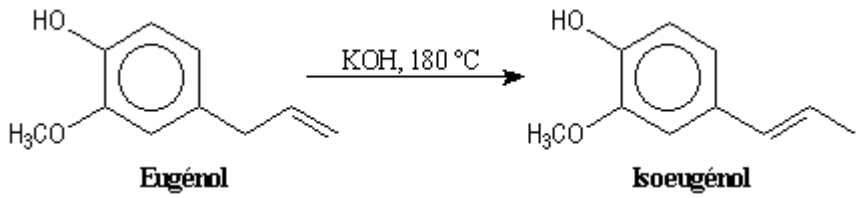
Le résorcinol est le 1,3-dihydroxybenzène. Son oxydation est beaucoup plus difficile que celle de ses deux isomères, le catéchol et l'hydroquinone car il n'existe pas de quinone correspondante. L'oxydation du résorcinol conduit à CO_2 et H_2O.

Sa réaction avec l'anhydride phtalique fournit la fluorescéine qui est un colorant.

Eugénol

L'eugénol est présent à l'état naturel dans le clou de girofle auquel il donne son goût et son odeur caractéristique. Il est utilisé comme antiseptique par les chirurgiens dentistes.

Par action de l'hydroxyde de potassium à chaud, l'eugénol est isomérisé en isoeugénol. La coupure oxydante de la chaîne latérale de l'isoeugénol conduit à la vanilline selon une réaction mise au point au siècle dernier par Reimer et Tiemann



Eugénol **Isoeugénol**

Les phénols et l'environnement

Les phénols synthétiques étant plus toxiques que ceux existant à l'état naturel, une réduction des émissions s'impose. Les personnes manipulant du phénol doivent notamment éviter le contact cutané et l'inhalation de ces produits.

Milieu aquatique*:*

Le phénol est plus lourd que l'eau et tend à se déposer. Il se dissout lentement et, même dilué, continue de former des solutions toxiques. En raison de sa forte toxicité dans l'eau, le phénol figure dans la catégorie de risque de pollution de l'eau.

Atmosphère*:*

Les vapeurs de phénol sont plus lourdes que l'air et forment des mélanges explosifs sous l'effet de la chaleur. Le phénol s'oxyde à l'air, et ce processus d'oxydation est accéléré par la lumière ou par des impuretés à effet catalytique.

Sols*:*

Dans le sol, le phénol subit une dégradation microbienne aérobie ou anaérobie, de sorte que l'effet d'accumulation reste limité. L'accumulation est fonction de la présence de minéraux argileux (forte affinité avec l'oxyde d'aluminium).

Dégradation, produits de décomposition*:*

La biodégradation des phénols naturels est en général très bonne, de sorte qu'une accumulation dans la flore ou la faune est peu probable. La dégradation par des bactéries est intégrale jusqu'à formation de dioxyde de carbone (gaz carbonique). Dans le sol, une condensation avec formation d'acide humique peut se produire. En revanche, la dégradabilité des phénols synthétiques est plus faible, car nombre d'entre eux ont une action bactéricide. Plus les phénols contiennent d'atomes de chlore ou d'azote, plus leur toxicité est forte. Ainsi, le 'pentachlorophénol' est le plus toxique des chlorophénols, et le trinitrophénol (acide picrique) le plus toxique des nitrophénols.

Les métabolites des phénols peuvent également être très toxiques: la combustion incomplète de 2,4,5-trichlorophénol peut donner naissance à la dioxine TCDD. En règle générale, la dégradation biologique entraîne d'abord la formation de pyrocatéchine, de o-quinone et d'acide dicarboxylique, puis d'acide acétique et de CO_2. Dans l'organisme humain, le phénol est éliminé par voie urinaire après oxydation ou liaison conjuguée avec l'acide sulfurique ou l'acide gluconique.

Les vapeurs et solutions de phénol sont toxiques et pénètrent aisément dans l'organisme par voie cutanée. L'inhalation de vapeurs a un effet caustique sur les voies respiratoires et les poumons. Le contact cutané et oculaire avec des solutions de phénol entraîne de sévères brûlures (poison puissant pour le protoplasme). L'exposition prolongée entraîne une paralysie du système nerveux central ainsi que des atteintes rénales et pulmonaires. Cette paralysie peut finalement entraîner la mort. L'intoxication s'accompagne de symptômes tels que maux de tête, bourdonnements, vertiges, troubles gastriques et intestinaux, étourdissement, collapsus, empoisonnement, perte de conscience, respiration irrégulière, défaillance respiratoire, troubles cardiaques, et parfois convulsions. Selon HORN (1989), le phénol possède un potentiel tératogène et cancérogène. Selon le test d'Ames, le phénol n'a pas d'effets mutagènes.

Généralement, l'effet organoleptique des phénols halogénés (odeur et goût) permet d'éviter les lésions faisant suite à une ingestion par voie orale.

Végétaux:
Perturbation de la perméabilité passive; inhibition de la croissance

CHROMATOGRAPHIE A HAUTE PERFORMANCE

La chromatographie est une méthode de séparation des constituants d'un mélange même très complexe.

Il existe trois principaux types de chromatographie:

• la chromatographie en phase gazeuse (CPG)

• la chromatographie en phase liquide à haute performance (HPLC)

• la chromatographie en couche mince (CCM).

Les deux premières méthodes peuvent être assez largement décrites par des théories communes. Dans les deux cas, un fluide appelé phase mobile parcourt un tube appelé colonne. Cette colonne peut contenir des "granulés" poreux (colonne remplie) ou être recouverte à l'intérieur d'un film mince (colonne capillaire). Dans les deux cas, la colonne est appelée phase stationnaire. A l'instant initial, le mélange à séparer est injecté à l'entrée de la colonne où il se dilue dans la phase mobile qui l'entraîne à travers la colonne.

Si la phase stationnaire a été bien choisie, les constituants du mélange, appelés généralement les solutés, sont inégalement retenus lors de la traversée de la colonne.

De ce phénomène appelé rétention il résulte que les constituants du mélange injecté se déplacent tous moins vite que la phase mobile et que leurs vitesses de déplacement sont différentes. Ils sont ainsi élués de la colonne les uns après les autres et donc séparés.

Un détecteur placé à la sortie de la colonne couplé à un enregistreur permet d'obtenir un tracé appelé chromatogramme. En effet, il dirige sur un enregistreur un signal constant appelé ligne de base en présence du fluide porteur seul ; au passage de chaque soluté séparé il conduit dans le temps à l'enregistrement d'un pic.

Dans des conditions chromatographiques données, le "temps de rétention" (temps au bout duquel un composé est élué de la colonne et détecté), caractérise qualitativement une substance. L'amplitude de ces pics, ou encore l'aire limitée par ces pics et la prolongation de la ligne de base permet de mesurer la concentration de chaque soluté dans le mélange injecté.

I – 2 – 2 – La chromatographie liquide a haute performance

a) chargement b) injection

Figure 4 : les deux phases de l'injection avec une boucle

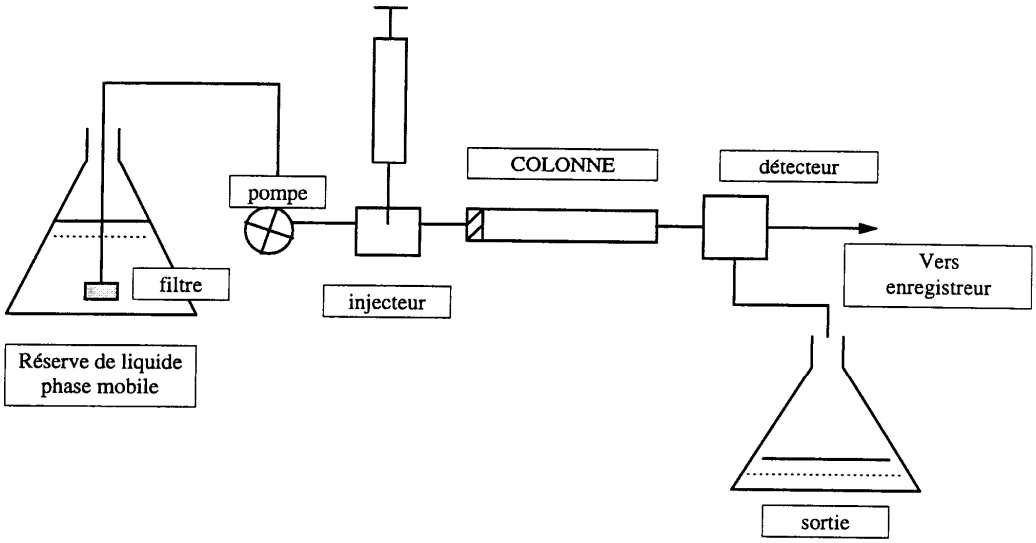


Figure 3 : principe de fonctionnement de l'HPLC

Les organes

a) Un réservoir de solvant (éluant) qui contient la phase mobile en quantité suffisante. Plusieurs flacons d'éluants (solvants de polarités différentes) sont disponibles pour pouvoir réaliser des gradients d'élution (mélange de plusieurs solvants à des concentrations variables) à l'aide de la pompe doseuse.

b) La pompe : elle est muni d'un système de gradient permettant d'effectuer une programmation de la nature du solvant. Elle permet de travailler:

- en mode isocratique, c'est-à-dire avec 100% d'un même éluant tout au long de l'analyse.

- en mode gradient, c'est-à-dire avec une variation de la concentration des constituants du mélange d'éluants.

Les pompes actuelles ont un débit variable de quelques □l à plusieurs ml/min.

c) Vanne d'injection : c'est un injecteur à boucles d'échantillonnage. Il existe des boucles de différents volumes, nous utiliserons une boucle de 20□l. Le choix du volume de la boucle se fait en fonction de la taille de la colonne et de la concentration supposée des produits à analyser. Le système de la boucle d'injection permet d'avoir un volume injecté constant, ce qui est important pour l'analyse quantitative.

d) La colonne

Une colonne est un tube construit dans un matériau le plus possible inerte aux produits chimiques, souvent en inox ou en verre. Sa section est constante, de diamètre compris entre 4 et 20 mm pour des longueurs généralement de 15 à 30 crn. Au delà, les importantes pertes de charges exigeraient des pressions de liquide beaucoup trop élevées.

e) La phase stationnaire

- La phase normale:

La phase normale est constituée de gel de silice. Ce matériau est très polaire. Il faut donc utiliser un éluant apolaire. Ainsi lors de l'injection d'une solution, les produits polaires sont retenus dans la colonne, contrairement aux produits apolaire qui sortent en tête.

L'inconvénient d'une telle phase, c'est une détérioration rapide au cours du temps du gel de silice, ce qui entraîne un manque de reproductibilité des séparations.

- La phase inverse :

La phase inverse est majoritairement composée de silice greffées par des chaînes linéaires de 8 ou 18 atomes de carbones (C8 et C18). Cette phase est apolaire et nécessite donc un éluant polaire (ACN, MeOH, H20). Dans ce cas, ce sont les composés polaires qui seront élués en premier.

Contrairement à une phase normale, il n'y a pas d'évolution de la phase stationnaire au cours du temps, et la qualité de la séparation est donc maintenue constante.

- La phase mobile :

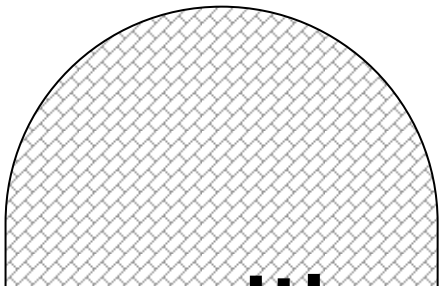
L'interaction plus ou moins forte entre la phase mobile et la phase stationnaire normale ou à polarité inversée se répercute sur les temps de rétention des solutés. La polarité de la phase stationnaire permet de distinguer deux situations de principe :

- si la phase stationnaire est polaire, on utilisera une phase mobile peu polaire la chromatographie est dite en phase normale ;

- si la phase stationnaire est très peu polaire, on choisira une phase mobile polaire (le plus souvent des mélanges de méthanol ou d'acétonitrile avec de l'eau), c'est la chromatographie en phase inverse. En modifiant la polarité de la phase mobile, on agit sur les facteurs de rétention k des composés.

Les silices greffées conduisent en général à une perte importante de polarité. Avec une **phase greffée**, l'ordre d'élution est opposé à celui auquel on est habitué avec les phase normales. Ainsi avec un éluant polaire, un composé polaire migre plus vite qu'un composé apolaire. Dans ces conditions les hydrocarbures sont fortement retenus. On réalise des gradients d'élution en diminuant au cours de la séparation la polarité de l'éluant (ex : mélange eau /acétonitrile dont la concentration en acétonitrile va en croissant au cours de l'élution).

On peut, en mélangeant plusieurs solvants, ajuster le pouvoir d'élution de la phase mobile [1],



I – RELATIONS EXTRATHERMODYNAMIQUES

I – 1 Généralités

Les premières études portant sur les relations structure / rétention chromatographique de différents solutés remontent à 1949 lorsque Martin, dans son article fondamental [2], suggéra qu'un substituant modifie le coefficient de partage d'un soluté par un facteur qui dépend de la nature de ce substituant et, simultanément, des phases stationnaire et mobile, mais non du reste de la molécule. Depuis pratiquement la naissance de la chromatographie on a pu observer des relations simples entre les paramètres de rétention et, par exemple, le nombre de carbones d'une série homologue [3].

A la suite de Green et al. [4], qui trouvèrent que les incréments portant sur le paramètre de rétention : $R_M = \log\left(\frac{1}{R_f} - 1\right)$, en chromatographie sur couche mince, associés aux substituants d'un certain nombre de composés benzénoïdes plans étaient additifs, Iwasa et al. [5] suggèrent en 1965 d'utiliser les données chromatographiques dans les études de relations structure / activité biologique. On a largement utilisé, depuis, la chromatographie pour la quantification de l'hydrophobicité d'agents bioactifs [6-9].

