الجمهورية الجزائرية الديمقراطية الشعبية
REPUBLIQUE ALGERIENNE DEMOCRATIQUE ET POPULAIRE
وزارةالتعليمالعالي و البحثالعلمي

BADJI MOKHTAR- ANNABA UNIVERSITY
UNIVERSITE BADJI MOKHTAR- ANNABA

جامعة باجي مختار-عنابة

Année 2021

FACULTE DES SCIENCES
DEPARTEMENT DE CHIMIE

THESE

Pour l'obtention du diplôme de doctorat en sciences

Option : Chimie analytique et environnement

Modélisation du facteur acentrique et du volume critique de plusieurs familles de composés

Presenté par: M. HADDAG Hamza

Devant le jury composé de :

Mme.	ALI MOKHNACHE Salima	Présidente	Professeur	Université d'Annaba
Mme	FERTIKH Nadia	Encadreur	Professeur	Université d'Annaba
Mme.	SOBHI Chafia	Examinateur	Maître de Conférences	Université de Skikda
M.	BOUDJEMA Kheir Eddine	Examinateur	Maître de Conférences	Université de Khenchla
M.	CHAOUKI Mourad	Examinateur	Maître de Conférences	Université d'Ouargla

Je dédie ce modeste travail

à ma famille

et à mes amis

Remerciements

En premier, je remercie Madame le professeur **FERTIKH Nadia** d'avoir accepté d'encadrer ce travail, pour ses conseils et son soutient.

Un grand merci à Madame le professeur **ALI MOKHNACHE Salima** qui en plus de m'avoir honoré en acceptant la présidence du jury, a été un soutient indéfectible tout au long de ces années.

Je tiens aussi à remercier les membres du jury, Madame **SOBHI Chafia** et Messieurs **BOUDJEMA Kheir Eddine** et **CHAOUKI Mourad** d'avoir accepté de faire partie de ce jury.

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

"Success is not final, failure is not fatal: it is the courage to continue that counts"

Winston Churchill

Résumé:

L'objectif de ce travail est de mettre en œuvre la procédure des relations quantitatives structure-propriété (QSPR) pour la prédiction du volume critique et du facteur acentrique de divers composés organiques, vues leurs importances dans le domaine de la chimie et du génie chimique.

L'optimisation des géométries des molécules, le calcul des descripteurs moléculaires et le choix parmi eux d'un sous-ensemble de variables explicatives pertinentes par algorithme génétique ont été effectuées par des logiciels spécialisés.

Les modèles obtenus par régression linéaire multiple et/ou réseaux de neurones artificiels sont stables, robustes avec d'assez bonnes capacités à prédire les propriétés de nouveaux composés dont les valeurs expérimentales venaient à manquer.

Par leurs simplicités, en comparaison aux méthodes de contribution de groupes aussi appliquées dans ce travail, nos deux modèles QSPR peuvent aisément remédier au problème de carence des données expérimentales.

Mots – clés: Propriétés critiques – Facteur acentrique – Modèles QSAR/QSPR – Descripteurs moléculaires – Contribution de groupes – Familles de composés organiques.

Abstract:

The aim of this work was to implement the Quantitative Structure-Property Relationships (QSPR) procedure for the prediction of critical volume and acentric factor of various organic compounds given their importance in chemistry and chemical engineering fields.

Specialized softwares have carried out molecular geometry optimization, molecular descriptors generation and the choice among them of a relevant explanatory variables subset by genetic algorithm.

The obtained models by multiple linear regression and / or artificial neural networks are stable, robust with fairly good ability to predict the properties of new compounds whose experimental values are lacking.

By their simplicity, in comparison to the group contribution methods also applied in this work, our two QSPR models can easily remedy the problem of experimental data lack.

Keywords: Critical properties - Acentric factor - QSAR / QSPR models - Molecular descriptors - Group contribution - Organic compounds families.

ملخص :

الهدف من هذا العمل هو استعمال العلاقات الكمية هيكل -خاصية (QSPR) للتنبؤ بالحجم الحاسم والمعامل اللامركزي لمركبات عضوية مختلفة نظرا لأهميتها في مجالات الكيمياء والهندسة الكيميائية.

الشكل الهندسي الأمثل للجزيئات، وحساب الواصفات الجزيئية والاختيار فيما بينها لمجموعة فرعية من المتغيرات المفسرة ذات صلة بواسطة الخوارزمية الجينية قد تم تنفيذها بواسطة برامج متخصصة.

النماذج التي تم الحصول عليها من خلال الانحدار الخطي المتعدد و / أو الشبكات العصبية الاصطناعية، مستقرة وقوية مع قدرة جيدة إلى حد ما للتنبؤ بخصائص المركبات الجديدة التي نفتقر إلى قيمها التجريبية.

من خلال بساطتهما، بالمقارنة مع طرق مساهمة المجموعات التي تم تطبيقها أيضًا في هذا العمل، يمكن لنموذجينا QSPR معالجة مشكلة نقص البيانات التجريبية بسهولة.

كلمات دالة: الخصائص الحرجة - معامل لامركزي - نماذج QSAR / QSPR - واصفات جزيئية - مساهمة المجموعات - عائلات المركبات العضوية.

SYMBOLES ET ABREVIATIONS

AD	Domaine d'application structurel
AM1	Austin Model 1
AAD	Moyenne des écarts absolus
AAD%	Moyenne des pourcentages des écarts absolus
A_i	Coefficients du viriel (cas de 1/V)
atm	Atmosphère
B_i	Coefficients du viriel (cas de P)
CNDO	Complete Neglect of Differential Overlap
CC_0	1er neurones de la couche cachée
CC_1	2ème neurones de la couche cachée
d	Statistique de Durbin-Watson
$\boldsymbol{d_{ij}}$	Distance topologique entre les atomes i et j
d_j	Incrément du groupe du 2ème ordre
E	Valeur propre
e_i	Résidu : différence entre les valeurs observée (y_i) et estimée ($\hat{y}_i$)
$\boldsymbol{e_{i\ STD}}$	Résidu standardisé de prédiction
ESpm05u	Moment spectral 05 de matrice de contiguïté de bord
ESpmku	Moments spectraux de la matrice de contiguïté de bord
EQM	Ecart quadratique moyen
EQMC	Ecart quadratique moyen calculé sur l'ensemble de calibrage
EQMP	Ecart quadratique moyen de prédiction
$EQMP_{ext}$	Ecart quadratique moyen calculé sur l'ensemble de validation externe
E_i	i^{ème} entrée du neurone
F	Statistique de Fisher
FIV	Facteur d'inflation de la variance
GA	Algorithme génétique (Genetic Algorithm)
GA-VSS	Genetic Algorithm for Variable Subsets Selection
GIC	Group Interaction Contribution
H	Matrice influence moléculaire
H	Opérateur hamiltonien
HATSkw	Descripteurs d'autocorrélation à levier pondéré de distance topologique k
HATS3v	Descripteur d'autocorrélation à levier pondéré de distance topologique 3/ pondérée par les volumes atomiques de van der Waals v
HOMO	Highest Occupied Molecular Orbital / Plus haute orbitale moléculaire occupée
$\boldsymbol{h_i}$	Eléments diagonaux de la matrice chapeau
$\boldsymbol{h_{ii}}$	Eléments diagonaux de la matrice influence moléculaire pour l'atome i
$\boldsymbol{h_{jj}}$	Eléments diagonaux de la matrice influence moléculaire pour l'atome j
$\boldsymbol{h^*}$	Valeur critique des leviers
J	Joule
K	Kelvin

SYMBOLES ET ABREVIATIONS

k	Pente de la droite de régression passant par l'origine pour les valeurs calculées par rapport aux valeurs observées
k'	Pente de la droite de régression passant par l'origine pour les valeurs observées par rapport aux calculées
L	Litre
Lkw	Indices de taille WHIM dirigés
L2 p	Seconde composante de l'indice de taille WHIM dirigé/ pondérée par les polarisabilités p
LMO	Cross-validation by Leave-Many-Out: Validation croisée par omission d'un ensemble d'observations
LOO	Cross-validation by Leave-One-Out: Validation croisée par omission d'une observation
LUMO	Lowest Unoccupied Molecular Orbital / Plus basse orbitale moléculaire inoccupée
M	Matrice moléculaire des coordonnées cartésiennes x, y, z des atomes
M	Masse moléculaire
m	Mètre
MATS2m	Autocorrélation de Moran de distance topologique égale à 2/pondérée par les masses atomiques
MATSkw	Autocorrélation de Moran de distance topologique égale à k/pondérée par la propriété atomique employée ***w***
MCG	Méthodes de Contribution de Groupes
MNDO	Modified Neglect of Differential Overlap
Morsw	Descripteurs 3D-MoRSE
Mor17 p	Signal 3D-MoRSE - 17/ pondéré par les volumes atomiques de van der Waals
Mor18v	Signal 3D-MoRSE - 18/ pondéré par les volumes atomiques de van der Waals
N	Nombre de données dans une MCG
n	Nombre de moles
n_{AT}	Nombre d'atomes dans la molécule
n_{ext}	Nombre de composés dans l'ensemble de validation
n_G	Nombre de groupes dans une MCG
n_i	Fréquence du groupe u_i
n_j	Fréquence du groupe d_i
n_k	Existance/absence du groupe t_k
nSK	Nombre d'atomes dans la molécule (hydrogènes exclus)
n_{tr}	Nombre de composés dans l'ensemble de calibrage
P	Pression
p	Nombre de descripteurs
Pa	Pascal
P_c	Pression critique
PLS	Moindres carrés partiels
PM3	Parametrized Model 3
PM6	Parametrized Model 6