En 1977 parurent les premières publications [10-12] dans lesquelles la méthodologie QSAR (pour : Quantitative Structure – Activity Relationships) était appliquée pour l'analyse des données de rétention chromatographique obtenues pour diverses séries de composés. Par la suite, le nombre de corrélations structure / rétention augmenta de façon exponentielle avec le temps, par suite de la disponibilité générale des micro-computers et de programmes de calculs statistiques appropriés. Par analogie avec Quantitative Structure Activity Relationships (QSAR) le terme Quantitative Structure Retention Relationships (QSRR) fut proposé [13] pour englober ce nouveau domaine de la chromatographie.

I – 2 Relations de type extrathermodynamique

Les conditions chromatographiques peuvent être modifiées en changeant la phase stationnaire et / ou mobile. Les variations des rétentions relatives qui s'ensuivent pour les composés étudiés, peuvent renseigner sur la capacité de ces composés d'entreprendre divers types d'interactions moléculaires sur une base thermodynamique [14]. Une autre approche consiste soit à sélectionner un groupe convenable de composés tests pour lesquels les données chromatographiques seront déterminées pour des conditions constantes, soit à ramener les données chromatographiques obtenues à des conditions standards normalisées.

Dans cette seconde approche les différences obtenues pour les données chromatographiques reflètent celles observées dans la structure des solutés. Les relations entre les données de

rétention chromatographiques et les grandeurs associées à la structure des solutés ne peuvent être expliquées sur une base thermodynamique stricte. De telles relations sont dites de type extrathermodynamique.

Le terme extrathermodynamique signifie que la science se situe en dehors de la structure formelle de la thermodynamique, quoique l'approche ressemble à celle de la thermodynamique en ce sens qu'il n'est pas nécessaire d'expliciter les mécanismes microscopiques lors des son utilisation [15]. Les approches extrathermodynamiques sont des combinaisons de modèles détaillés avec les concepts de la thermodynamique. Comme elle implique la construction de modèles, cette sorte d'approche n'a pas la rigueur de la thermodynamique, mais peut fournir des informations inaccessibles autrement. Les relations linéaires d'enthalpie libre (LFER, pour Linear Free Energy Relationships) constituent des manifestations des relations extrathermodynamiques. Quoique les LFER ne soient pas une conséquence nécessaire de la thermodynamique, leur existence suggère un lien entre les grandeurs corrélées, et la nature de ce lien peut être étudiée [16].

I – 3 Application à la chromatographie

Les paramètres de rétention chromatographiques utilisés dans les études de corrélations sont supposés proportionnels aux variations de l'enthalpie libre associée au processus de partage chromatographique. Cependant, toutes les données ne conviennent pas nécessairement pour les études QSRR. La relation de Gibbs :

$$\Delta G = \Delta H - T\,\Delta S \qquad (1)$$

relie les variations de l'enthalpie libre, ΔG, à celles de l'enthalpie, ΔH, et de l'entropie, ΔS ; T est la température absolue.

L'existence de relations extrathermodynamiques entre le système réel et le modèle adopté, suppose une variation constante soit de l'entropie, soit de l'enthalpie, ou que leurs variations soient reliées linéairement [17].

$$\Delta H = T_c\,\Delta S + \Delta G_{T_C} \qquad (\text{pour } T = T_c) \qquad (2)$$

Lorsque la compensation enthalpie – entropie est établie pour une famille de composés, dans une transformation chimique donnée, les valeurs de T_C et ΔG sont invariantes, et T_c est appelée "température de compensation".

En utilisant la relation (1), on peut réécrire (2) de façon à exprimer la variation de l'enthalpie libre ΔG_T, mesurée à une température T, pour le processus d'équilibre, sous la forme :

$$\Delta G_T = \Delta H\left(1 - \frac{T}{T_C}\right) + \frac{T\,\Delta G_{T_C}}{T_C} \qquad (3)$$

En chromatographie liquide, le paramètre de rétention k', ou facteur de capacité, est relié à la constante d'équilibre thermodynamique K de liaison du soluté, selon :

$$k' = K\,\Phi \qquad (4)$$

où Φ représente le rapport des phases de la colonne. La variation d'enthalpie libre pour le processus chromatographique est exprimée par :

$$\Delta G = -R\,T\,\mathrm{Ln}\,K = -R\,T\,\mathrm{Ln}\left(k'/\Phi\right) \qquad (5)$$

En portant (5) dans (1), on obtient pour le paramètre de rétention k' :

$$\mathrm{Ln}\,k' = -\frac{\Delta H}{R\,T} + \frac{\Delta S}{R} + \mathrm{Ln}\,\Phi \qquad (6)$$

Si, pour la plage de température considérée, le mécanisme du processus est invariant et l'enthalpie constante, alors la variation de Ln k' en fonction de 1/T est linéaire, et la pente de la droite permettra d'attribuer la variation d'enthalpie d'un soluté donné.

La combinaison des équations (3) et (5) fournit :

$$\mathrm{Ln}\,k'_T = -\frac{\Delta H}{R}\left(\frac{1}{T} - \frac{1}{T_c}\right) - \frac{\Delta G_{T_C}}{R\,T_c} + \mathrm{Ln}\,\Phi \qquad (7)$$

en désignant par k'_T le paramètre de rétention à la température T.

Cette dernière équation montre que, pour divers solutés, le k' mesuré à une température T donnée pour différentes conditions, varie linéairement en fonction de ΔH lorsque la compensation enthalpie – entropie a lieu. Dans ce cas, la rétention réversible des solutés par la phase stationnaire met essentiellement en jeu le même mécanisme.

Notons que la pente de la courbe de compensation ($\mathrm{Ln}\,k'_T$ en fonction de ΔH ; eq. (7)) permet d'atteindre la température de compensation T_c.

II – PHENOLS

Les phénols sont des composés importants biologiquement et du point de vue de l'environnement. Non seulement ils possèdent d'importantes fonctions physiologiques et certaines activités pharmaceutiques, mais ils peuvent encore influencer la saveur de certaines boissons. D'autres composés phénoliques sont utilisés pour le tannage, en cosmétique, dans l'industrie organique (fabrication de matières plastiques, produits pharmaceutiques, explosifs ….), ainsi que pour le développement photo, ce qui en fait d'importants polluants potentiels de l'environnement [18,19].

Les méthodes chromatographiques sont souvent utilisées pour leur analyse.

La rétention chromatographique est le résultat d'un processus de distribution concurrentiel du soluté entre les phases stationnaire et mobile, dans lequel le partage entre ces 2 phases est largement déterminé par la structure moléculaire. En se basant sur cette approche, différents auteurs [9, 19-21] se sont attachés à décrire des modèles de régression linéaire ou multilinéaire pour la prévision de la rétention chromatographique, en utilisant différents types de descripteurs moléculaires (structural, topologique, électronique, géométrique …. [22,23]), ainsi que des propriétés physico-chimiques [19, 24, 25]). On peut noter également les nombreuses tentatives pour relier quantitativement les paramètres de rétention (ou facteurs de capacité) k' à la composition de la phase mobile, particulièrement pour ce qui concerne la chromatographie liquide haute performance à polarité de phase inversée (CLHP-PI) mettant en jeu des phases mobiles hydro-organiques binaires [17, 26-37].

Dans cette partie nous nous intéresserons à quelques phénols diversement substitués, analysés par CLHP-PI avec une phase mobile méthanol – eau. Cet éluant étant couramment utilisé en CLHP – PI, la disponibilité, dans ce cas, d'un modèle de rétention du soluté est importante pour l'optimisation chromatographique.

Nous évaluerons les performances de deux modèles, choisis parmi les nombreux modèles de la littérature, lorsqu'ils sont appliqués à 10 phénols non congénères.

III – LES MODELES

Le facteur de capacité k' des phénols sera, pratiquement, calculé à partir de la relation :

$$k' = \frac{t_R - t_0}{t_0} = \frac{t_R}{t_0} - 1 = \frac{1}{R_F} - 1 \qquad (8)$$

t_R étant le temps de rétention du soluté et t_0 le temps de rétention nulle ; R_F étant égal à t_0/t_R.

Le modèle proposé par Kowalska [29] prend en compte, en les quantifiant, les phénomènes d'associations entre des molécules de la phase mobile (auto-association et association mixte). Il permet, pour une température donnée, de relier simplement R_F à la composition volumique (méthanol / eau) de la phase mobile.

D'après cette auteure, l'équation suivante serait valide sur une large gamme de fractions volumiquesx_1 du méthanol (x_2 étant la fraction volumique de l'eau) :

$$R_F = A\sqrt{x_1} + B\sqrt{x_2} + C \qquad (9)$$

quand on l'applique à des isomères de composés aromatiques di-hydroxylés en nombre restreint.

Nous la testerons avec les dérivés phénoliques étudiés qui, s'ils sont également limités en nombre, n'en présentent pas moins divers types de substituants.

Ce modèle ne considérant pas l'influence de la température nous avons, à titre de comparaison, appliqué le modèle proposé par Horváth et al. [28] qui semble plus général puisqu'il fait intervenir la fraction volumique x du co-solvant organique, la température absolue T de la colonne, ainsi d'ailleurs que la température de compensation T_c.

La relation analytique entre log k' et ces différents paramètres s'exprime [28] par :

$$\log k' = A_1 x \left(1 - \frac{T_c}{T_0}\right) + \frac{A_2}{T} + A_3 \qquad (10)$$

Les coefficients A_i (i = 1 à 3) sont obtenus par régression linéaire de log k' avec les fonctions appropriées de la composition de la phase mobile et de la température.

Rappelons que la détermination de la température T_c suppose une compensation enthalpie-entropie, qui prouve la constance du mécanisme du processus chromatographique dans le domaine de température considéré. Cette condition est vérifiée lorsque la variation de $\mathrm{Ln}\, k'_T$ en fonction de ΔH (kcal.mol^{-1}) est linéaire, la variation d'enthalpie, pour l'intervalle de la température d'étude, étant obtenue à partir de la pente de la droite d'équation (6) :

$$\text{Ln } k' = -\frac{\Delta H}{R\,T} + \frac{\Delta S}{R} + \text{Ln } \Phi \qquad (6)$$

IV – MATERIEL ET METHODE

Le chromatographe utilisé est un système Philips (Pye Unicam) constitué d'une pompe PU 4010, d'un injecteur Rhéodyne 7125, d'un détecteur UV / visible à longueur d'onde variable PU 4200 et d'un enregistreur PM 8251. Une boucle de 20 µL est remplie avec une seringue Hamilton de 25 µL.

Une colonne de remplissage en inox (L : 25 cm ; diam. Int. = 4,6 mm) contenant des radicaux octadécyl-silyle (partisil ODS), dont le diamètre des particules support est 10 µm, a été utilisée pour les analyses. La colonne est placée dans une enveloppe en verre où circule de l'eau thermostatée, ce qui permet de fixer la température désirée à 0,1 °C près.

Les analyses ont été réalisées en régime isochratique avec des phases mobiles constituées de mélanges (méthanol + eau) dans les rapports (V :V) 15 : 85, 25 : 75, 50 : 50, 70 : 30 et 85 : 15, pour un débit réglé à 2 mL / min.

L'effet de la température sur la rétention a été examiné pour l'intervalle 12 – 52 °C (285 – 325 K) en faisant varier la température par bonds de 10 °, la phase mobile étant un mélange volume à volume méthanol – eau.

Nous avons pris pour t_0, le temps de rétention de ^2H_2O qui absorbe à 190 nm.

Les structures des 10 phénols considérés sont représentées dans la figure 1.

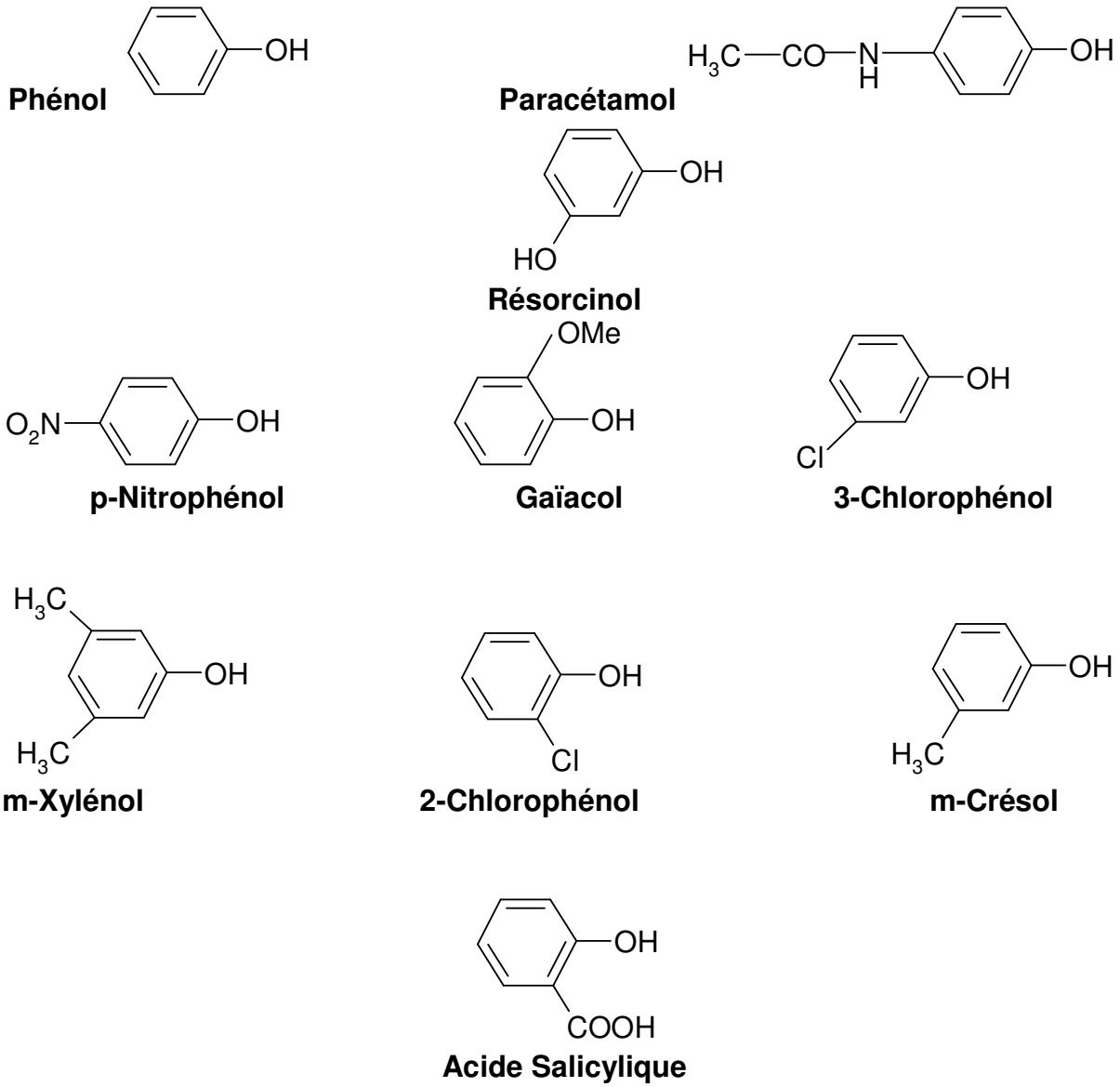


Figure 1 – *Noms et structures des phénols étudiés.*

V – RESULTATS ET DISCUSSION

Les valeurs de k' mesurées à 22 °C pour différents ratios volumiques des constituants de la phase mobile (Tableau I) permettent de déduire les valeurs correspondantes de R_F (éq. 8).