SYMBOLES ET ABREVIATIONS

PMC	Perceptron multicouches
P_r	Pression réduite
PRESS	Predicted Residual Error Sum of SquaresSomme des carrés des erreurs de prédiction
P_{sat}	Pression de vapeur saturante
P_{vp}	Pression de vapeur
P_{vpr}	Pression de vapeur réduite
Q^2	Coefficient de prédiction interne
Q^2_{BOOT}	Coefficient de prédiction par la technique du bootstrap
Q^2_{EXT}	Coefficient de prédiction externe
$Q^2_{L10\%O}$	Coefficient de prédiction par omission de 10% des observations
$Q^2_{L20\%O}$	Coefficient de prédiction par omission de 20% des observations
q_{ij}	J^{ème} coordonnée cartésienne de l'atome i
$\bar{q}_j$	Moyenne de la j^{ème} coordonnée
QSAR	Quantitative Structure/ Activity Relationships (Relations Quantitatives Structure/ Activité)
QSPR	Quantitative Structure/ Propriety Relationships (Relations Quantitatives Structure/ Propriété)
QSTR	Quantitative Structure/ Toxicity Relationships (Relations Quantitatives Structure/ Toxicité)
R	Matrices influence/distances
R	Constante des gaz parfaits
$R1p+$	Autocorrélation maximale R de distance topologique égale à 1/ pondérée par les polarisabilités atomiques
$R1u+$	Autocorrélation maximale R de distance topologique égale à 1 / non pondérée
$Rkw+$	Autocorrélation maximale R de distance topologique égale à k / pondérée par la propriété atomique employée w
R_C	Coefficient de corrélation critique pour la loi normale
R^2	Coefficient de détermination
R^2_{adj}	Coefficient de détermination ajusté
r	Coefficient de corrélation
r_{ij}	Distance interatomique
RESN	Résidu normalisé
RHF	Restricted Hartree-Fock
RLM	Régression linéaire multiple
RMSE	Racine de l'écart quadratique moyen (Root Mean Squared Error)
RNA	Réseaux de neurones artificiels
S_{ij}	Intégrale de recouvrement

SYMBOLES ET ABREVIATIONS

s	Erreur standard (SE)
SCE	Somme des carrés des écarts
SCT	Somme des carrés totale
s_{jk}	Eléments de la matrice de covariance des coordonnées atomiques pondérées
SVM	Support Vector Machines / Machines à vecteurs support
T	Température
T_b	Température d'ébullition
T_{br}	Température d'ébullition réduite
T_c	Température critique
T_{fp}	Température de fusion
t	t de Student
t_k	Incrément du groupe du 3ème ordre
$\mathbf{t_m}$	Composante m obtenue par analyse en composantes principales de la matrice de covariance des coordonnées atomiques pondérées
T_r	Température réduite
UNIFAC	UNIversal Functional Activity Coefficient / Coefficient d'activité fonctionnel universel
u_i	Incrément du groupe du 1er ordre
V	Volume
v	Volume atomique de van der Waals
V_c	Volume critique
$V_{C\,CAL}$	Valeur calculée du volume critique
$V_{C\,EXP}$	Valeur expérimentale du volume critique
$V_{C\,PRED}$	Valeur prédite du volume critique
V_m	Volume molaire
V_r	Volume réduit
WHIM	Weighted Holistic Invariant Molecular descriptors / Descripteurs moléculaires à invariant holistique pondéré
W_i	Pondération atomique physicochimique
W_{ij}	Poids du réseau de neurone
w	Propriété atomique employée pour pondérer le graphe moléculaire
w_i	Propriété atomique de pondération de l'atome i
w_j	Propriété atomique de pondération de l'atome j
$\bar{w}$	Valeur moyenne de la propriété atomique de pondération dans la molécule
X	Matrice des valeurs des descripteurs de l'ensemble de calibrage
x_i	Variable indépendante
Y_i^{EST}	Valeur estimée de la propriété

SYMBOLES ET ABREVIATIONS

Y_i^{EXP}	Valeur expérimentale de la propriété
y_i	Valeur observée
$\hat{y}i$	Valeur estimée
$\hat{y}_{(i)}$	Valeur prédite
Z	Facteur de compressibilité
Z_C	Facteur de compressibilité critique
°C	Degrés Celsius
α	Niveau de significativité
θ	Paramètre de la méthode de Lydersen
δ	Fonction delta de Dirac
δ_{ij}	Fonction delta de Kronecker
Δ	Somme des deltas de Kronecker
ΔH_v	Enthalpie de vaporisation
η_L	Viscosité liquide
ρ_C	Densité critique
σ^2	Variance
ω	Facteur acentrique
ω_{CAL}	Valeur calculée du facteur acentrique
ω_{EXP}	Valeur expérimentale du facteur acentrique
ω_{PRED}	Valeur prédite du facteur acentrique
$\omega_{PRED/LOO}$	Valeur prédite du facteur acentrique par la procédure LOO
Ψ	Fonction d'onde

LISTE DES TABLEAUX

PARTIE A: BASES THEORIQUES

PARTIE B: APPLICATIONS

LISTE DES FIGURES

PARTIE A: BASES THEORIQUES

PARTIE B: APPLICATIONS

TABLES DES MATIERES

INTRODUCTION GENERALE

Il peut sembler, dans le domaine de la chimie et du génie chimique que les données importantes telles que les pressions de vapeurs, enthalpies de vaporisations, densités, capacités calorifiques ... etc., nécessaires pour la conception des procédés sont facilement disponibles. Cependant, quand la littérature est consultée, souvent très peu ou pas de données peuvent être trouvées. Cela devient donc le travail de l'ingénieur ou du chimiste d'estimer ces données au meilleur de sa connaissance. En conséquence, beaucoup de corrélations utiles et relativement précises ont été développées pour prévoir les propriétés mentionnées ci-dessus. Le problème est que la plupart de ces corrélations (en particulier, les corrélations basées sur les principes des états correspondants) exigent la connaissance du point critique du composé, quoique les propriétés près du point critique ne soient, la plupart du temps, pas nécessaires pour l'application pratique. Le point critique sert comme point de référence le plus généralement utilisé dans les méthodes des états correspondant.

Les propriétés critiques (c.-à-d. température critique, pression critique et volume critique V_C , sont d'une grande importance pratique car ils doivent être connus pour êtres utilisés dans les corrélations basées sur la loi des états correspondants. Cependant, il y a là aussi un manque de données des propriétés critiques dans littérature car ces données sont difficiles ou souvent impossibles à mesurer.

Avec les propriétés critiques, le facteur acentrique ω est une des constantes des composés purs utilisé généralement pour l'estimation des propriétés. La détermination de ses valeurs se fait à partir des données expérimentales de la pression de vapeur et des paramètres des points critiques. Le facteur acentrique représente la non sphéricité d'une molécule et de ce fait, il est très utilisé pour la détermination de nombreuses propriétés thermodynamiques (facteur de compressibilité, pression de vapeur (Ambrose et Patel, 1984), enthalpie de vaporisation, coefficients de l'équation du viriel) et dans les études des équilibres de phases des substances (Poling et al., 2001; Gharagheizi et al., 2015).

Les données des propriétés critiques et du facteur acentrique sont généralement disponibles uniquement pour les petites molécules de stabilité thermique suffisante et la synthèse des produits d'intérêts, suffisamment pur, et les mesures de ces données sont coûteuses et prennent du temps. Dans de nombreux cas, les produits chimiques se dégradent

ou sont dangereux à manipuler, ce qui rend les mesures expérimentales difficiles ou impossibles. Par conséquent, les méthodes d'estimation sont d'une grande valeur.

Parmi les méthodes d'estimations les plus utilisés on trouve les modèles basés sur les relations quantitatives structure-propriété (QSPR pour Quantitative Structure-Property Relationship) qui sont des relations déduites, par différentes approches, entre la structure chimique représentée par des descripteurs moléculaires et les propriétés à prédire (Gharagheizi F. & Mehrpooya M., 2008). Nous avons aussi celle basées sur la structure moléculaires mais appartenant à la classe des méthodes empiriques et sont également connues en tant que méthodes d'addition de groupe se sont les méthodes de contribution de groupes (MCG) (Qiang W & al., 2012; Gharagheizi F et al., 2011).

Le présent travail vise à construire des modèles QSPR fiables pouvant prédire les volumes critiques et les facteurs acentriques d'une large gamme de substances chimiques. Les résultats (les prédictions) de ces modèles, une fois construits et évalués, seront comparés aux résultats de différentes méthodes de contribution de groupes.

Le mémoire réunissant les résultats auxquels nous avons aboutis est séparé en deux grandes parties:

- PARTIE A: BASES THEORIQUES: où, après un rappel sur la thermodynamique, les méthodes de contribution de groupes sont définies et les différents modèles MCG passés en revus. Nous avons aussi détaillé les démarches nécessaires à la construction et l'évaluation de la qualité des modèles QSPR tels que: l'optimisation de la géométrie des molécules, la régression linéaire multiple (MLR), les réseaux de neurones artificiels (RNA) et les statistiques y afférentes.
- PARTIE B: APPLICATIONS: où nous avons appliqué les techniques précédemment mentionnées pour l'obtention du modèle du volume critique par une QSPR linéaire (MLR) et du facteur acentrique par deux modèles QSPR, un linéaire et l'autre établi par RNA. Pour les deux grandeurs modélisées (ω et V_C) une comparaison avec les MCG a été faite.

Enfin une conclusion générale résumant l'essentiel des résultats obtenus et une annexe clôturent ce mémoire.

PARTIE A: BASES THEORIQUES

I.1 Propriétés des corps purs:

Bien que les chimistes et ingénieurs chimistes traitent normalement des mélanges, les propriétés des composés pures sont à la base d'une grande partie du comportement observé. Par exemple, les modèles de propriétés destinés à la gamme entière de composition doivent donner les propriétés des composés pures aux limites de la composition pure. En outre, les propriétés constantes des composés pures sont employées souvent comme base pour des modèles tels que des corrélations des états correspondants pour des équations d'état (*P,V,T*). Elles sont aussi souvent employées dans les règles dépendantes des compositions des mélanges pour les paramètres décrivant ces mélanges (Poling *et al.*, 2001).

I.1.1 Surface d'état:

Un corps pur peut exister sous trois phases: *solide, liquide, vapeur.* La figure 1 donne le nom des différentes transitions de phase possibles entre les états solide, liquide et gazeux.