Tableau I – *Valeurs de k' mesurées en fonction de la température (t) et de la fraction volumique (x) du méthanol.*

	t = 22 °C ; x(méthanol)					x = 0,50 ; t (°C)			
	0,15	**0,25**	**0,50**	**0,70**	**0,85**	**12**	**32**	**42**	**52**
m- Xylénol	-	7,780	2,266	1,036	0,529	2,556	1,939	1,595	1,451
3- Chlorophénol	9,920	6,551	2,228	1,012*	0,543	1,429	1,789	1,465*	1,319
m- Crésol	6,907	4,644	1,524	2,216	0,795*	1,714	1,351	1,156	1,063
2- Chlorophénol	6,633	5,303	1,642*	0,819	0,595	2,167	1,448	1,220	1,116
Gaïacol	5,604	3,407	1,243*	0,739	0,591	1,346	1,105*	0,985	0,914
Phénol	3,193*	2,410*	1,109	0,695	0,551	1,394*	0,998	0,875	0,826
Paracétamol	2,640	1,609	0,764	0,564	0,509	0,848*	0,677	0,614	0,579
Résorcinol	1,688*	1,262	0,717*	0,507	0,480	0,781	0,637	0,586	0,544
p- Nitrophénol	3,623	3,611	1,363	0,590	0,417	1,858	1,225	0,994	0,882
Acide Salicylique	0,285	0,275	0,119	0,109	0,160	0,145	0,101	0,0883	0,066

***** Valeurs arbitrairement choisies pour le contrôle des performances du modèle de Horváth.

En portant les valeurs ainsi trouvées dans (9) on obtient, pour chaque soluté, un système linéaire en A, B, C, surdéterminé, que l'on traite par la méthode des moindres carrés.

Les jeux respectifs des constantes (A, B, C) de l'équation (9) sont réunis dans le tableau II.

Tableau II – *Jeux de constantes A, B, C (éq. (9)) pour les 10 phénols.*

	A	B	C
m- Xylénol	0,943	0,269	- 0,146
3- Chlorophénol	0,230	-0,764	0,722
m- Crésol	0,322	- 0,638	0,604
2- Chlorophénol	0,245	- 0,726	0,707
Gaïacol	0,322	- 0,638	0,604
Phénol	0,450	- 0,320	0,371
Paracétamol	1,055	- 0,320	0,371
Résorcinol	0,674	0,132	$- 0{,}630 \cdot 10^{-3}$
p- Nitrophénol	0,259	- 0,755	0,781
Acide Salicylique	0,321	$1{,}396 \cdot 10^{-2}$	0,633

Ils permettent d'obtenir les valeurs calculées de k', soit k'_c, lesquelles sont comparées dans la figure (2), aux valeurs mesurées.

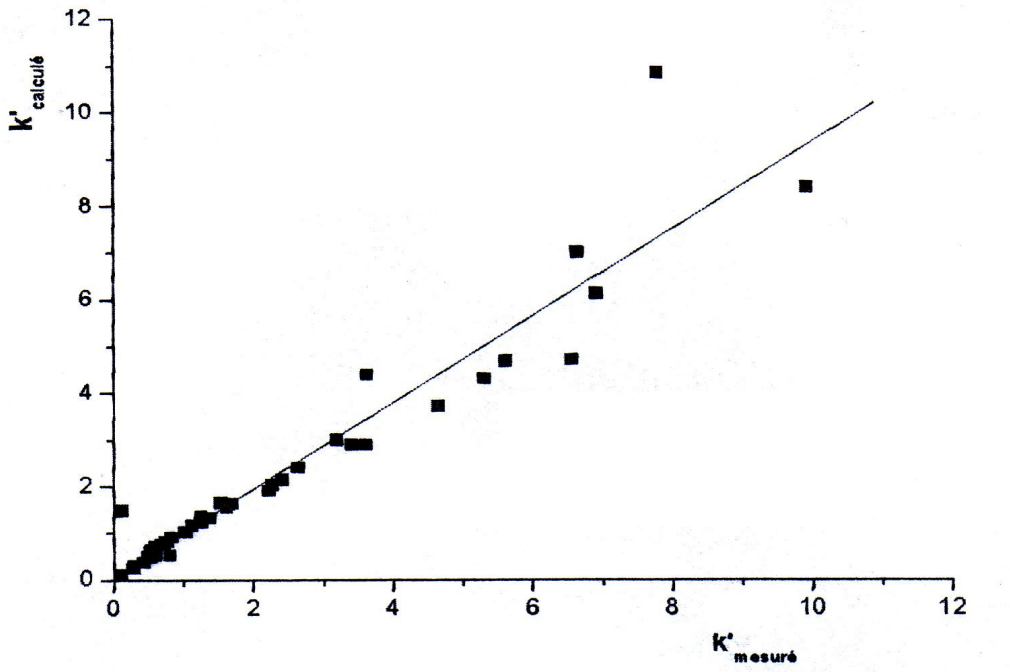


Figure 2 – *Corrélation entre les k' mesurés et calculés par le modèle de Kowalska.*

Ainsi, les valeurs estimées par le modèle de Kowalska reproduisent celles mesurées au niveau de 10 %. Cette erreur relative ne permet pas de reproduire correctement l'ordre d'élution expérimental. Plusieurs inversions ont été relevées, comme il est précisé ci-après :

x_{CH_3OH}	**0,15**	**0,25**	**0,50**	**0,70**	**0,85**
Inversion (s)	2- Chlorophénol 3- Chlorophénol	aucune	p- Nitrophénol Gaïacol	2- Chlorophénol 3- Chlorophénol m- Xylénol	aucune

Pour chaque phénol, nous avons noté une variation (quasi) linéaire de Ln k_T en fonction de 1/T. Le traitement des données par la méthode des moindres carrés permet de déterminer la pente $\left(b = -\frac{\Delta H}{R}\right)$ de la droite d'équation (6), et par conséquent ΔH. Les valeurs trouvées, ainsi que les coefficients de corrélation linéaire, r, sont rassemblés dans le tableau suivant :

Tableau III – *Pente* $\left(b = - \frac{\Delta H}{R} \right)$ *de la droite d'équation (6) et coefficients, r, de Bravais – Pearson.*

	b = - ΔH/R	r	ΔH (kcal.mol^{-1})
m- Xylénol	1408,0	0,9960	- 2,780
3- Chlorophénol	1499,0	0,9975	- 2,978
m- Crésol	1170,0	0,9975	- 2,325
2- Chlorophénol	1575,0	0,9803	- 3,130
Gaïacol	981,1	0,9985	- 1,949
Phénol	1198,0	0,9828	- 2,380
Paracétamol	912,8	0,9965	- 1,814
Résorcinol	860,4	0,9980	- 1,710
p- Nitrophénol	1750,0	0,9823	- 3,477
Acide Salicylique	1842,0	0,9970	- 3,660

Les courbes représentant, pour les 10 phénols, Ln k'$_T$ mesuré à une température déterminée T et pour différentes conditions, en fonction de la variation d'enthalpie correspondante ne sont pas linéaires (Figurer 3-a). La compensation enthalpie – entropie apparaît seulement après élimination du p- nitrophénol et de l'acide salicylique (Figure 3-b). Ces 2 composés ne sont plus considérés dans la suite de ce travail.

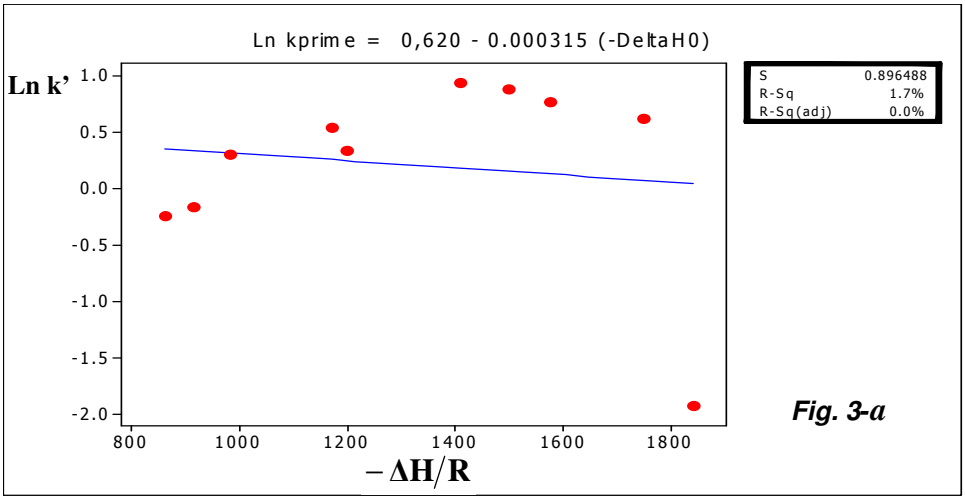


Fig. 3-a

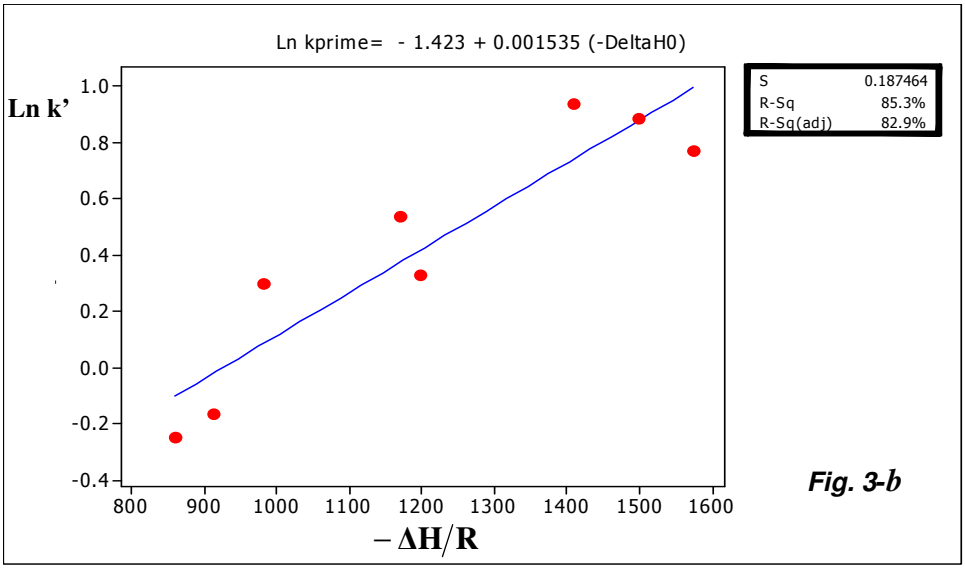


Fig. 3-b

Figure 3 – *Variation de Ln k' à 305 K en fonction de* $b = -\frac{\Delta H}{R}$.

***a/** pour les phénols étudiés ;*
***b/** après élimination du p- nitrophénol et de l'acide salicylique.*

Pour éviter tout effet statistique indésirable, nous avons pris pour température de référence T, dans l'équation (7), la température proche de la moyenne harmonique des températures expérimentales utilisées pour l'évaluation des enthalpies à partir de l'équation (6), soit : T = 305 K.

Rappelons que la moyenne harmonique $\overline{x}_h$ d'une série statistique de n valeurs positives x_1, ….. , x_n est égale à l'inverse de la moyenne arithmétique des inverses :

$$\overline{x}_h = \frac{n}{\sum_{i=1}^{n} (1/x_i)} \qquad (11)$$

Nous avons pu ainsi déterminer une température de compensation proche de 500 K ; nous la supposerons constante et égale à cette valeur (T_c = 500 K).

Le groupe des 71 observations correspondant aux 8 premiers phénols du tableau II a été séparé en deux ensembles : un ensemble d'estimation (59 éléments) et un ensemble d'essai (12 éléments étoilés). L'ensemble d'estimation a été utilisé pour l'obtention des coefficients A_1 – A_3, par régression linéaire de log k' avec les fonctions appropriées de la composition de la phase mobile et de la température. Ces coefficients rassemblés dans le Tableau IV, ont d'abord été utilisés pour calculer les 59 valeurs de k' fournies par le modèle. La figure 4 compare les k' calculés aux k' mesurés.