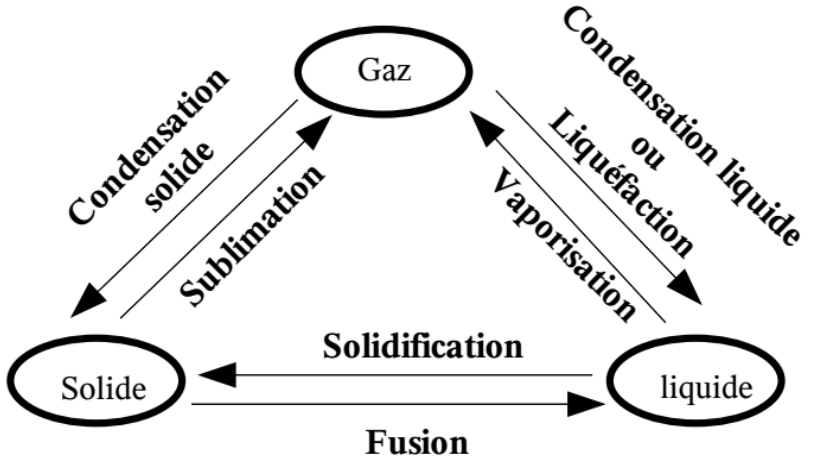


Figure 1: Transitions entre les états solide, liquide et gazeux.

Ces phases n'existent que pour certaines valeurs de la pression *P*, la température *T* et le volume *V*. La représentation de ces états (fig. 2) s'effectue en généralisant le tracé des isothermes, chaque isotherme étant contenue dans un plan (Bruhat, 1933).

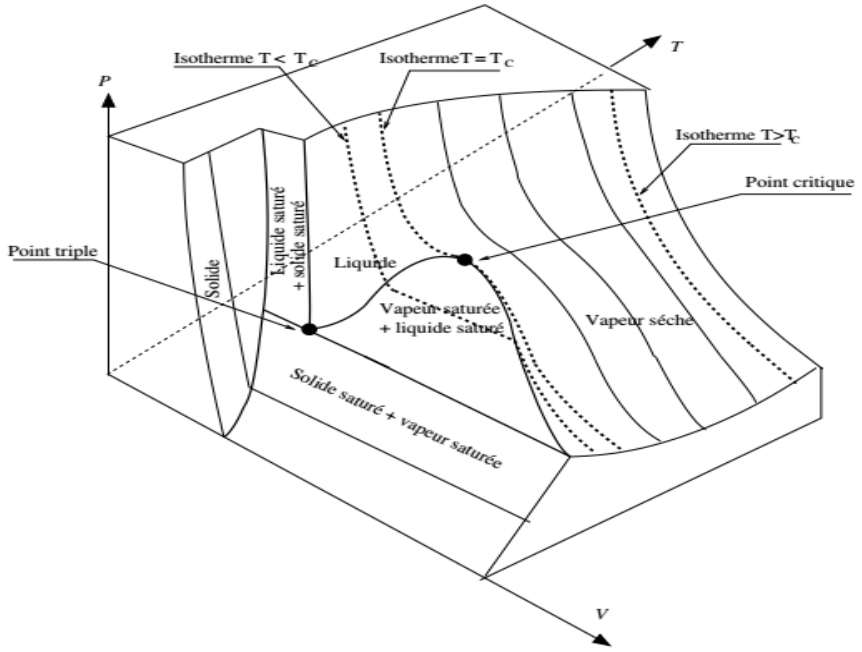


Figure 2: Représentation des trois phases d'un corps pur

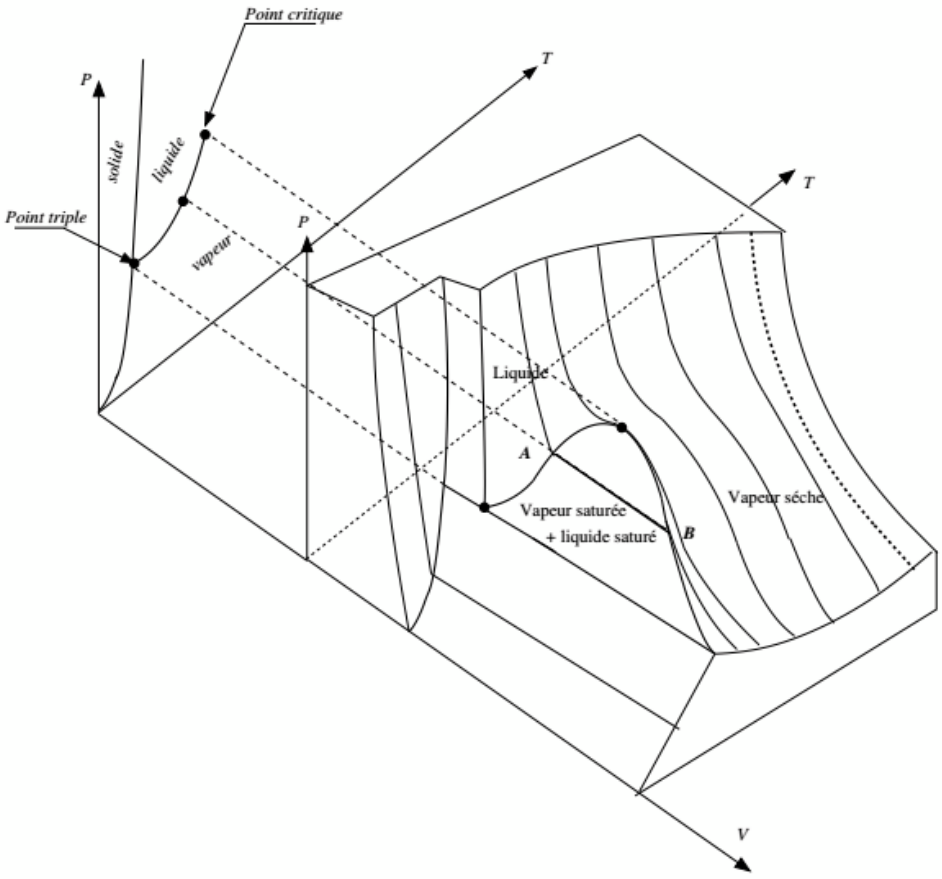


Figure 3: Représentation d'un état d'équilibre liquide-vapeur et projection dans le plan (P; T)

Considérons un mélange liquide-vapeur à la température *T*. Les différents états de cet équilibre sont représentés par le segment *AB* (fig. 3), dont la projection dans le plan (*P; T*) donne un point: à chaque température d'un mélange liquide-vapeur, il ne peut correspondre qu'une pression.

I.1.2 Pression de vapeur saturante:

A une température *T* donnée, l'équilibre d'un liquide et de sa vapeur n'est possible que sous une pression *P* qui est la *pression de vapeur saturante*. La courbe représentant la *pression de vapeur saturante* en fonction de la température, distingue les états du fluide:

- si $P < P_{sat}$ le fluide est dans l'état gazeux;
- si $P > P_{sat}$ le fluide est dans l'état liquide;
- si $P = P_{sat}$ le fluide est en équilibre liquide-vapeur.

I.2 Caractéristiques du point critique:

La courbe de pression de vapeur s'arrête brusquement au point critique (phénomène qui a été en premier découvert en 1822 par De La Tour (1822; 1823); où pression, volume et température sont dit critiques): la liquéfaction du gaz ne peut s'observer qu'en dessous de la température critique T_C et c'est Andrews (1869) qui a découvert les conditions essentiels pour la liquéfaction des gaz.

Ultérieurement, des recherches ont menés au concept que chaque gaz a une température, au-dessus de laquelle il ne peut pas être liquéfié indépendamment de la pression appliquée. Ceci a mené à la découverte du point critique pour lequel la température critique (T_C) est définie comme la température minimum d'un gaz à laquelle il ne peut être liquéfié quelque soit la pression même très haute. La pression critique (P_C) (pression de vapeur) est la plus basse pression qui liquéfiera le gaz à sa température. Le volume molaire critique (V_C) est le volume d'une mole de la substance à la température critique et à la pression critique. La pression critique, le volume critique et la température critiques sont les valeurs de la pression, volume molaire et de la température thermodynamique (Kelvin) à laquelle les densités des phases coexistantes liquides et gazeuses deviennent identiques. Le facteur de compressibilité critique (Z_C) peut être calculé à partir de l'équation 1. D'autres définitions incluent également la densité critique (ρ_C), qui est directement calculée du volume critique (éq. 2).

$$Z_C = \frac{P_C V_C}{R T_C} \quad (1)$$

$$\rho_C = \frac{M}{V_C} \quad (2)$$

Une excellente analyse du point critique, des appareils expérimentaux et des corrélations est fournie par Kobe et Lynn (1953).

Pour l'eau (fig. 4); au point où les trois phases coexistent à l'équilibre thermodynamique (point triple): T = 0,01 °C, P = 0,006 bar; et au point critique: T = 374 °C, P = 218 bar. (Smith *et al.*, 2005).

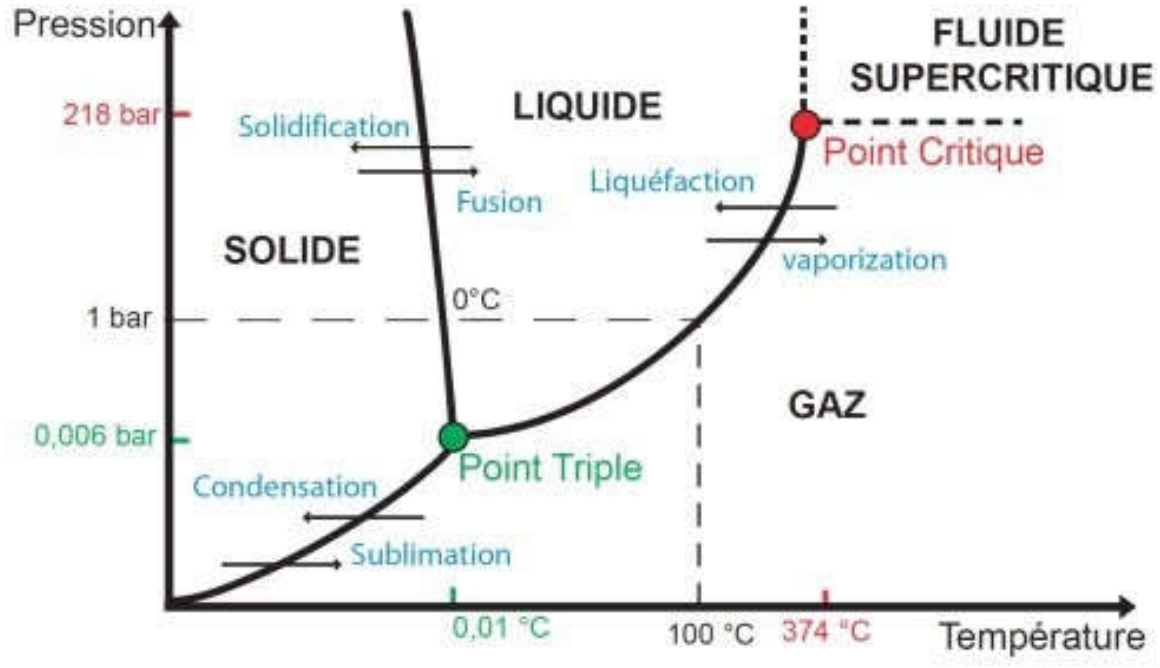


Figure 4: Représentation de l'équilibre liquide-vapeur de l'eau dans le plan (P; T)

Le tableau 1 (Foussard *et al.*, 2010) donne les coordonnées du point critique de quelques corps pur.