Les équations de régression et les paramètres statistiques obtenus dans cas se présentent comme suit :

$$\mathbf{k'_{cal} = 0{,}134 + 0{,}902\, k'_{mes}} \qquad (12)$$

n = 59 ; S = 0,2339 ; r = 0,993

Tableau IV – *Coefficients A_i (i = 1 à 3) de l'équation (9) pour l'ensemble d'estimation des phénols.*

A_i	A_1	A_2	A_3
m- Xylénol	3,220202	1528,452	- 3,839338
3- CP	2,983042	1477,537	- 3,776043
m- Crésol	2,612924	1293,236	- 3,384252
2- CP	2,625580	1376,006	- 3,626678
Gaïacol	2,348141	1218,916	- 3,284247
Phénol	1,855347	1027,972	- 2,855112
Paracétamol	1,685203	1040,243	- 3,081672
Résorcinol	1,338046	818,0536	- 2,493453

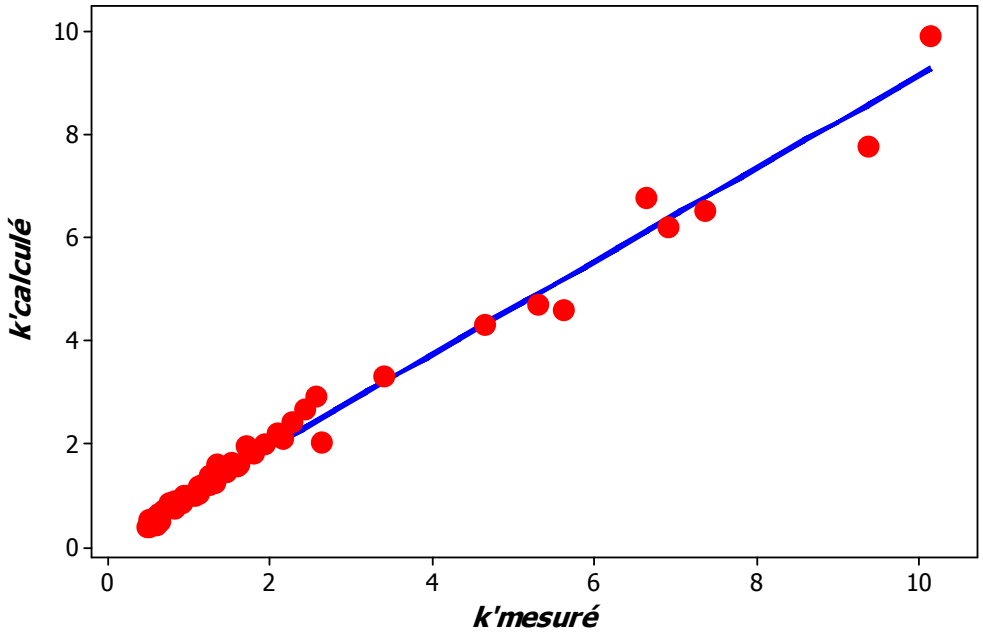


Figure 4 – *k' des 59 données d'estimation calculées par le modèle de Horváth en fonction des k' mesurés.*

Nous avons ensuite utilisé les coefficients A_1 – A_3 pour la validation du modèle sur l'ensemble d'essai.

La droite de régression de k'_{cal} en k'_{mes} pour les 12 observations d'essai apparaît sur la figure 5 ; son équation et les paramètres statistiques y afférents sont donnés ci après :

$$k'_{cal} = 0{,}1579 + 0{,}912\, k'_{mes} \qquad (13)$$

n = 12 ; S = 0,107 ; r = 0,988

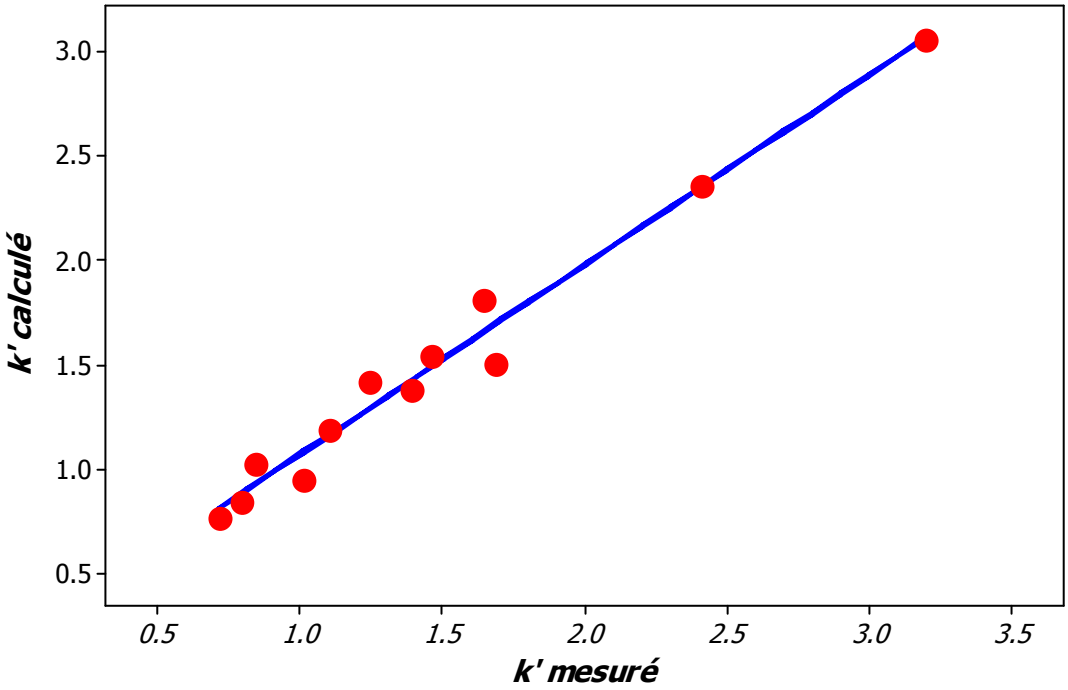


Figure 5 – *k' des 12 données d'essai calculés d'après le modèle de Horváth en fonction des k' mesurés.*

VI – CONCLUSION

L'optimisation des conditions de séparation en CLHP à polarité de phase inversée nécessite l'utilisation de modèles. La littérature en fait ressortir plusieurs, basés sur des relations fonctionnelles différentes. Ces modèles sont souvent évalués selon le critère statistique d'ajustement de la rétention d'un seul soluté pour différentes compositions de la phase mobile.

Une telle approche ne permet pas de s'assurer de la justesse des hypothèses de base des modèles de rétention.

Les performances de 2 modèles de la littérature ont été reliées à la qualité de l'ajustement des données de rétention d'un ensemble de phénols non congénères, obtenues en faisant varier les conditions de séparation (composition de la phase mobile ; température).

Le modèle de Kowalska, qui semble facile d'utilisation, est basé sur les associations possibles des molécules de la phase mobile qui peuvent être modifiées selon la température de travail dont il n'est pas tenu compte dans le modèle. Les mauvais critères statistiques (coefficient de détermination ; erreur standard …) peuvent refléter un écart sensible aux hypothèses de base du modèle.

L'élimination de 2 phénols a permis de justifier les hypothèses du modèle d'Horváth qui tient compte de la température des expérimentations, et fait intervenir une température dite de compensation déterminée expérimentalement. L'amélioration des critères statistiques de base n'est pas pour autant un gage d'une bonne capacité prédictive du modèle.

APPROCHE QSAR

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

Les structures des molécules ont été obtenues à l'aide du logiciel de modélisation moléculaire Hyperchem 7.5 [38], et les géométries finales à l'aide de la méthode semi empirique AM1 du même logiciel. Tous les calculs ont été menés dans le cadre du formalisme RHF sans interaction de configuration. Les structures moléculaires ont été optimisées à l'aide de l'algorithme Polak-Ribiere avec pour critère une racine du carré moyen du gradient égale à 0,001 kcal.mol^{-1}. Les géométries obtenues ont été transférées dans les logiciels informatiques [38-39] utilisés pour le calcul de plus de 1700 descripteurs appartenant à 20 classes différentes.

Le logiciel Ecalc [40] est muni d'une interface graphique qui permet à l'utilisateur d'introduire les molécules, puis de calculer les indices électrotopologiques.

II- SELECTION D'UN SOUS-ENSEMBLE DE DESCRIPTEURS SIGNIFICATIFS

Des logiciels informatiques spécialisés permettent le calcul de plus de 3000 descripteurs moléculaires appartenants 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 réduit de variables explicatives, on peut citer : les méthodes de pas à pas (méthode descendante; méthode ascendante, et méthode dite stepwise), ainsi que les algorithmes évolutifs et génétiques.

En général, la comparaison se fait à l'avantage des algorithmes génétiques (GA) que nous avons appliqués dans le présent travail, et que nous rappelons succinctement

II-1 Principe

Dans la terminologie des algorithmes génétiques, le vecteur binaire $\underset{\sim}{I}$, appelé "chromosome", est un vecteur de dimension p où chaque position (un "gène") correspond à une valeur (1 si elle figure dans le modèle, 0 sinon). Chaque chromosome représente un modèle basé sur un ensemble de variables explicatives.

On commence par définir le paramètre statistique à optimiser (par exemple maximiser Q^2 en utilisant la validation croisée par "leave-one-out" ; cf. infra), avec la taille P de la population du modèle (par exemple, P = 100), et le nombre maximum de variables L permises pour le modèle (par exemple, L = 10) ; le minimum de variables permises est généralement supposé égal à 1. De plus, une probabilité de croisement p_c (habituellement élevée, $p_c > 0,9$),

et une probabilité de mutation p_M (habituellement faible, $p_M < 0,1$) doivent être également définies.

Après définition des principaux paramètres, la mise en œuvre de l'algorithme génétique est démarrée, son évolution comprend trois étapes principales.

II – 2 Initialisation aléatoire du modèle

La population est constituée au départ de modèles aléatoires avec des variables comprises entre 1 et L, puis les modèles sont ordonnés eu égard au paramètre statistique sélectionné – la qualité du modèle – (le meilleur modèle est en première position, le plus mauvais en position P) ;

II – 3 Etape de croisement

A partir de la population, on sélectionne des paires de modèles (aléatoirement, ou avec une probabilité proportionnelle à leur qualité). Puis, pour chaque paire de modèles on conserve les caractéristiques communes, c'est-à-dire les variables exclues dans les 2 modèles restent exclues, et les variables intégrées dans les 2 modèles sont conservées. Pour les variables sélectionnées dans un modèle et éliminées dans l'autre, on en essaye un certain nombre au hasard que l'on compare à la probabilité de croisement p_c : si le nombre aléatoire est inférieur à la probabilité de croisement, la variable exclue est intégrée au modèle et vice versa. Finalement, le paramètre statistique du nouveau modèle est calculé : si la valeur de ce paramètre est meilleure que la plus mauvaise pour la population, le modèle est intégré à la population, à la place correspondant à son rang ; dans la cas contraire, il n'est pas pris en compte. Cette procédure est répétée pour de nombreuses paires (100 fois par exemple).

II – 4 Etape de mutation

Pour chaque modèle de la population (c'est-à-dire pour chaque chromosome) p nombres aléatoires sont éprouvés, et, un à la fois, chacun est comparé à la probabilité de mutation, p_M, définie : chaque gène demeure inchangé si le nombre aléatoire correspondant excède la probabilité de mutation, dans le cas contraire, on le change de 0 à 1 ou vice versa. Les faibles valeurs de p_M permettent uniquement peu de mutations, conduisant à de nouveaux chromosomes peu différents des chromosomes générateurs.

Après obtention du modèle transformé, on en calcule le paramètre statistique : si cette valeur est meilleure que la plus mauvaise de la population, le modèle est intégré à la population, à la place correspondant à son rang ; dans le cas contraire il n'est pas pris en compte.

Cette procédure est répétée pour tous les chromosomes, c'est-à-dire P fois.

II – 5 Conditions d'arrêt

Les étapes 2 et 3 sont répétées jusqu'à la rencontre d'une condition d'arrêt (par exemple un nombre maximum d'itérations défini par l'utilisateur), ou qu'il est mis fin arbitrairement au processus.

Une caractéristique importante de la sélection d'un ensemble réduit de variables par algorithme génétique est qu'on n'obtient pas nécessairement un modèle unique, mais le résultat consiste habituellement en une population de modèles acceptables ; cette caractéristique, parfois considérée comme un désavantage, fournit une opportunité pour procéder à une évaluation des relations avec la réponse à différents points de vue.

Notons que le taille du modèle est fixée par la valeur optimale de la fonction FIT de KUBINYI [41], calculée selon :

$$FIT = \frac{R^2}{1 - R^2} \frac{n - p - 1}{(n + p)^2} \qquad (1)$$

p désignant le nombre de variables du modèle et R^2 le coefficient de détermination.
Ce critère permet de comparer entre modèles construits sur un même nombre n de données, mais avec un nombre de variable p différent.

III – DEVELOPPEMENT DES MODELES

Les techniques les plus courantes pour établir des modèles QSRR utilisent l'analyse de régression (régression linéaire multiple : MLR ; projection des structures latentes par les moindres carrés partiels : PLS), les réseaux neuronaux, et les méthodes de classification.

Nous avons utilisé la MLR et les réseaux de neurones artificiels (RNA). En imposant des transformations linéaires entre descripteurs moléculaires et propriétés étudiées, la MLR peut influencer négativement les capacités prédictives du modèle. Par contre, avec les réseaux de neurones il n'est nul besoin de postuler un modèle. Les réseaux de neurones ont la capacité de représenter n'importe quelle dépendance fonctionnelle qu'ils découvrent par eux-mêmes. Ainsi, la découverte et l'exploitation des dépendances non-linéaires de haut niveau peuvent améliorer la capacité de prédiction de la variable d'intérêt.

III – 1 La régression linéaire multiple (MLR)

Supposons qu'on ait mesuré sur n individus (k+1) variables représentées par des vecteurs de $\Re^n : \underset{\sim}{y}, \underset{\sim}{x}_1, \underset{\sim}{x}_2, \ldots, \underset{\sim}{x}_k$; $\underset{\sim}{y}$ est la variable dépendante ou à expliquer (propriété physique d'intérêt) et les $\underset{\sim}{x}_j$ les variables explicatives ou encore prédicteurs (descripteurs moléculaires). On cherche alors à reconstruire $\underset{\sim}{y}$ au moyen des $\underset{\sim}{x}_j$ par une formule linéaire.

On pose :

$$\underset{\sim}{y} = \beta_0 \underset{\sim}{\mathbf{1}} + \underset{\sim}{\mathbf{X}}(j)\,\beta(j) + \varepsilon(j) \qquad (2)$$

$\underset{\sim}{y}$ est un vecteur de dimension n contenant la propriété physique d'intérêt des hydrocarbures considérés, $\underset{\sim}{1}$ est un vecteur unité, c'est-à-dire une matrice colonne formée d'éléments égaux à 1, $\underset{\sim}{X}(j)$ indique la matrice (n×j), et $\varepsilon(j)$ correspond aux résidus qui doivent suivre une distribution Normale, posséder une espérance mathématique nulle et une matrice de dispersion I σ^2 [42]. Les estimateurs {β} sont calculés en utilisant la technique des moindres carrés ordinaires.

III – 2 Les réseaux de neurones

Les réseaux de neurones ont été étudiés depuis les années 40 [43]. Les idées de base de cette technique viennent de la recherche cognitive, d'où vient le nom 'réseaux de neurones'.

La technique inspirait beaucoup de chercheurs à cette époque, mais beaucoup de l'intérêt disparaît après un article de Minsky et Papert [44], finalement relancée au début des années 80 après un quasi-oubli d'une vingtaine d'années. La cause de l'intérêt soudain était l'apparition de nouvelles architectures de réseaux de neurones.

III – 2 -1 Le neurone artificiel :

L'élément de base d'un réseau de neurones est, bien entendu, le neurone artificiel. Un neurone (figure 6) contient deux éléments principaux :

- Un ensemble de poids associés aux connexions du neurone, et
- Une fonction d'activation (Figure7).

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

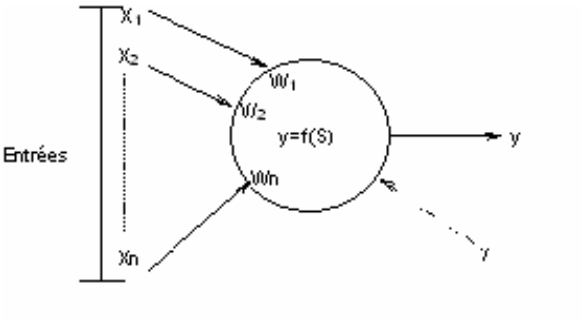
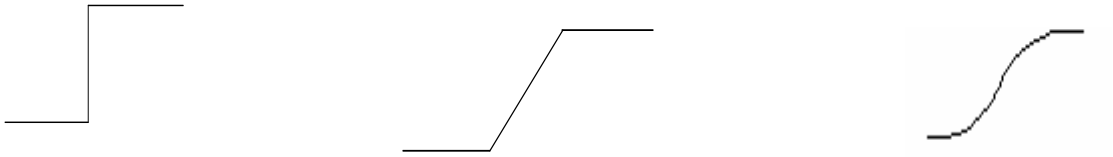


Figure – 6 *le neurone artificiel générique.*

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



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

Figure – 7 *Fonctions d'activation.*

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

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

III – 2 - 2. Propriétés des réseaux de neurones :

Un réseau de neurones se compose de neurones qui sont interconnectés de façon à ce que la sortie d'un neurone puisse être l'entrée d'un ou plusieurs autres neurones. Ensuite il y a des entrées de l'extérieur et des sorties vers l'extérieur [45].

Rumelbart et al. donnent huit composants principaux d'un réseau de neurones [45] :

- Un ensemble de neurones.
- Un état d'activation pour chaque neurone (actif, inactif,…).

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

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

Les réseaux de neurones sont souvent appelés des "boites noires" car la fonction mathématique qui est représentée devient vite trop complexe pour l'analyser et la comprendre directement. Cela est notamment le cas si le réseau développe des représentations distribuées [45], c'est-à-dire que plusieurs neurones sont plus ou moins actifs et contribuent à une décision. Une autre possibilité est d'avoir des représentations localisées, ce qui permet d'identifier le rôle de chaque neurone plus facilement. Les réseaux de neurones ont quand même une tendance à produire des présentations distribuées.

III – 2 -3.Les différents types de réseaux de neurones

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

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

Le réseau de Hopfield [46] est un réseau avec des sorties binaires où tous les neurones sont interconnectés avec des poids symétriques. C'est-à-dire que le poids du neurone N_i au neurone N_j est égal au poids du neurone N_j au neurone N_i. Les poids sont donnés par l'utilisateur. Les poids et les états des neurones permettent de définir l'«énergie» du réseau.

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

Les cartes auto-organisatrices de Kohonen [47] sont utilisées pour faire des classifications automatiques des vecteurs d’entrées.

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

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

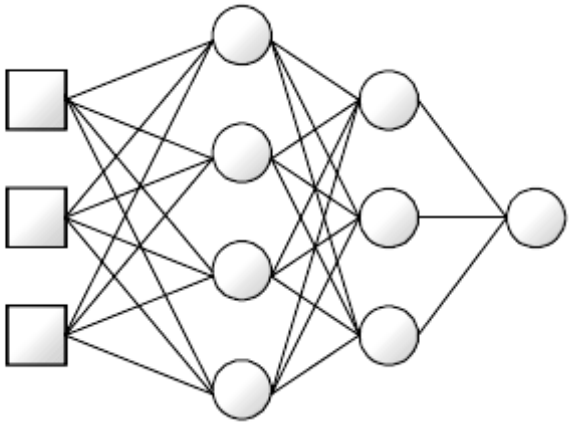
III – 2 -.3-1 Les réseaux multicouches ou perceptron multicouches (PMC)

Les réseaux multicouches (PMC) (figure 8) se composent des entrées, une couche de sortie et zéro ou plusieurs couche cachées [45]. Les connexions sont permises seulement d’une couche inférieure (plus proche des entrées) vers une couche supérieure (plus proche de la couche de sortie). Il est aussi interdit d'avoir des connexions entre des neurones de la même couche.

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

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

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



Les entrées Couches cachées Couche de sortie

Figure – 8 *Structure générale du perceptron multicouches*

Il a été prouvé [48] qu'il existe toujours un réseau de neurones de ce type avec trois couches seulement (les entrées, couche de sortie et une couche cachée) qui peut approximer une fonction f : $[0.1]^n \Rightarrow R^n$ avec n'importe quelle précision $\varepsilon > 0$ désirée. Un problème consiste à trouver combien de neurones cachés sont nécessaires pour obtenir cette précision. Un autre problème est de s'assurer a priori qu'il est possible d'apprendre cette fonction.

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

III – 2 -.4. Apprentissage :

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

Il existe trois types d'apprentissages principaux. Ce sont l'apprentissage supervisé, l'apprentissage non-supervisé, et l'apprentissage par tentative (graded training en anglais) [48].

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

Dans le cas de l'apprentissage non-supervisé le réseau décide lui-même quelles sont les bonnes sorties. Cette décision guidée par un but interne au réseau qui exprime une configuration idéale à atteindre par rapport aux exemples introduits. Les cartes auto-organisatrices de Kohonen sont un exemple de ce type de réseau [47].

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

Il existe plusieurs règles pour chaque type d'apprentissage. L'apprentissage supervisé est le type le plus utilisé. Pour ce type d'apprentissage la règle la plus utilisée est celle de Widrow-Hoff. D'autres règles d'apprentissage sont par exemple la règle de Hebb, la règle de perceptron, la règle de Grossberg etc …. [45, 48,49].

III – 2 -.4.- 1 L'apprentissage de Widrow-Hoff :

La règle d'apprentissage de Widrow-Hoff est une règle qui permet d'ajuster les poids d'un réseau de neurones pour diminuer à chaque étape l'erreur commise par ce réseau de neurones (à condition que le facteur d'apprentissage soit bien choisi).

Un poids est modifié en utilisant la formule suivante :

$$w_{k+1} = w_k - \alpha \delta_k x_k \qquad (3)$$

Où :

w_k est le poids à l'instant k ;

w_{k+1} le poids à l'instant k-1 ;

α est le facteur d'apprentissage ;

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

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

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

Deux versions améliorées de cet apprentissage existent, la version 'par lois' et la version 'par inertie' (momentum en anglais) [48], dont l'une utilise plusieurs exemples pour calculer la moyenne des changements requis avant de modifier le poids et l'autre empêche que le changement du poids au moment k ne devienne beaucoup plus grand qu'au moment k-1.

III – 2 -.4.- 2 L'apprentissage par rétro-propagation du gradient (Levenberg-Marquardt backpropagation)

L'algorithme d'apprentissage par rétro-propagation du gradient (figure9) est un algorithme itératif qui a pour objectif de trouver le poids des connexions minimisant l'écart commis par le réseau sur l'ensemble d'apprentissage. Cette minimisation par une méthode du gradient conduit à l'algorithme d'apprentissage par rétro-propagation.

La procédure d'apprentissage se décompose en deux étapes. Pour commencer, les valeurs d'entrées sont présentées au réseau, qui propage ensuite ces valeurs jusqu'à la couche de sortie et donne ainsi la réponse au réseau. A la deuxième étape les bonnes sorties

correspondantes sont présentées aux neurones de la couche de sortie qui calculent l'écart, modifient leurs poids et rétro-propagent l'erreur jusqu'aux entrées pour permettre aux neurones cachés de modifier leurs poids de la même façon. Le principe de modification des poids est normalement l'apprentissage de Widrow-Hoff.

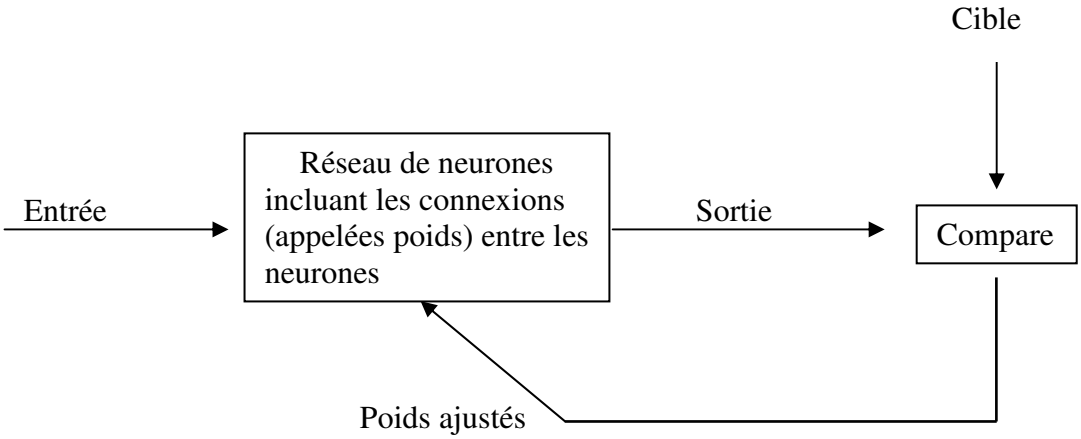


Figure – 9 *Apprentissage par un algorithme de rétro-propagation*

Généralement pour le calcul de l'écart on utilise l'erreur quadratique moyenne *EQM* (*écart quadratique moyen*) définie par la relation :

$$EQM = \frac{\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2}{n} \qquad (4)$$

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

III – 2 –5 Critères d'arrêt

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

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

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

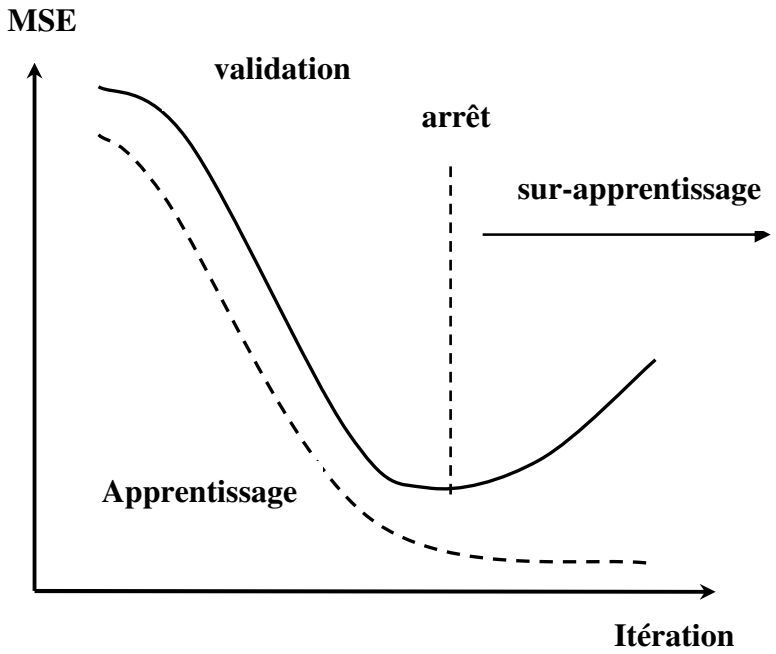


Figure – 10 *Illustration de l'arrêt précoce*

III – 2 –6 Construction d'un modèle

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

Les quatre grandes étapes de la création d'un réseau de neurones sont détaillées comme suit :

III – 2 –6- 1 Construction de la base de données

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

Afin de développer une application à base de réseaux de neurones, il est nécessaire de disposer de deux bases de données, une pour effectuer l'apprentissage et l'autre pour tester le réseau obtenu et déterminer ses performances. Pour cette raison nous avons partagé notre base des données (tableau I) aléatoirement en trois sous-ensembles comme suit :

- Un ensemble de 69 composés pour l'apprentissage du réseau de neurones.

- Un deuxième de 20 composés pour la validation.

- Et un troisième de 20 composés choisis aléatoirement de l'ensemble d'apprentissage pour le test.