Tableau 1. Point critique de quelques corps purs.

	N_2	**O_2**	**CO_2**	**He**
T_C *(K)*	126,2	155	304,2	5,2
P_C *(bar)*	34	51	74	2,3

L'existence d'un point critique pour la transition liquide-vapeur fait qu'il est possible de convertir une vapeur en liquide (et réciproquement) sans transition de phase. Il faut effectuer les opérations suivantes:

- apport de chaleur isochore, qui a pour effet d'augmenter la température (et donc la pression), jusqu'à atteindre une température supérieure à la température critique;
- compression isotherme, de façon à descendre sous le volume du point critique;
- retrait de chaleur isochore, pour amener la température du système sous la température critique. Ces trois opérations permettent de contourner le point critique.

Le système reste donc monophasique du début à la fin: le passage de la vapeur au liquide s'effectue de façon continue.

I.3 Equations d'état du corps pur:

Soit un seul corps pur, constituant une seule phase fluide. Cela indique qu'il existe deux variables que l'on peut se donner indépendamment pour caractériser l'état du fluide, par exemple la température et la pression. Toutes les autres variables sont fonctions des deux qui ont été initialement choisies. Si on considère, par exemple, le volume occupé par une quantité déterminée de substance, on a une relation $f(P, V, T) = 0$ qui est l'équation d'état.

I.3.1 Le gaz parfait:

Pour un nombre n de moles de gaz, on a:

$$\boldsymbol{PV = nRT} \tag{3}$$

avec R constante molaire des gaz, égale à environ 8,315 J $mol^{-1}K^{-1}$. Une méthode simple d'application de la formule (1) consiste à se rappeler que dans les conditions normales de température et de pression (T = 273 K, P = 1 atm = 101 325 Pa) le volume occupé par 1 mol de gaz parfait est 22,4 L. Du point de vue pratique, l'équation des gaz parfaits donne la valeur de la variable d'état à calculer avec une approximation d'autant meilleure que l'on se trouve loin du point critique du fluide. Elle donne, le plus souvent du moins, des ordres de grandeur acceptables. L'équation (3) a le grand mérite de la simplicité.

I.3.2 Equation d'état d'un gaz réel:

La fonction $f(P, V, T) = 0$ est différente de celle du gaz parfait. La figure 5 donne la représentation de quelques isothermes (P; V) pour 1 mol de fluide réel CO_2 en coordonnées

dites de Clapeyron (ou d'Andrews). Plusieurs caractéristiques du fluide réel sont apparentes sur cette figure:

- l'isotherme P (V) n'est continue qu'aux températures supérieures à la température critique (31 °C) ;
- il existe un écart considérable aux pressions élevées entre l'isotherme expérimentale et celle qui est calculée d'après la loi des gaz parfaits (exemple à 57,8 °C = 331 K) ;
- en-dessous de la température critique apparaissent le domaine du liquide et le domaine diphasique liquide-vapeur.

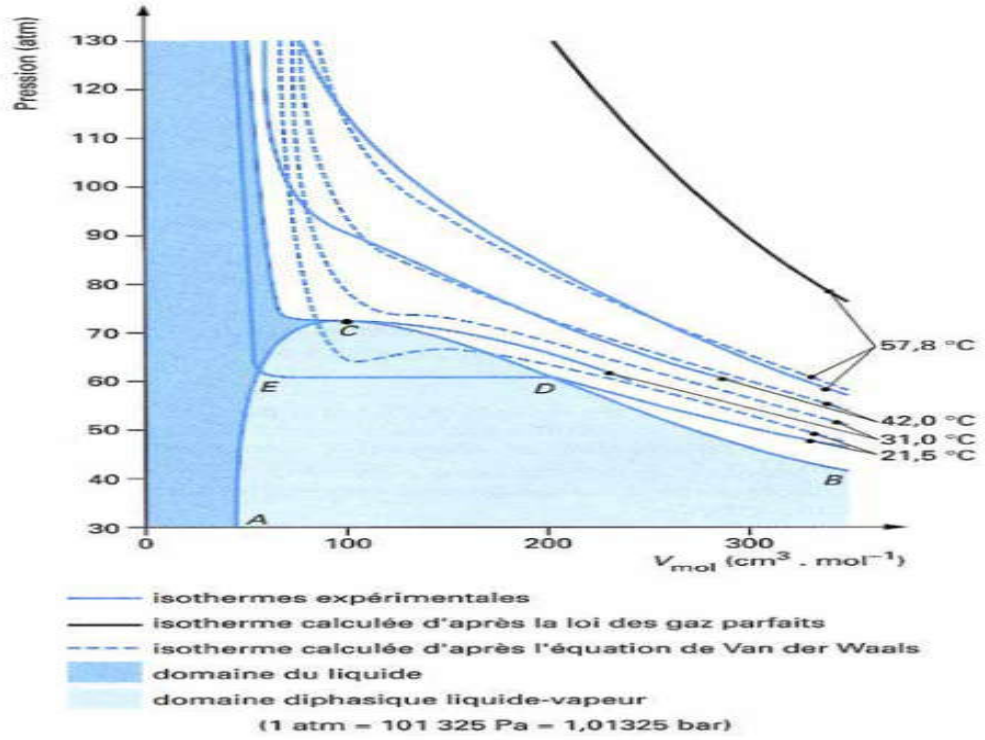


Figure 5: Réseau d'isothermes du dioxyde de carbone en coordonnées (P, V) (Hougen et al., 1962)

I.3.3 Equation de Van der Waals:

Elle s'écrit pour n moles de fluide:

$$\left(P+\frac{n^2a}{V^2}\right)(V-nb)=nRT \qquad (4)$$

avec $\frac{a}{V^2}$ (pour 1 mol) pression interne et b le covolume.

a et **b** sont des paramètres indépendants des variables d'état, mais dépendants de la nature du fluide. Le tableau 2 donne quelques valeurs de ces paramètres.

L'équation de Van der Waals représente une approximation meilleure que celle des gaz parfaits dans des conditions proches de la liquéfaction ainsi que le montre la figure 5. Sa relative simplicité (bien qu'elle soit du 3ème degré en V) permet un usage courant. On remarque qu'elle peut s'exprimer en variables réduites (éq. 5). En effet, en précisant qu'au point critique: $f\left(P_C, V_C, T_C\right) = 0; \left(\frac{\partial P}{\partial V}\right)_C = 0; \left(\frac{\partial^2 P}{\partial V^2}\right)_C = 0$

On trouve les relations:

$$a = \frac{9}{8} R T_C V_C;\ b = \frac{V_C}{3};\ Z_C = \frac{P_C V_C}{R T_C} = \frac{3}{8} \tag{5}$$

Il est à noter que le **facteur de compressibilité critique** Z_c est très supérieur à ce que donne l'expérience (≈ 0,27).

Tableau 2. Constantes a et b de l'équation de Van der Waals pour quelques fluides (James et Lord, 1992).

Fluide	$a \times 10$ (Pa m^6 mol^{-2})	$b \times 10^5$ (m^3 mol^{-1})
Ammoniac	4,225	3,707
Argon	1,363	3,219
Azote	1,408	3,913
Dioxyde de carbone	3,640	4,267
Dioxyde de soufre	6,803	5,636
Eau	5,536	3,049
Éthane	5,562	6,380
Hélium	3,457 × 10^{-2}	2,370
Hydrogène	0,2476	2,661
Krypton	2,349	3,978
Méthane	2,283	4,278
Monoxyde de carbone	1,505	3,985
Néon	0,2135	1,709
Oxygène	1,378	3,183
Propène	8,490	8,272
Xénon	4,250	5,105

Une autre conséquence de l'équation de Van der Waals est la possibilité d'obtenir des pressions négatives. Par exemple, pour 1 mol d'eau à 298 K dans un volume de 4 × 10^{-2} L, les

données du tableau 2 permettent de calculer $P \approx -842$ atm ≈ -842 bar. Expérimentalement, des pressions négatives de l'ordre de – 100 atm (≈ – 100 bar) ont été observées pour l'eau.

I.3.4 Loi des états correspondants:

Proposé par Van der Waals en 1873, la loi des états correspondants exprime que la généralisation des propriétés d'équilibre qui dépendent de certaines forces intermoléculaires sont liées aux propriétés critiques d'une manière universelle. Les états correspondants constituent la base la plus importante et la plus simple pour l'élaboration des corrélations et des méthodes d'estimations. En 1873, Van der Waals a montrée qu'elle était théoriquement valide pour toutes les substances pures dont les propriétés *PVT* pouvaient être exprimées par une équation d'état à deux constantes telle que l'équation 4.

On définit les **variables d'état réduites** comme les rapports des valeurs de ces variables aux valeurs dans l'état critique: $\boldsymbol{P_r = \frac{P}{P_C};\ T_r = \frac{T}{T_C};\ V_r = \frac{V}{V_C}}$

Deux fluides sont dits dans des **états correspondants** si deux de leurs **variables d'état réduites** sont les mêmes. La loi des états correspondants établit que toute autre variable réduite dépendant de l'équation d'état est alors la même. Autrement dit, l'équation d'état entre variables réduites $f(P_r, V_r, T_r) = 0$ est indépendante de la nature du fluide. Cette loi est expérimentale, et n'a donc qu'une valeur approchée, mais est cependant plus générale que d'autres équations d'état.

En utilisant les coordonnées réduites, l'équation 4 devient:

$$\left(P_r + \frac{3}{V_r^2}\right)\left(V_r - \frac{1}{3}\right) = \frac{8}{3}T_r \qquad (6)$$

Dans cette relation n'intervient plus aucune constante relative au fluide, comme le veut la loi des états correspondants.

I.3.5 Autres équations d'état:

De très nombreuses équations d'état algébriques ont été proposées. Ce nombre même prouve qu'aucune d'elles n'est totalement satisfaisante. Certaines font intervenir de nombreux paramètres. Parmi les équations d'état à deux paramètres les plus couramment utilisées (pour 1 mol), on peut citer:

- l'**équation de Peng et Robinson: (**Peng et Robinson, 1976)

$$\left[P+\frac{a(T)}{V(V+b)+b(V-b)}\right](V-b)=RT \tag{7}$$

- **l'équation de Redlich-Kwong: (**Redlich et Kwong, 1949).

$$\left[P+\frac{a}{\sqrt{T}V(V+b)}\right](V-b)=RT \tag{8}$$

a, *b* fonctions uniquement du fluide ;

- **l'équation de Redlich-Kwong-Soave:** (Soave, 1972)
  Représente une généralisation de la relation (8):

$$\left[P+\frac{a(T)}{V(V+b)}\right](V-b)=RT \tag{9}$$

Les équations (7), (8) et (9) peuvent être considérées comme s'apparentant à la relation (4) de Van der Waals par complication du terme de pression interne. On notera cependant qu'elles restent cubiques par rapport au volume. Dans le cas où *a* et *b* ne dépendent que de la nature du fluide, il est relativement facile de se livrer au même exercice qu'avec l'équation de Van der Waals, c'est-à-dire de chercher à exprimer ces paramètres en fonction des variables critiques. C'est ainsi que dans le cas de (8), on trouve les relations (Reid *et al.*, 1987):

$$a=0,4275\frac{R^2T_C^{2,5}}{P_C};\ b=0,0866\frac{RT_C}{P_C};\ Z_C=\frac{P_CV_C}{RT_C}=0,33 \tag{10}$$

à comparer avec les équations (5). On notera en particulier la valeur améliorée du facteur de compressibilité Z_C.

I.3.6 Equations du viriel:

On peut prendre le facteur de compressibilité:

$$Z=\frac{PV}{RT} \tag{11}$$

comme fonction d'état soit de *P*, soit de 1/*V*. Pour 1 mol de gaz parfait *Z* est évidemment égal à 1. Pour 1 mol de gaz réel, il s'écarte de cette valeur selon un développement limité soit de *P*, soit de 1/*V*. Ce développement s'appelle la **formule du viriel** et les coefficients A_i: A_1 , A_2 , A_3

...etc. dans le cas de $1/V$ (éq.12); ou B_i: B_1 , B_2 , B_3 ...etc. (cas de P (éq.13)) des différents termes, les coefficients du viriel.

$$Z = 1 + \frac{A_1(T)}{V} + \frac{A_2(T)}{V^2} + \frac{A_3(T)}{V^3} + \ldots \quad (12)$$

$$Z = 1 + B_1(T)P + B_2(T)P^2 + B_3(T)P^3 + \ldots \quad (13)$$

Le gros avantage de ces développements est de permettre (évidemment lorsque les coefficients sont connus en fonction de la température (Dymond et Smith, 1980)) de faire les calculs en les limitant à la précision voulue. Le tableau 3 **(**Claudel, 1996**)** donne les valeurs en fonction de la température du coefficient A_1 pour les mêmes gaz que ceux du tableau 2. La figure 6 (Nelson et Obert, 1954) montre les isothermes de Z en fonction de la pression réduite P_r à différentes températures réduites T_r. On notera que la logique voudrait que soit porté en ordonnée le rapport. Mais comme Z_C doit être une constante indépendante de la nature du constituant, la loi des états correspondants reste valable avec Z. Beaucoup a été écrit au sujet de l'équation du viriel ; voir particulièrement Mason et Spurling (1968) et Dymond et Smith (1980).

Tableau 3. Premier coefficient du viriel pour quelques gaz.

Gaz	α*	β*	γ*	Domaine de température T (K)
Ammoniac	44,3	23,6	766,6	273 à 573
Argon	154,2	119,3	105,1	80 à 1 000
Azote	185,4	141,8	88,7	75 à 700
Dioxyde de carbone	137,6	87,7	325,7	220 à 1 100
Dioxyde de soufre	134,4	72,5	606,5	265 à 473
Eau	33,0	15,2	1 300,7	293 à 1 250
Éthane	311,7	230,6	227,8	200 à 600
Hélium	114,1	98,7	3,25	7 à 150
Hydrogène	315,0	289,7	9,47	14 à 400
Krypton	189,6	148,0	145,3	110 à 700
Méthane	206,4	159,5	133,0	110 à 600
Monoxyde de carbone	202,6	154,2	94,2	90 à 573
Néon	81,0	63,6	30,7	44 à 973
Oxygène	152,8	117,0	108,8	90 à 400
Xénon	247,0	192,9	199,8	160 à 650

$$* A\left(cm^3 \cdot mol^{-1}\right) = \alpha - \beta \exp\left(\frac{\gamma}{T}\right)$$

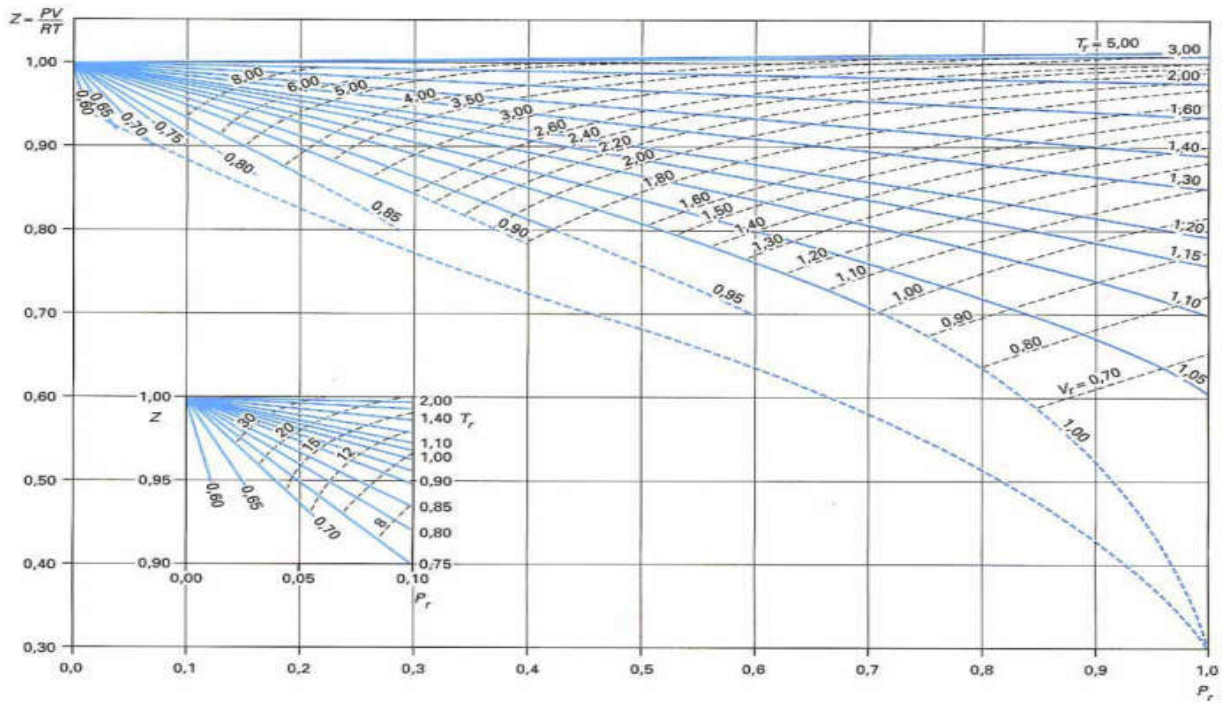


Figure 6: Facteur de compressibilité Z en fonction de la pression réduite P_r à différentes températures réduites T_r

Comme vue précédemment, la température (T_C), la pression (P_C) et le volume (V_C), critiques sont des constantes des composés pures de très grand intérêt pour l'équilibre vapeur-liquide. Elles sont aussi employées dans beaucoup de corrélations des propriétés des gaz et des liquides:

- Propriétés volumétriques: la corrélation de Tsonopoulos (1974) (éq. (14)) pour l'estimation du deuxième coefficient du viriel (A_1 dans l'équation (12)) en est un exemple.

$$\frac{A_1 P_C}{RT_C} = f^{(0)} + \omega f^{(1)} + a f^{(2)} + b f^{(3)} \qquad (14)$$

où ω est le facteur acentrique, T_C et P_C la température et la pression critiques respectivement, avec:

$$f^{(0)} = 0,1445 - 0,330/T_r - 0,1385/T_r^2 - 0,0121/T_r^3 - 0,000607/T_r^8 \qquad (15a)$$

$$f^{(1)} = 0,0637 + 0,331/T_r^2 - 0,423/T_r^3 - 0,008/T_r^8 \qquad (15b)$$

$$f^{(2)} = 1/T_r^6 \tag{15c}$$

$$f^{(3)} = -1/T_r^8 \tag{15d}$$

ici T_r est la température réduite.

- Propriétés thermodynamiques: comme la méthode de Riedel (1954) pour le calcul de l'enthalpie de vaporisation (ΔH_v) par l'équation suivante:

$$\Delta H_v = 1,093 R T_C T_{br} \frac{\ln P_C - 1,013}{0,930 - T_{br}} \tag{16}$$

avec P_C la pression critique et T_{br} la température d'ébullition réduite qui est le résultat de la division de la température d'ébullition par température critique.