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

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

$$x_{norm} = 2 \times \frac{\left(x_j - x_{\min}\right)}{\left(x_{\max} - x_{\min}\right)} - 1 \qquad (5)$$

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

III – 2 –6- 2 Définition de la structure du réseau

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

III – 2 – 6 - 3 Nombre de couches et de neurones cachés

Mis à part les entrées et la couche de sortie, il faut décider du nombre de couches cachées. Sans couche cachée, le réseau n'offre que de faibles possibilités d'adaptation. Néanmoins, il a été démontré qu'un Perceptron Multicouches avec une seule couche cachée pourvue d'un nombre suffisant de neurones, peut approximer n'importe quelle fonction avec la précision souhaitée [50].

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

III – 2 – 6 - 4 Présentation de l'environnement utilisé

Dans cette optique, le logiciel MATLAB [51], qui contient un module consacré au développement de réseaux de neurones, a été retenu ; un PC Dell P4 avec une Ram de 512 et une vitesse de 3.4 GHZ a été utilisé.

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

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

Les fonctions sigmoïde de transfert, tangente hyperbolique et linéaire, ont été adoptées comme fonctions d'activation pour les couches cachée et de sortie.

Nous présentons l'algorithme du réseau de neurones utilisé dans la page suivante :

Algorithme du réseau de neurones utilisé :

```
P= [les descripteurs];
T= [la propriété physique étudiée];
N = 89 ;       % tous les composés
N1 =69 ;       % Composés d'apprentissage
N2 =20 ;       % Composés de validation
P0= (P)';      % Transposition de la matrice P
T0= (T)';      % Transposition de la matrice T
[pn,minP,maxP,tn,minT,maxT] = premnmx(P0,T0);     % Normalisation entre [-1,+1]
P1n= (pn)';
T1n= (tn)';
% Apprentissage
P1=Pn(1:N1,:)';     % Descripteurs normalisés d'apprentissage
T1=Tn(1:N1,:)';     % Propriété physique normalisée d'apprentissage
T10=T(1:N1,:)';
%  Test
[R, Q] = size (P1);
iitst = [3:3:Q 2:42:Q];   % Choix aléatoire de 11 composés du test
test.P = P (:,iitst);
T20=T10 (:,iitst);
%  Validation
val.P = Pn(N1+1:N,:)';   % Descripteurs normalisés de validation
val.T = Tn(N1+1:N,:)';   % Propriété physique normalisée de validation
T30=T (N1+1:N, :)';
net = newff(minmax(P),[ S1 S2],{ TF1 TF2}, BTF);     % Création d'un réseau
 % S1 : Neurones de la couche cachée – S2 : la sortie (=1)
 % TF1, TF2 : Fonctions de transferts – BTF : Fonction de transfert de rétro-propagation
net.trainParam.epochs =500;            % Nombre d'itération
net.trainParam.goal= 0.0000001;      % Erreur désirée
net = init(net);        % Initialisation du réseau
[net,tr]=train (net, P1, T1, [], [], val);      % Entrainement du réseau
plotperf(tr)
a1n=sim(net,P1);       % Simulation du réseau pour les données d'apprentissage
[a1]=postmnmx(a1n,minT0,maxT0);%Remettre les résultats d'apprentissage à leurs valeurs réels
E1= T10-a1; % Calcul de l'erreur
a2n=sim(net,test.P);  % Simulation du réseau pour les données du test
[a2]=postmnmx(a2n,minT0,maxT0);%Remettre les résultats du test à leurs valeurs réels
E2= T20-a2; % Calcul de l'erreur
a3n=sim(net,val.P);   % Simulation du réseau pour les données de validation
[a3]=postmnmx(a3n,minT0,maxT0);%Remettre les résultats de validation à leurs valeurs réels
E3= T30-a3;  % Calcul de l'erreur
```

IV – Paramètres d'évaluation de la qualité de l'ajustement

Deux paramètres sont couramment utilisés :

- Le coefficient de détermination multiple :

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

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

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

$$\sigma_N = \sqrt{\frac{1}{n} \sum_1^n (y_i - \hat{y}_{(i)})^2} \quad (7)$$

IV – 1 Robustesse du modèle

La stabilité du modèle a été explorée en utilisant la "validation croisée par omission d'une observation" (LOO : cross-validation by leave-one-out) [52]. Elle consiste à recalculer le modèle sur (n – 1) hydrocarbures, le modèle obtenu servant alors à estimer le paramètre physique du composé éliminé noté $\hat{y}_{(i)}$. On répète le procédé pour chacun des n hydrocarbures.

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

$$PRESS = \sum_1^n (y_i - \hat{y}_{(i)})^2 \quad (8)$$

est une mesure de la dispersion de ces estimations. On l'utilise pour définir le coefficient de prédiction :

$$Q^2_{LOO} = \frac{SCT - PRESS}{SCT} \quad (9)$$

Contrairement à R^2, qui augmente avec le nombre de paramètres de la régression, le facteur Q^2_{LOO} affiche une courbe avec maximum (ou avec palier), obtenu pour un certain nombre de variables explicatives, puis décroît par la suite de façon monotone. Ce fait confère une grande importance au coefficient Q^2_{LOO}. Une valeur de $Q^2_{LOO} > 0{,}5$ est, généralement, considérée comme satisfaisante, et une valeur supérieure à 0,9 est excellente [53].

IV – 2 Détection des observations aberrantes

Elle a été basée sur la non – satisfaction à trois au moins (pour n'en privilégier aucun) des six tests statistiques couramment utilisés pour la détection de telles observations en analyse de régression :

1°/ Les résidus ordinaires e_i, différences entre les valeurs observées (y_i) et estimées par le modèle $(\hat{y}_i)$.

2°/ Les résidus normalisés d_i, obtenus en divisant les e_i par l'écart type s de l'équation de régression.

3°/ Le résidu studentisé interne r_i, est le résidu d'une prédiction divisé par son écart type propre $\left(r_i = e_i / s\sqrt{1 - h_{ii}}\right)$.

4°/ Les leviers, h_{ii}, permettent de juger de l'influence d'une observation i dans la détermination de l'équation de régression.

5°/ La statistique représentée par le symbole DFITS :

$$\text{DFITS} = \frac{1}{p}\sqrt{\left(\frac{h_{ii}}{1 - h_{ii}}\right)}\, t_i \qquad (10)$$

permet de mesurer l'influence d'une observation i sur la valeur ajustée ou prédite. Belsley, Kuh et Welsch [23] considèrent qu'une observation pour laquelle DFITS > $2\sqrt{p/n}$ (p étant le nombre de paramètres de la régression) est inhabituelle.

6°/ La distance de Cook D_i :

$$D_i = \frac{1}{p}\frac{h_{ii}}{1 - h_{ii}} d_i^2 \qquad (11)$$

permet d'étudier l'influence d'une observation i sur les coefficients de régression estimés par les moindres carrés. Cook [55] et Weisberg [56] suggèrent de comparer D_i au paramètre de Fisher $F_{(0,5,p,n-p)}$ et de contrôler les observations avec distances de Cook > $F_{(0,5,p,n-p)}$. Comme $F_{(0,5,p,n-p)} \approx 1$, on considère que les observations pour lesquelles $D_i > 1$ sont influentes.

IV – 3 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 QSRR, selon la méthode habituelle. Ce procédé est répété plusieurs fois (100 dans notre cas).

IV – 4 Validation externe

En plus du test de randomisation, il est intéressant [57], 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 EQMC)
- Ensemble de validation croisée (appelée également EQMP)
- Ensemble de prédiction externe (désignée par EQMPext).

Ces valeurs RMSE sont mieux adaptées, pour juger de la qualité d'un modèle que les valeurs de R^2 et Q^2 seules, qui constituent de bons tests uniquement pour des données réparties régulièrement.

$$EQMC = \sqrt{\frac{\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2}{n}} \qquad (12)$$

$$\sigma_N = EQMP = \sqrt{\frac{\sum_{i=1}^{n}\left(y_i - \hat{y}_{(i)}\right)^2}{n}} = \sqrt{\frac{PRESS}{n}} \qquad (7)$$

$$EQMP_{ext} = \sqrt{\frac{\sum_{i=1}^{n_{ext}}\left(y_i - \hat{y}_{(i)}\right)^2}{n_{ext}}} \qquad (13)$$

PARTIE EXPERIMENTALE

FACTEUR DE CAPACITE k'

I-1- modèle hybride algorithme génétique / réseaux de neurones artificiels

I - 1- 1 choix des paramètres statistiques

On commencera par choisir la taille du modèle. Le graphe suivant montre une valeur maximale du FIT égale à 3, qui fixe ainsi le nombre optimal de descripteurs.

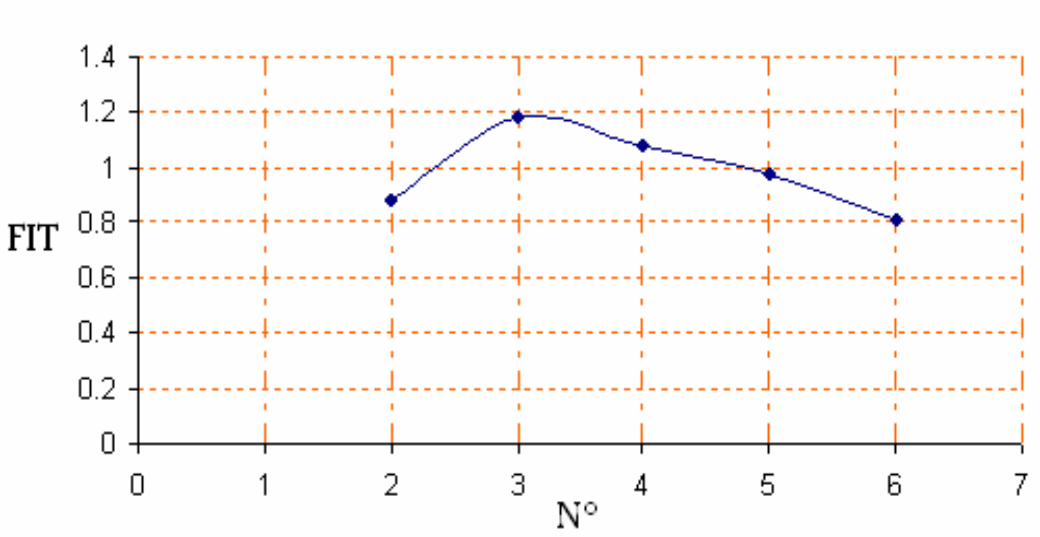


Figure 11 – *Variation du FIT en fonction du nombre de descripteurs pourK'.*

Les trois descripteurs retenus pour le modèle sont présentés dans le tableau V.

Tableau V - Descripteurs moléculaires intervenant dans la modélisation de facteurs de capacité [23].

N°	Descripteur	Classe	Signification
1	X.T	Composition de la phase mobile et la température	/
2	HATS7e	Indice d'auto corrélation 2D	Descripteur (GETAWAYS)
3	RTe+	Indice pondéré maximal	Descripteur (GETAWAYS)

Ces descripteurs sont utilisés pour la configuration du réseau de neurones, qui est perfectionnée en phase d'apprentissage ; les paramètres de fonctionnement sont déterminés de façon à obtenir une bonne adéquation entre les valeurs simulées et les données d'apprentissage, combinée à une généralisation correcte de ces simulations.

I - 1- 2 Choix du nombre de couches cachées

Quelle que soit la problématique étudiée, l'utilisation d'une seule couche cachée a permis d'obtenir de meilleures configurations des réseaux de neurones.

I - 1- 3 Choix du nombre de neurones dans la couche cachée

Le choix de ce nombre est très important. On fixe a priori un nombre d'itérations (50 par exemple), puis on discrétise l'ensemble des valeurs possibles pour le nombre de neurones cachés (par exemple entre 2 et 18 avec un pas de 2). On fixe la valeur du nombre de neurones de la couche cachée par la valeur minimale de l'erreur quadratique moyenne EQM.

Le graphe EQM= f(nombre de neurones) de la figure suivante permet de visualiser cet impact et de fixer à 10 le nombre de neurones de la couche cachée.

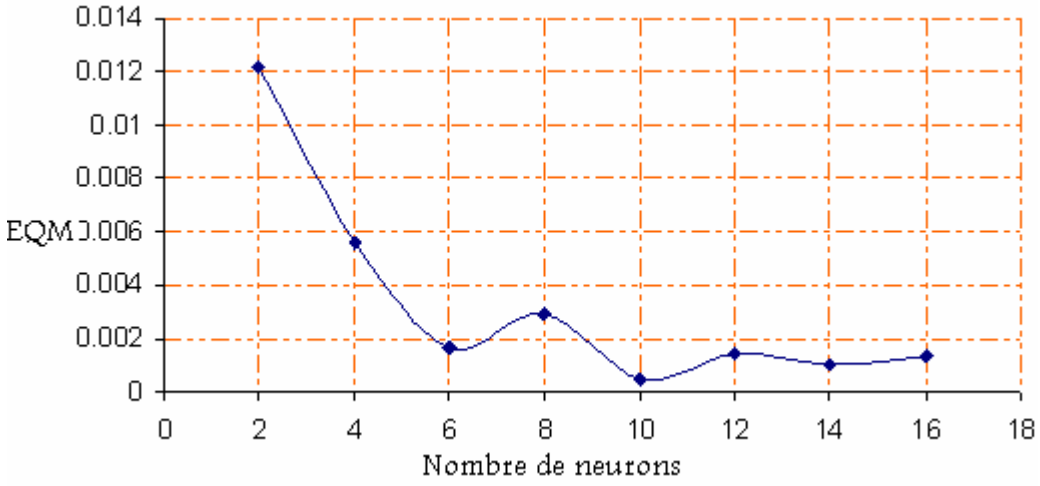


Figure 12 – *Choix du nombre de neurones de la couche cachée.*

I - 1- 4 Choix de la fonction de transfert

Les réseaux de neurones les plus adaptés à notre étude ont l'architecture suivante :

- Fonction de transfert tangente hyperbolique (tansig) pour la couche cachée
- Fonction de transfert linéaire (purelin) pour la couche de sortie.

I - 1- 5 Choix des paramètres d'apprentissage

Ces paramètres sont également importants et ont permis d'affiner la configuration des réseaux de neurones pour obtenir les meilleures prédictions.