- Propriétés de transport: par exemple Przezdziecki et Sridhar (1985) ont proposé une méthode de calcul de la viscosité liquide (η_L) à basse température en utilisant l'équation:

$$\eta_L = \frac{V_0}{E\left(V_m - V_0\right)} \tag{17}$$

où V_m est le volume molaire du liquide et les paramètres V_0 et E sont définis comme suit:

$$E = -1,12 + \frac{V_C}{12,94 + 0,10M - 0,23P_C + 0,0424T_{fp} - 11,58\left(T_{fp}/T_C\right)} \tag{18a}$$

$$V_0 = 0,0085\omega T_C - 2,02 + \frac{V_m}{0,342\left(T_{fp}/T_C\right) + 0,894} \tag{18b}$$

pour les équations (18a) et (18b): V_C est le volume critique, M la masse moléculaire et V_m le volume molaire du liquide à la température de fusion T_{fp}.

La détermination expérimentale des valeurs des constantes critiques peut être un chalenge voir impossible (Ambrose et Young, 1995; Nannoolal *et al.*, 2007), particulièrement pour les plus gros composés qui peuvent se dégrader à leurs températures critiques très élevées (Anselme et Teja, 1990; Skander et Chitour, 2007).

I.3.7 Le facteur acentrique:

Avec les propriétés critiques, le facteur acentrique ω est une des constantes des composés purs utilisé généralement pour l'estimation des propriétés. Il a été à l'origine défini par Pitzer (1955). La détermination de ses valeurs se fait à partir des données expérimentales de la pression de vapeur et des paramètres des points critiques en utilisant l'équation:

$$\omega = -\log P_{vpr} - 1 \tag{19}$$

dans laquelle P_{vpr} est la pression de vapeur réduite ($P_{vpr} = P_{vp} / P_C$), sachant que la pression de vapeur P_{vp} est mesurée pour une température réduite égale à 0,7.

Pour les gaz monoatomiques, ω est essentiellement nul, et pour le méthane sa valeur est encore très petite. Cependant, ω croît avec la masse moléculaire des hydrocarbures, de même qu'avec la polarité et ses valeurs (toutes positives) peuvent aller jusqu'à 1,5.

Une des méthodes fiables décrite dans la littérature (Poling *et al.*, 2001) pour l'estimation de ω est résumée par l'équation suivante:

$$\omega = -\frac{\ln\left(P_C / 1{,}01325 + F^{(0)}\left(T_{br}\right)\right)}{F^{(1)}\left(T_{br}\right)} \tag{20}$$

où les fonctions $F^{(0)}$ et $F^{(1)}$ sont :

$$F^{(0)}\left(T_{br}\right) = \frac{-5{,}97616\tau + 1{,}29874\tau^{1{,}5} - 0{,}60394\tau^{2{,}5} - 1{,}06841\tau^{5}}{T_{br}} \tag{21a}$$

$$F^{(1)}\left(T_{br}\right) = \frac{-5{,}03365\tau + 1{,}11505\tau^{1{,}5} - 5{,}41217\tau^{2{,}5} - 7{,}46628\tau^{5}}{T_{br}} \tag{21b}$$

avec $\tau = \left(1 - T_{br}\right)$.

Le facteur acentrique représente la non sphéricité d'une molécule. De ce fait, le facteur acentrique est très utilisé pour la détermination de nombreuses propriétés thermodynamiques (facteur de compressibilité (éq. (22), pression de vapeur (Ambrose et Patel, 1984), enthalpie de vaporisation (éq. (16), coefficients de l'équation du viriel (éq. (14)) et dans les études des équilibres de phases des substances (Poling *et al.*, 2001; Gharagheizi *et al.*, 2015).

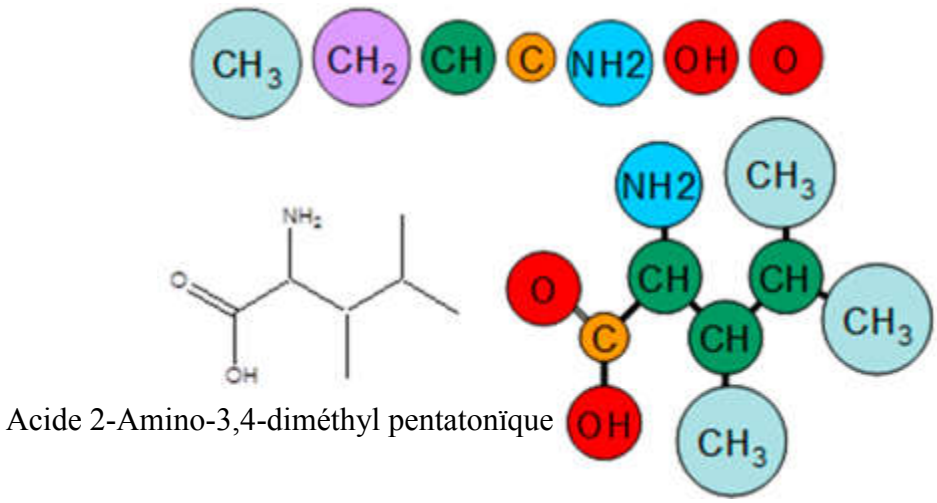
$$Z_C = 0,291 - 0,080\omega \quad (22)$$

Si, à présent, ω est très largement utilisé pour caractériser la complexité d'une molécule du point de vue de la géométrie et de la polarité (Poling *et al.*, 2001), les grandes valeurs du facteur acentrique de certains composés polaires (ω >0,4) ne sont pas significatives dans l'acception originelle de cette propriété.

Comme montré par Liu et Chen (1996), la sensibilité aux erreurs dans les informations d'entrées lors de l'estimation des propriétés est très grande. La procédure recommandée pour obtenir une valeur inconnue du facteur acentrique en utilisant directement l'équation (19) en employant une corrélation très précise pour la pression de vapeur tel que l'équation de Wagner (1973, 1977) et sa modification par Ambrose et Ghiassee (1987) et pour de hautes pressions une extension de l'équation d'Antoine (1888) proposée par le Thermodynamics Research Center pour quelques intervalles de température (TRC, 1999). Une autre approche plus fiable est d'employer des valeurs expérimentales précises de T_C, P_C et T_b dans l'équation (20).

II.1 Définition:

Les méthodes de contribution de groupes (MCG) appartiennent à la classe des méthodes empiriques et sont également connues en tant que méthodes d'addition de groupe. La supposition à la base de cette méthode est qu'une propriété d'une molécule peut être décrite comme la somme des contributions venant de différents blocs représentant différents groupes fonctionnels (Figure 7).



Acide 2-Amino-3,4-diméthyl pentatonïque

Figure 7: Représentation schématique d'une molécule fragmentée en ses blocs constitutifs.

Les MCG peuvent être classées selon le nombre et l'ordre des groupes utilisés dans les modèles de corrélation. Elles sont ainsi classées sous différents ordres: premier, deuxième, ou troisième. Le premier ordre décrit les groupes fonctionnels simples dans les molécules organiques, tandis que les deuxième et troisième ordres permettent une description beaucoup plus détaillée de la structure des molécules comprenant d'autres groupes et fragments qui ne sont pas décrits dans le premier ordre. Les deuxième et troisième ordres agissent en tant qu'une correction de la déviation des paramètres fournis dans le premier ordre afin d'améliorer la précision de la méthode. Une propriété peut être estimée par une somme de termes ou de contributions représentant les différents groupes présents dans une molécule comme suit:

$$\textbf{Propriété} = F\left[\sum_i n_i u_i + w\left(\sum_j n_j d_j\right) + z\left(\sum_k n_k t_k\right)\right] \qquad (23)$$

Dans cette équation; n_i et n_j sont les fréquences des groupes u_i et d_i respectivement. Ces derniers sont aussi désignés comme incréments dont les valeurs sont calibrées en utilisant des données expérimentales de la propriété. Si le fragment k (t_k son incrément et sa valeur est

calibré comme u_i et d_i) est présent dans la molécule; n_k prend la valeur 1 sinon 0. A la constante w est assigné une valeur de zéro pour une approximation du premier ordre et d'unité dans le cas de l'approximation du deuxième ordre où les deux (premier et deuxième ordre) contributions de groupes sont impliquées. Pour w=1 et si le $3^{ème}$ niveau d'approximation (groupes du troisième ordre) est utilisé z devrait prendre 1 comme valeur; sinon 0. Enfin, la fonction F est choisie par le concepteur de la méthode et nous verrons par la suite que beaucoup de chercheurs ont adopté des fonctions similaires.

Depuis les premières élaborations des méthodes de contribution de groupes par Riedel (1949) et Lydersen (1955), beaucoup de méthodes ont été développées pour l'estimation des propriétés critiques et du facteur acentrique (Qiang *et al.*, 2012; Gharagheizi *et al.*, 2011; Nannoolal, 2006; Turner *et al.*, 1998).

Le tableau 4 donne une vue d'ensemble sur la chronologie des plus importants développements des méthodes de contribution de groupes pour les propriétés critiques et le facteur acentrique. Il est essentiel de mentionner que la méthode de Constantinou et Gani (1994) est la seule qui présente (mais dans une version modifiée de 1995 (Constantinou *et al.*, 1995) une tentative pour la prédiction du facteur acentrique pour un grand nombre de composés.

Tableau 4. Quelques méthodes de contribution de groupes pour les propriétés critiques et le facteur acentrique.

Méthode et référence	Code
Lydersen (1955)	Ly55
Joback et Reid (1987)	J&R87
Constantinou et Gani (1994) Constantinou *et al.* (1995)*	C&G94
Marrero-Morejon et Pardillo-Fontdevilla (1999)	M&P99
Marrero-Morejon et Gani (2001)	M&G01

* pour le facteur acentrique

En raison de leur importance pratique et théorique, l'estimation des propriétés critiques et du facteur acentrique a beaucoup suscité l'intérêt des chercheurs à travers le monde. En plus des méthodes données dans le tableau 4, de nombreuses publications scientifiques traitent de l'utilisation des corrélations QSPR (Quantitative Structure Property Relationships) pour

l'estimation des propriétés critiques ainsi que le facteur acentrique. La puissance de ces dernières approches a été démontrée dans beaucoup de cas. (Carande *et al.*, 2015; Sola *et al.*, 2008).