- ♣ Indice de performance choisi: EQM (pour Mean Square Error)
- ♣ L'algorithme de rétropropagation est celui de Levenberg-Marquardt, le rapport vitesse d'exécution / mémoire requise étant le meilleur. La fonction d'apprentissage Matlab de cet algorithme est *trainlm*.

L'apprentissage du réseau de neurones représente un fragile équilibre entre tous ces paramètres, d'où la difficulté pour l'atteindre. Une fois cet apprentissage achevé, le réseau de neurones devient un outil viable et peut être utilisé pour la simulation de nouvelles données. Le tableau VI précise la structure optimale du réseau de neurones.

Tableau – **VI** Structure optimale du réseau de neurones.

Nombre d'entrées	03 (les descripteurs)
Nombre de sorties	01 (facteur de capacité)
Nombre de couches cachées	Une couche cachée
Nombre de neurones dans la couche cachée	10
Algorithme d'apprentissage	Rétro propagation du gradient de l'erreur
Fonctions d'apprentissage	Tangente hyperbolique (couche cachée) Linéaire (couche de sortie)

I - 1- 6 Résultats et discussion

I - 1- 6 – 1 Evaluation de la qualité de l'ajustement

Nous avons deux paramètres utilisés pour évaluer la qualité de l'ajustement; la valeur du cœfficient de détermination R^2 = 99, 31% qui explique très bien la variabilité de K' en fonction des descripteurs choisis; la racine de l'erreur quadratique moyenne de prédiction σ_N=0,15, dont la petite valeur indique un modèle très hautement significatif, que justifie la grande valeur du paramètre de Fisher : F=3154,83.

I - 1- 6 – 2 Vérification de la qualité de l'ajustement

La qualité de l'ajustement a été vérifiée en procédant à une validation croisée par « Leave –one –out ». La figure (13) de la page suivante, qui reproduit les valeurs prédites de k' en fonction de celles observées, fait ressortir un bon ajustement, d'ailleurs confirmé par la grande valeur de Q^2 (=99.2%)

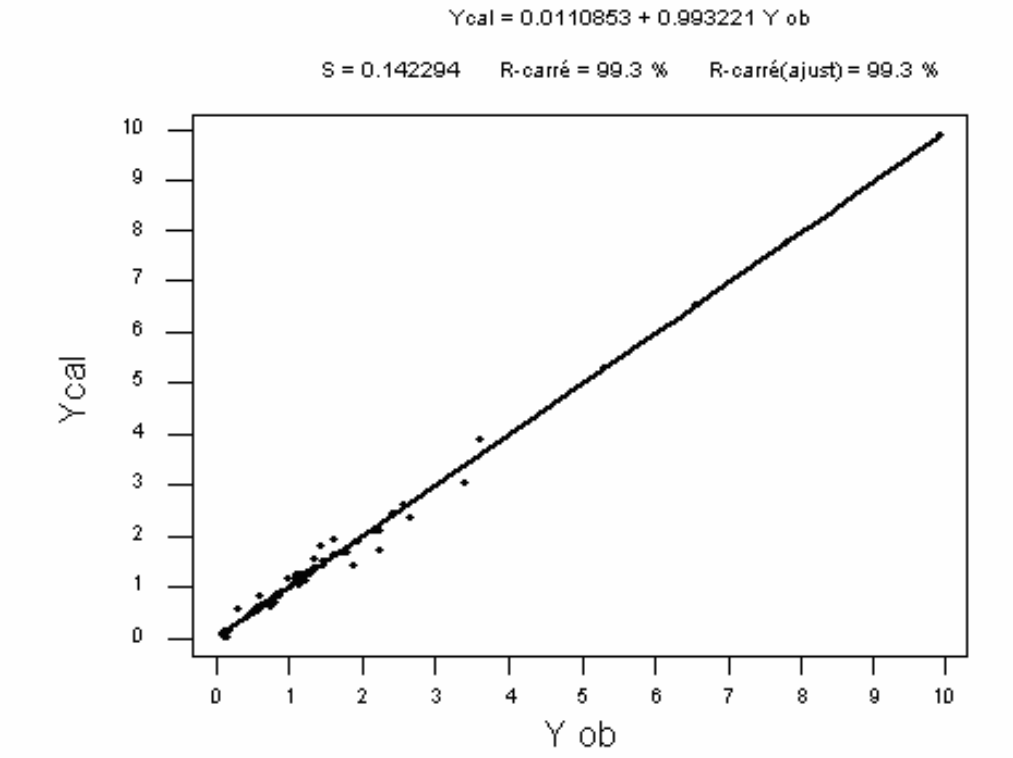


Figure 13 – *Graphe des valeurs prédites* K' *en fonction des valeurs observées* K'.

I-2 Validation externe

L'évaluation de la capacité de généralisation du réseau est réalisée sur la base de la validation externe, la performance du réseau est alors mesurée par le coefficient de régression R^2.

Les résultats obtenus montrent que les valeurs prédites (tableau - VII) sont très proches des valeurs observées (figure-14). La valeur de R^2 est égale à 99,3 % relative est de ce qui confirme que le modèle neuronal décrit de façon adéquate la relation entre les facteurs de capacité prédites et observées.

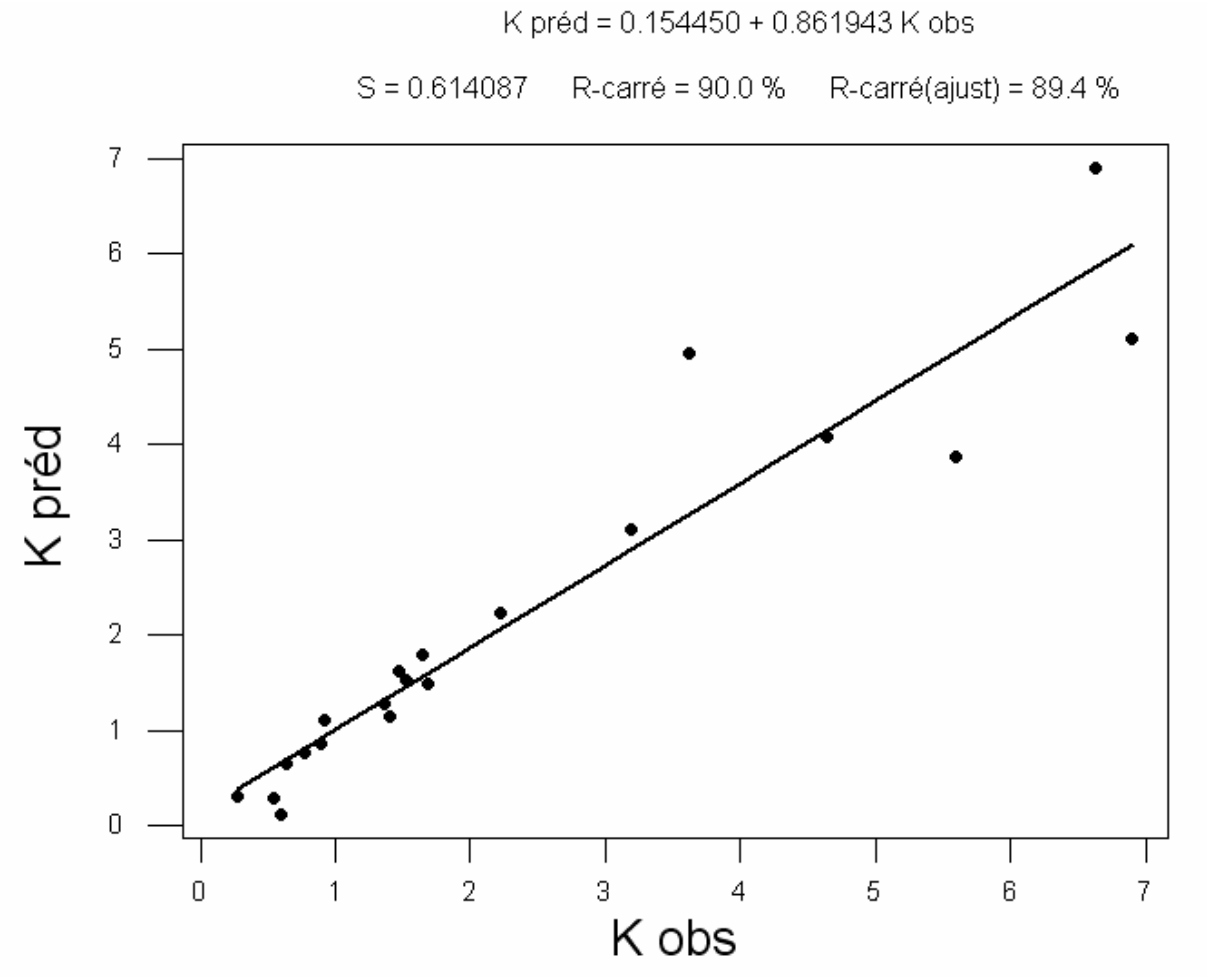


Figure 14 – *Graphe des K' prédites en fonction des K' observées pour validation*

La comparaison des résidus de prédiction $e_{(i)}$(tableau VII, p 57) avec deux fois la valeur de S (= 1.228) montre qu'il n'y a pas d'observation aberrante.

Les résultats confirment ainsi la faisabilité de l'approche neuronale comme technique de modélisation de facteur de capacité.

Tableau - VII Les valeurs k' observées, prédites et les erreurs pour l'ensemble de validation externe par RNA.

	k'c	k'	$e_{(i)}$
70	2,226	2,2369	-0,0109
71	1,465	1,6189	-0,1539
72	4,644	4,082	0,562
73	1,524	1,5216	0,0024
74	1,642	1,7898	-0,1478
75	0,591	0,1173	0,4737

76	0,914	1,1063	-0,1923
77	1,394	1,1449	0,2491
78	0,764	0,7641	-1.00E-04
79	1,688	1,4791	0,2089
80	0,637	0,6518	-0,0148
81	1,363	1,2752	0,0878
82	0,882	0,8647	0,0173
83	0,275	0,3038	-0,0288
84	0,543	0,2951	0,2479
85	6,633	6,8981	-0,2651
86	5,604	3,8755	1,7285
87	3,193	3,1001	0,0929
88	6,907	5,0999	1,8071
89	3,623	4,9546	-1,3316

Les valeurs RMSE sont réunies ci-après :

EQMC = 0,14	(69objets)	R^2 = 99,31	
EQMP = 0,15	(69objets)	Q^2 = 99,20	
EQMP (ext) = 0.44	(20 objets)	Q^2 (ext) = 89,84	

Les faibles valeurs des RMSE montrent une bonne capacité prédictive du modèle et une possibilité d'extension suffisante (valeurs proches ou similaires).

CONCLUSION GENERALE

V- CONCLUSION GENERALE

L'optimisation des conditions de séparation en CLHP à polarité de phase inversée nécessite l'utilisation de modèles. La littérature en fait ressortir plusieurs, basés sur des relations fonctionnelles différentes. Ces modèles sont souvent évalués selon le critère statistique d'ajustement de la rétention d'un seul soluté pour différentes compositions de la phase mobile.

Une telle approche ne permet pas de s'assurer de la justesse des hypothèses de base des modèles de rétention.

Les performances de 2 modèles de la littérature ont été reliées à la qualité de l'ajustement des données de rétention d'un ensemble de phénols non congénères, obtenues en faisant varier les conditions de séparation (composition de la phase mobile ; température).

Le modèle de Kowalska, qui semble facile d'utilisation, est basé sur les associations possibles des molécules de la phase mobile qui peuvent être modifiées selon la température de travail dont il n'est pas tenu compte dans le modèle. Les mauvais critères statistiques (coefficient de détermination ; erreur standard …) peuvent refléter un écart sensible aux hypothèses de base du modèle.

L'élimination de 2 phénols a permis de justifier les hypothèses du modèle d'Horváth qui tient compte de la température des expérimentations, et fait intervenir une température dite de compensation déterminée expérimentalement. L'amélioration des critères statistiques de base n'est pas pour autant un gage d'une bonne capacité prédictive du modèle.

Nous avons appliqué la méthodologie QSRR pour relier les indices de rétentions de composés de phénols étudiés, à des descripteurs moléculaires expérimentaux reflétant certaines particularités des molécules considérées.

Le mélange pris en compte comprend 89 données .

Les modèles QSRR ont été établis en utilisant l'analyse de régression multilinéaire et /ou les réseaux de neurones standards à 3 couches (les entrées, une couche cachée et une couche de sortie), avec algorithme d'apprentissage par rétro- propagation du gradient (Levenberg- Marquard)).

Les 89 données de base ont été éclatées aléatoirement en deux ensembles disjoints, invariants pour tous les modèles :

- un ensemble principal de 69 données utilisées pour le calcul et, les essais du modèle ;
- un ensemble de 20 données pour la prédiction externe.

La taille du modèle est fixée par la valeur optimale de la fonction FIT de KUBINYI. La sélection des variables explicatives a été réalisée par algorithme génétique, dans la version MOBYDIGS de TODESCHINI [58], en maximisant Q^2_{L00}.

Les statistiques réunies ci-après permettent de faire des comparaisons, et de tirer plusieurs conclusions comparons les résultats trouvés par RNA.

		(I)	(II)	(III)	(IV)	(V)	(VI)	(VII)	(VIII)
		R^2 (%)	**σ_N**	**Q^2 (%)**	**F**	**Q^2_{ext} (%)**	**RMSE**	**Points * aberrants**	**Points ** aberrants**
k’	(RNA)	99,31	0,15	99,20	3154,83	89,84	0,15 -0,44	-	75-86

Points de l’ensemble d’essai (*) et de l’ensemble de prédiction externe (**).

ANEXE

	Y	(X.T)	HATS7e	RTe+
A2	7,78	73,75	0,068	-3,42
A3	2,226	147,499	0,068	-3,42
A4	1,036	206,5	0,068	-3,42
A5	0,529	250,749	0,068	-3,42
A6	2,556	142,498	0,068	-3,42
A7	1,939	152,5	0,068	-3,42
A8	1,595	157,5	0,068	-3,42
A9	1,451	162,499	0,068	-3,42
B1	9,92	44,25	0,131	-3,36
B2	6,551	73,75	0,131	-3,36
B3	2,228	147,499	0,131	-3,36
B4	1,012	206,5	0,131	-3,36
B5	0,543	250,749	0,131	-3,36
B6	1,429	142,498	0,131	-3,36
B7	1,789	152,5	0,131	-3,36
B8	1,465	157,5	0,131	-3,36
B9	1,319	162,499	0,131	-3,36
C1	6,907	44,25	3,63	0,059
C2	4,644	73,75	3,63	0,059
C3	1,524	147,499	3,63	0,059
C4	2,216	206,5	3,63	0,059
C5	0,795	250,749	3,63	0,059
C6	1,714	142,498	3,63	0,059
C7	1,351	152,5	3,63	0,059
C8	1,156	157,5	3,63	0,059
C9	1,063	162,499	3,63	0,059
D1	6,633	44,25	0,131	-3,36
D2	5,303	73,75	0,131	-3,36
D3	1,642	147,499	0,131	-3,36
D4	0,819	206,5	0,131	-3,36
D5	0,595	250,749	0,131	-3,36
D6	2,167	142,498	0,131	-3,36
D7	1,448	152,5	0,131	-3,36
D8	1,22	157,5	0,131	-3,36
D9	1,119	162,499	0,131	-3,36
E1	5,604	44,25	3,05	0,06
E2	3,407	73,75	3,05	0,06
E3	1,243	147,499	3,05	0,06
E4	0,739	206,5	3,05	0,06
E5	0,591	250,749	3,05	0,06
E6	1,346	142,498	3,05	0,06
E7	1,105	152,5	3,05	0,06
E8	0,985	157,5	3,05	0,06
E9	0,914	162,499	3,05	0,06
K1	3,193	44,25	0,111	-2,74
K2	2,41	73,75	0,111	-2,74
K3	1,109	147,499	0,111	-2,74
K4	0,695	206,5	0,111	-2,74
K5	0,551	250,749	0,111	-2,74
K6	1,394	142,498	0,111	-2,74
K7	0,998	152,5	0,111	-2,74
K8	0,875	157,5	0,111	-2,74
K9	0,826	162,499	0,111	-2,74

L1	2,64	44,25	0,066	-2,29
L2	1,609	73,75	0,066	-2,29
L3	0,764	147,499	0,066	-2,29
L4	0,564	206,5	0,066	-2,29
L5	0,509	250,749	0,066	-2,29
L6	0,848	142,498	0,066	-2,29
L7	0,677	152,5	0,066	-2,29
L8	0,614	157,5	0,066	-2,29
L9	0,579	162,499	0,066	-2,29
M1	1,688	44,25	0,214	-2,12
M2	1,262	73,75	0,214	-2,12
M3	0,717	147,499	0,214	-2,12
M4	0,507	206,5	0,214	-2,12
M5	0,48	250,749	0,214	-2,12
M6	0,781	142,498	0,214	-2,12
M7	0,637	152,5	0,214	-2,12
M8	0,586	157,5	0,214	-2,12
M9	0,544	162,499	0,214	-2,12
N1	3,623	44,25	2,991	0,048
N2	3,611	73,75	2,991	0,048
N3	1,363	147,499	2,991	0,048
N4	0,59	206,5	2,991	0,048
N5	0,417	250,749	2,991	0,048
N6	1,858	142,498	2,991	0,048
N7	1,225	152,5	2,991	0,048
N8	0,994	157,5	2,991	0,048
N9	0,882	162,499	2,991	0,048
P1	0,285	44,25	2,042	0,059
P2	0,275	73,75	2,042	0,059
P3	0,119	147,499	2,042	0,059
P4	0,109	206,5	2,042	0,059
P5	0,16	250,749	2,042	0,059
P6	0,145	142,498	2,042	0,059
P7	0,101	152,5	2,042	0,059
P8	0,083	157,5	2,042	0,059
P9	0,066	162,499	2,042	0,059

REFERENCES BIBLIOGRAPHIQUES

REERENCES BIBLIOGRAPHIQUES

1- *Wikipedia internet.*

2– A.J.P. Martin, *some theoretical aspects of partition chromatography*, Biochem. Soc. Symp.., 3, 4-20, 1950.

3– R. Kaliszan, *Quantitative structure – chromatographic retention relationships*, Wiley-Interscience, New York, pp 303, 1987.

4– J. Green, S. Marcinkiewicz, D. Mc Hale, *Paper chromatography and chemical structure. III. The correlation of complex and simple molecules. The calculation of R_M values for to copherols, vitamins K, ubiquinones and ubichromenols from R_M (phenol). Effects of insaturation and chain branching*, J. Chromatogr., 10, 158-183, 1963.

5– J. Iwasa, T. Fujita, C. Hansch, *Substituent constants for aliphatic functions obtained from partition coefficients*, J. Med. Chem., 8, 150-153, 1965.

6– E. Tomlinson, *Chromatography hydrophobic parameters in correlation analysis of structure-activity relationships*, J. Chromatogr., 113, 1-45, 1975.

7– R. Kaliszan, *Chromatography in studies of quantitative structure-activity relationships,* J. Chromatogr., 220, 71-83, 1981.

8– R. Kaliszan, High performance liquid *chromatography as a source of structural information for medicinal chemistry,* J. Chromatogr., 22, 362-370, 1984.

9– R. Kaliszan, *Quantitative relationships between molecular structure and chromatographic retention. Implications in physical, analytical and medicinal chemistry,* Crit. Rev. Anal. Chem., 16, 323-328, 1986.

10– R. Kaliszan, H. Foks, *The relationship between the R_M values and the connectivity indices for pyrazine carbothioamide derivatives,* Chromatographia, 10, 346-349, 1977.

11 – R. Kaliszan, *Correlation between the retention indices and connectivity indices of alcohols and methyl esters with complex cyclic structure,* Chroamtographia, 10, 529-531, 1977.

12 – Y. Michotte, L. Massart, *Molecular connectivity and retention indexes,* J. Pharm. Sci., 66, 1630-1632, 1977.

13 – B. K. Chen, Cs. Horváth, *Evaluation of substituent contribution to chromatographic retention: Quantitative structure-retention relationships,* J. Chromatogr., 171, 15-28, 1979.

14 – P.J. Schoenmakers, H.A.H. Billiet, L. De Galan, *The solubility parameter as a tool in understanding liquid chromatography,* Chromatographia, 15, 205-214, 1982.

15 – E. Tomlinson, *Boxes in boxes: Cases for extrathermodynamics,* British Pharmaceutical Conference Science Award Lecture, Bringhton, 1981.

16 – C. Reichardt, *Solvent and solvent effects in organic chemistry,* Verlag Chemie, Weinheim, New York, 1990

17 – W. Melander, D.E. Campbell, Cs. Horváth, *Enthalpy-entropy compensation in reversed-phase chromatography*, J. Chromatgr., 158, 215-218, 1978.

18 –C. Nistor, J. Emnéus, L. Gorton, A. Ciucu, *Improved stability and altered selectivity of tyrosinase based graphite electrodes for detection of phenolic compounds*, Anal. Chim. Acta, 387, 309-326, 1999.

19 – Seung Ki Lee, Yulia Polyakova, Kyung Ho Row, *Evaluation of predictive retention factors for phenolic compounds with QSPR equations*, J. Liq. Chromatogr. & Rel. Tech., 27(4), 629-639, 2004.

20– Q.S. Wang, L. Zhang, M. Zhang, X.D. Xing, G.Z. Tang, *A system for predicting the retention of o-alkyl,n-(1-methylthioethylideneamino) phosphoramidates on RP–HPLC*, Chromatographia, 49 (7/8), 444-448, 1999.

21 – H, Nacer, D. Messadi, *Relation structure - rétention chromatographique pour les phénols*, Rev. COST, 4, 81-87, 2006.

22 – M. Karelson, *Molecular Descriptors in QSAR / QSPR*, Wiley–Interscience, New York, 2000.

23 – R. Todeschini, V. Consonni, *Handbook of Molecular Descriptors*, Wiley–VCH, Weinheim 2000.

24 – J.G. Dorsey, M.G. Khaledi, *Hydrophobicity estimations by reversed-phase liquid chromatography: implications for biological partitioning processes*, J. Chromatogr. A, 656, 485-499, 1993.

25 – W.J. Lambert, *Modelling oil-water partitioning and membrane permeation using reversed-phase chromatography*, J. Chromatogr. A, 656, 469-484, 1993.

26 – R.R. Krug, W.G. Hunter, R.A. Grieger, *Enthalpy-entropy compensation. I. Some fundamental statistical problems associated with the analysis of Van't Hoff and Arrhenius data*, J. Phys. Chem., 80, 2335-2341, 1976.

27 – R.R. Krug, W.G. Hunter, R.A. Grieger, *Enthalpy-entropy compensation. II. Separation of the chemical from the statistical effects*, J. Phys. Chem., 80, 2341-2351, 1976.

28 – W. Melander, B.K. Chen, Cs. Horváth, *Mobile phase effects in reversed-phase chromatography. VII. Dependence of retention on mobile phase composition and column temperature*, J. Chromatgr., 318, 1-10, 1985.

29 – T. Kowalska, *Physico-chemical modelling of solute retention in reversed-phase HPLC with methanol-water mobile phase*, Chromatographia, 27, 628-630, 1989.

30– T. Kowalska, P. Kus', *On the mechanism of solute retention in reversed-phase HPLC systems with methanol-water eluent*, Chromatographia, 29, 583-586, 1990.

31 – W. Zapala, K. Kaczmarski, T. Kowalska, *Comparison of different retention models in normal - and reversed-phase liquid chromatography with binary mobile phases*, J. of Chromatog. Sci., 40, 575-580, 2002.

32 – C.H. Lochmuller, C. Reese, A.J. Aschman, S.J. Breiner, *Current strategies for prediction of retention in high-performance liquid chromatography*, J. Chromatogr. A, 656, 3-18, 1993.

33 – M. Jaroniac, *Partition and displacement models in reversed-phase liquid chromatography with mixed eluents*, J. Chromatogr. A, 656, 37-50, 1993.

34 – R. Tijssen, P.J. Schoenmakers, M.R. Bohmer, L.K. Koopal, H.A.H. Billiet, *Lattice models for the description of partitioning / adsorption and retention in reversed-phase liquid chromatography including surface and shape effects*, J. Chromatogr. A, 656, 135-196, 1993.

35 – T. Kowalska, W. Prus, *On contingency in the modelling of the solute retention: The schoenmakers model as an example*, Acta Chromatogr., 7, 210-217, 1997.

36 – A. Vailaya, Cs. Horváth, *Retention in reversed-phase chromatography: partition or adsorption?*, J. Chromatogr. A, 829, 1-27, 1998.

37 – J. Ko, J.C. Ford, *Comparison of selected retention models in reversed-phase liquid chromatography*, J. Chromatgr. A., 913, 3-13, 2001.

38- Hyperchem™ Release 7.5 for windows, Molecular Modelling system, 2000.

39- R. Todeschini, V. Consonni, M. Pavan, DRAGON, Software for the calculation of Molecular Descriptors. Release 5.3 for Windows, Milano, 2005.

40- E-calc computes all the E-stat values and displays them in a convent form of the screen. The computational parts of this program have been taken from:

-Molconn-Z™ software. Lowell H., Hall Associates Consulting, 2 Devis street, Quincy, MA02170 for DOS version only.

-Sci. QSAR^{TM} 2D. Sci. Vision, Inc, 200 Wheeler Rd, Burlington, MA, 01803 for PC versions.

41- H.Kubinyi ,Quant. Struct. – Act. Relat., 13, , 285. 1994

42- D.C. Montgomery, E.A. Peck, Introduction to linear Regression Analysis, Second Edition, Wiley-Interscience Publication, New York, 1992.

43- Mc Culloch-Pitts. a logical Calculus at the ideas imminent in Nervous Activity. Bulletin at math. Biophysics.,Vol. 5, p.115-133. 1943

44- M. Minsky,S. Papert, Perceptrons. Massachusetts: MIT press, 1969.

45- D. E. Rumelbart, J. L. Mcclelland et al . Parallel Distributed processing Vol. 1. Massachusetts: MIT press, 547 p, 1988.

46- J. J. Hopfield. Neural Networks and physical systems with emergent collective computational abilities. Proceedings of the National Academy of sciences. USA..
Vol.79. p. 2554-58. 1982

47- T. Kohonen Self-organization and associative memory. Bulletin: Springer-Verlag. 984.

48- R. Hecht-Nielson Neurocomputing. Addison-Wesly Publishing Company.
. 433 p. 1990

49- F. Fogelmen-Soulié. Méthodes connexionnistes pour l'apprentissage. Actes des journées Nationales sur l'intelligence Artificielle. Paris: Teknea.. p. 275-293. 1988

50- K. Hornik, Approximation capabilities of multilayer feedforward networks, Neural Networks,). 251-257. 4 (1991)

51- Matlab Version 7.0.0.19920 (Release 14) The Language of Technical Computing The MathWorks, Inc. May 06, 2004.

52- N.R Draper, H. Smith, Applied Regression Analysis, Third Edition, Wiley series in Probability and Statistics, New york, 1998.

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

54- D.A. Belsley, E. Kuh, R.E. Welsch, Regression Diagnostics : Identifying Influential Data and Sources of Collinearity, Wiley, New York, 1980.

55- D. Cook, Detection of Influential Observations in linear Regression. Technometrics. 19, 15-18, (1977)

56- S. Weisberg, Applied linear Regression. J. Wiley, Inc., New York, 1980.

57- R. Todeschini, V. Consonni, Handbook of Molecular Descriptors, R. Mannhold, H. Kubinyi, H. Timmerman eds., Wiley- VCH, Verlage Gmbh, Weinheim, 2000.

58– R. Todeschini, D. Ballabio, Consonni, A. Mauri and M. Pavan Milano Chemometrics and QSAR Research Group Moby Digs Professional – Version 1.0 – 2004.