II.2 Riedel (1949) et Lydersen (1955):

Guldberg (1890) était le premier à observer que la température critique peut être approximée par l'équation 24, qui peut également être désignée comme la *règle de Guldberg*:

$$T_C = 1,5T_b \tag{24}$$

Riedel (1949) et Lydersen (1955) avaient proposé des modifications de la règle de Guldberg, sous la forme de l'équation 25.

$$T_C = \frac{T_b}{\theta} \tag{25}$$

La valeur du θ est généralement différente pour chaque composé et peut être calculée en sommant les contributions structurales. Riedel a proposé 22 groupes simples du premier ordre employant l'équation 26 pour calculer θ pour l'estimation de la température critique (éq. 25). Pour la pression critique Riedel a employé l'équation 27.

$$\theta = 0,574 + \sum_i n_i u_i \tag{26}$$

$$P_C = \frac{M}{0,33 - \sum \left(n_i u_i\right)^2} \tag{27}$$

Lydersen (1955) a étendu la méthode de Riedel en incorporant un plus grand ensemble de groupes et de données expérimentales. Lydersen a également proposé une équation quadratique pour estimer θ (éq. 28) pour l'estimation de la température critique. Les équations 27 (avec 0,34 au lieu de 0,33) et 29 sont employées pour l'estimation de la pression et du volume critique respectivement. Ces deux dernières équations sont devenues les équations standards utilisées par beaucoup d'autres chercheurs.

$$\theta = 0,567 + \sum_i n_i u_i - \left(\sum_i n_i u_i\right)^2 \tag{28}$$

$$V_C = 40 + \sum_i n_i u_i \tag{29}$$

Les méthodes de Riedel et de Lydersen sont parmi les plus anciennes méthodes de contribution de groupes et leur extension basée sur un plus grand ensemble de données a été présentée par Joback et Reid (1987).

II.3 Joback et Reid (1987):

Joback et Reid ont examiné différents types d'équations d'estimation exigeant des contributions de groupes et ont choisi les équations 30, 31 et 32 pour la prédiction de la température, de la pression et du volume critiques respectivement. Ils ont supposé qu'il n'y avait pas d'interactions entre les groupes et de ce fait que les paramètres dépendants de la structure sont déterminés en additionnant la fréquence de chaque groupe multiplié par sa contribution.

$$T_C = \frac{T_b}{0,584 + 0,965\sum_i n_i u_i - \left(\sum_i n_i u_i\right)^2} \tag{30}$$

$$P_C = \frac{1}{\left(0,113 + 0,0032 N_A - \sum_i n_i u_i\right)^2} \tag{31}$$

où N_A est le nombre d'atomes de la molécule.

$$V_C = 17,5 + \sum_i n_i u_i \tag{32}$$

Ils ont employé seulement 41 groupes moléculaires, ce qui simplifie trop la structure moléculaire rendant de ce fait plusieurs types d'isomères non différenciables. De façon générale c'est insuffisant pour capturer les effets structuraux des molécules organiques et c'est la raison principale de la pauvre fiabilité de la méthode. La régression linéaire multiple a été effectuée en utilisant 409, 392 et 310 composés pour la température critique, pression critique et volume critique respectivement afin de calibrer ces 41 groupes. Ces groupes sont semblables à ceux de Lydersen (1955) avec l'inclusion du =N- (cyclique) mais avec l'omission de >Si< et >B-.

Ce n'est pas particulièrement utile dans les estimations des propriétés (notamment dans le cas des propriétés critiques où les ensembles de données sont relativement petits) de prendre les données des petits composés ou des composés qui sont les premiers de leurs séries homologues qui sont habituellement facilement disponibles. Des estimations sont habituellement effectuées pour des composés plus grands, complexes ou multifonctionnels.

Le seul avantage de la méthode est sa simplicité d'emploi, cependant elle est désavantagée par sa gamme de composés relativement petite, son faible pouvoir prédictif et des extrapolations douteuses. Beaucoup d'auteurs ont malgré ça continué le travail de Joback et Reid se servant des groupes identifiés comme point de départ.

II.4 Constantinou et Gani (1994):

Les approximations (contributions) du deuxième ordre ou deuxième niveau sont une manière de fournir les informations supplémentaires sur de la structure moléculaire du composé pour une estimation sensiblement améliorée des propriétés. En même temps, si ces contributions ne sont pas disponibles, une estimation moins précise est encore possible employant seulement les groupes du premier ordre. Constantinou *et al.* (1993), (1994) et Constantinou (1993) ont fourni une méthode additive d'estimation des propriétés basée sur les opérateurs conjugués et applicable aux composés organiques. Cependant, la génération des formes conjuguées est une manière non triviale et exige un environnement de calcul symbolique.

Constantinou et Gani (1994) ont appliqué la méthode de Constantinou *et al.* (1994) basé sur les formes conjuguées du deuxième ordre au concept de contribution de groupes. La méthode a proposé une estimation des propriétés effectuée à deux niveaux. Le niveau de base est des contributions des groupes fonctionnels du premier ordre et le deuxième niveau est des groupes du deuxième ordre qui ont les groupes du premier ordre comme blocs à subdiviser. Ainsi, leur méthode tient compte d'une approximation du premier ordre (utilisant les groupes du premier ordre) et d'une approximation du deuxième ordre plus précise (utilisant les groupes du premier et du deuxième ordre).

Constantinou et Gani ont aussi proposé d'employer en tant que groupes du premier ordre l'ensemble des groupes utilisés généralement pour l'estimation des propriétés des mélanges (ou groupes UNIFAC). Un inconvénient à ce choix est qu'un groupe dans un composé aliphatique cyclique ou non-cyclique est considéré avec la même contribution. Ces

groupes ne peuvent pas distinguer des configurations spéciales telles que les groupes multiples situés près l'un de l'autre, les structures de résonance... etc. Par conséquent, chaque groupe a une contribution simple indépendamment du type de composé impliqué. Au total, il y a 78 groupes du premier d'ordre, semblables à ceux employés par Joback et Reid; et la plupart des nouveaux groupes étant des subdivisions et quelques uns étant aussi bien superflus.

Puisque leur estimation a été principalement basée seulement sur des informations de la structure moléculaire, l'idée était d'inclure un niveau d'approximation différent. Ainsi Constantinou et Gani ont présenté les groupes du deuxième ordre pour fournir des informations structurales additionnelles au sujet du composé. Leur objectif ultime était d'accroître l'exactitude et la fiabilité et d'agrandir le domaine d'applicabilité de l'estimation des propriétés et de prévoir les effets de proximité et les différences entre isomères avec fiabilité.

La méthode utilise une équation logarithmique pour l'estimation de la température critique (équation 33). Le modèle extrapole correctement mais des déviations élevées sont à signaler en conséquence à l'omission du point d'ébullition normal.

Abildskov (1994) a effectué une étude limitée de cette méthode pour environ 100 composés et constaté qu'en incluant des approximations du deuxième ordre que l'estimation était aussi souvent améliorée que dégradée. Excepté les composés cycliques, l'amélioration étaient rarement plus de 1 à 2%. Il a conclu que l'utilisation des groupes du deuxième ordre ne peut pas être toujours valable et qu'il n'y a aucun moyen de savoir quand les employer.

Les modèles de la pression et du volume critiques sont présentés dans les équations 34 et 35 respectivement:

$$T_C = 181,128 \ln\left(\sum_i n_i u_i + w\left(\sum_j n_j d_j\right)\right) \tag{33}$$

$$P_C = \frac{1}{\left(0,10022 + \sum_i n_i u_i + w\left(\sum_j n_j d_j\right)\right)^2} + 1,3705 \tag{34}$$

$$V_C = -0,00435 + \sum_i n_i u_i + w\left(\sum_j n_j d_j\right) \tag{35}$$

285, 269, 251 points de données expérimentales ont été employés dans le calibrage par régression pour la température, la pression et le volume critiques respectivement.

L'équation 36 est le modèle d'estimation du facteur acentrique élaboré par Constantinou *et al.* (1995). Comme signalé au paravent, c'est la seule approche par la méthode de contribution de groupes pour le facteur acentrique. Dans ce modèle, 181 valeurs expérimentales ont été utilisées pour le calibrage.

$$\omega = 0,4085\left(\ln\left(\sum_i n_i u_i + w\left(\sum_j n_j d_j\right) + 1,1507\right)\right)^{(1/0,5050)} \tag{36}$$

II.5 Marrero-Morejon et Pardillo-Fontdevilla (1999):

Pardillo et Gonzalez-Rubio (1997) étaient les premiers à proposer une nouvelle approche structurale appelée la contribution des interactions des groupes (GIC pour Group Interaction Contribution) qui considère la contribution des interactions entre les groupes de liaison au lieu de la simple contribution des groupes. Basé sur l'approche décrite (GIC), Marrero et Pardillo (1999) ont proposé une nouvelle méthode pour estimer les points d'ébullition et les constantes critiques des composés organiques purs.

Marrero et Pardillo ont choisi 39 groupes simples, qui peuvent également être considérés comme des groupes du premier ordre, afin de constituer un ensemble cohérent d'interactions de groupes qui permet de traiter une large variété de composés organiques. Ces groupes sont semblables à ceux de la méthode de Joback et de Reid, précédemment présentée, avec l'omission du =NH et du =N- (non-cyclique). Les équations des modèles sont également celles utilisées par Joback et Reid. Cette duplication des modèles apporte également les mêmes inconvénients, c.-à-d. des résultats d'extrapolations peu fiables à cause de l'utilisation du terme quadratique dans le modèle de la température critique. Les modèles sont comme suit:

$$T_C = \frac{T_b}{0,5851 - 0,9286\sum_i n_i u_i - \left(\sum_i n_i u_i\right)^2} \tag{37}$$

$$P_C = \frac{1}{\left(0,1285 - 0,0059 N_A - \sum_i n_i u_i\right)^2} \tag{38}$$

$$V_C = 25,1 + \sum_i n_i u_i \tag{39}$$

La définition structurale proposée des interactions de groupe devrait être une définition de contributions de liaison parce qu'il n'y a aucune interaction physique entre les groupes mais plutôt une liaison entre deux groupes définis. Quelques contributions de liaison n'ont pas été calculées à cause du manque de valeurs expérimentales des propriétés pour les composés impliqués. En outre, les groupes de la méthode de Joback et Reid, où le domaine d'applicabilité est réduit et où les groupes ont été mal définis, ont été la base pour déterminer les contributions de liaison. En raison de l'approche de contribution de liaisons, le domaine d'applicabilité de la méthode est sévèrement restreint mais fournit cependant une estimation significativement améliorée dans le cas des isomères par rapport à la méthode de Joback et Reid.

II.6 Marrero-Morejon et Gani (2001):

Marrero-Morejon et Gani (2001) ont proposé une nouvelle méthode de contribution de groupes basée sur trois niveaux d'approximations. Le premier (prévu pour traiter les composés simples et monofonctionnels), avec un grand ensemble de groupes simples peut capturer partiellement les effets de proximité mais ne peut pas distinguer les isomères. Le deuxième niveau permet une meilleure description des composés polyfonctionnels et une différentiation des isomères. Cependant, les groupes du deuxième ordre ne peuvent pas fournir une bonne représentation des composés contenant plus d'un cycle et dans certains cas des composés polyfonctionnels à chaîne ouverte de plus de quatre atomes de carbone dans la chaîne principale. Un autre niveau est donc exigé pour fournir une meilleure description pour ces types de composés. Le troisième niveau permet l'estimation des gros composés polyfonctionnels non cycliques ainsi que les hétérocycliques complexes. Les équations 40, 41, et 42 sont les modèles d'estimation de Marrero-Morejon et Gani pour T_C, P_C et V_C respectivement:

$$T_C = 231,239 \ln\left[\sum_i n_i u_i + w\left(\sum_j n_j d_j\right) + z\left(\sum_k n_k t_j\right)\right] \tag{40}$$

$$P_C = 5,9827 + \frac{1}{\left[0.108998 + \sum_i n_i u_i + w\left(\sum_j n_j d_j\right) + z\left(\sum_k n_k t_k\right)\right]^2} \tag{41}$$

$$V_C = 7,95 + \sum_i n_i u_i + w\left(\sum_j n_j d_j\right) + z\left(\sum_k n_k t_k\right) \quad (42)$$

Avec un nombre extrêmement grand de groupes (respectivement: 125, 79 et 33 groupes de premier, deuxième et troisième ordres) la méthode est fort complexe et en considérant que, par exemple, seulement 783 données expérimentales ont été employées dans la régression pour la températures critiques (une moyenne d'approximativement trois données par groupe); les incrément (ou contributions) calibrés son discutables.

II.7 Comparaison et discussion des méthodes présentées:

Le tableau 5 résume les caractéristiques des différentes méthodes emblématiques déjà présentées. Pour chaque propriété (et pour chaque méthode); le ratio de calibrage ou la moyenne du nombre de composés pour calculer l'incrément du groupe c.-à-d. sa contribution à la propriété à estimer est donné par $N/n_G \approx$. On remarque que sa valeur décroit au fil des années (des méthodes) en cause est la sophistication de ces nouvelle méthodes qui, malgré l'augmentation du nombre de données expérimentales utilisées (cas de M&G01), où de plus en plus de groupes sont définis (M&P99 par rapport à Ly55 et J&R87) avec l'ajout de niveaux d'approximations (C&G94 et M&G01 par rapport aux autres méthodes). Le cas de la méthode de Constantinou *et al.* (1995), par exemple, est frappant où le calibrage de l'unique groupe à base de soufre (CH_2S) n'est fait qu'avec deux composés; et pour les 10 groupes incluant des halogènes seulement 17 composés des 181 que compte la base de données expérimentales en contiennent.

Pour comparer les précisions des différentes méthodes de contributions de groupe les métriques qui les caractérisent sont les déviations ou les erreurs d'estimations qui sont condensées dans le tableau 6 et se calculent comme suit:

La moyenne des écarts absolus:

$$AAD = \frac{1}{N}\sum_{i=1}^{N}\left|Y_i^{EXP} - Y_i^{EST}\right| \quad (43)$$

La moyenne des pourcentages des écarts absolus:

$$AAD\% = \frac{100}{N}\sum_{i=1}^{N}\frac{\left|Y_i^{EXP} - Y_i^{EST}\right|}{Y_i^{EXP}} \quad (44)$$

L'écart standard:

$$SD = \sqrt{\frac{\sum_{i=1}^{N}\left(Y_i^{EXP} - Y_i^{EST}\right)^2}{N}} \tag{45}$$

avec pour un composé *i*, Y_i^{EXP} et Y_i^{EST} sont respectivement les valeurs expérimentales et estimées de la propriété.

Tableau 5. Caractéristiques des méthodes emblématiques de contribution de groupes.

Code	n_{GR}	T_C			P_C			V_C		
		N	n_G	$N/n_G\approx$	N	n_G	$N/n_G\approx$	N	n_G	$N/n_G\approx$
Ly55	36	396	36	11	288	35	8	205	34	6
J&R87	40	409	40	10	392	40	10	310	39	8
C&G94	67; 48 [a]	285	67; 48	2	269	67; 47	2	251 (181) [c]	65; 47 (47; 27) [c]	2 (2) [c]
M&P99	167	391	166	2	345	158	2	189	134	1
M&G01	125;79; 33 [b]	783	123; 78; 32	3	775	124; 79; 33	3	762	125; 78; 33	3

N: nombre de données ; n_G: nombre de groupes; [b; (a)] (1^{er}; $2^{ème}$); $3^{ème}$ ordres; *$N/n_G\approx$: approximation du ratio de définition.* [c] description de la méthode de Constantinou *et al.* (1995) pour ω

Tableau 6. Métriques caractérisant différentes méthodes de contribution de groupes.

Code	T_C (K)			P_C (*k*Pa)			V_C (*c*m^3 mol^{-1})		
	AAD (K)	AAD%	SD (K)	AAD (*k*Pa)	AAD%	SD (*k*Pa)	AAD (*c*m^3 mol^{-1})	AAD%	SD (*c*m^3 mol^{-1})
Ly55	8,2	1,4	-	334	8,9	-	10	3,1	-
J&R87	4,8	0,8	6,9	213	5,2	330	7,5	2,3	13,2
C&G94	4,85	0,85	6,98	114	2,89	204	6 (0,010)*	1,79 (3)*	10 (0,015)*
M&P99	2,79	0,48	4,39	107	2,92	172	4,56	1,45	6,68
M&G01	4,93	0,8	6,99	80	2,3	140	7,33	1,8	10,74

***** Valeurs de la méthode de Constantinou *et al.* (1995) pour ω

La méthode de Lydersen (1955) présente pour toutes les propriétés traitées les plus grands écarts.

La méthode de Constantinou *et al.* (1995) étant la seule pour l'estimation de ω aucune discussion de sa performance ne peut être faite contrairement aux autres méthodes pour les constantes critiques. En parcourant le tableau 6 on remarque:

- $\boldsymbol{P_C}$: M&G01 est la meilleure méthode avec une moyenne des écarts de 80 *k*Pa (2,3 % en moyenne des pourcentages des écarts). Néanmoins, ce résultat n'est atteint qu'en utilisant les trois niveaux d'approximations disponibles ce qui fait de cette méthode la plus complexe et qui a empêché et découragé toute implémentation dans le DDBSP (base de donnée et logiciel d'estimation des propriétés physiques et thermodynamiques de la Dortmund Data Bank).
- $\boldsymbol{T_C}$: M&G01 et C&G94 sont de performances similaires et J&R87 un peu moins.
- $\boldsymbol{V_C}$: C&G94 et M&G01 ont presque la même précision (*AAD%* = 1,79% ≈ 1,80% pour les deux méthodes respectivement) ce qui prouve comme déjà mentionné dans la littérature que le 3ème niveau n'a pas apporté l'amélioration escomptée bien au contraire. J&R87 est la moins performante.

Les deux dernières constantes critiques ($\boldsymbol{T_C}$ et $\boldsymbol{V_C}$) sont bien estimées par M&P99 mais cette méthode a la plus petite gamme de composés (167 pour $\boldsymbol{T_C}$ et 189 pour $\boldsymbol{V_C}$) pour le calibrage des incréments qui sont basés sur des groupes définis par J&R87 comme vu précédemment avec les défaut inhérents à leurs définitions qui parfois n'ont pas de base théorique. Ces deux inconvénients, donc, réduisent son domaine d'application. De plus M&P99 est caractérisée par les plus faibles ratios (2 et 1 pour $\boldsymbol{T_C}$ et $\boldsymbol{V_C}$ respectivement) ce qui rend les valeurs des incréments douteuses. Puisque cette méthode considère seulement les interactions entre liaisons voisines, les capacités prédictives se dégradent fréquemment pour les gros composés, les composés polyfonctionnels et polycycliques où il est plus approprié de considérer le potentiel intermoléculaires et non pas les liaisons (Nannoolal Y, 2006).

Pour classer et utiliser les composés chimiques, il est indispensable de connaître leurs propriétés physico-chimiques et leurs activités biologiques, mais les essais expérimentaux notamment biologiques s'avèrent très onéreux. De plus, pour des raisons d'éthique ils doivent être limités. Aussi, a-t-on recours à des méthodes alternatives comme les méthodes chémoinformatiques qui ont pour but d'anticiper le comportement des molécules ou des systèmes moléculaires. Les techniques QSAR (Quantitative Structure-Activity Relationships) et QSPR (Quantitative structure property relationship) sont parmi les plus utilisées. La méthode QSAR modélise la relation entre la structure et l'activité de la molécule. Cette méthode a été introduite au XIX^{ème} siècle. En effet, dès 1868-1869, Crum-Brown chimiste et Fraser pharmacologiste écrivaient déjà dans leur papier (Blake, 1868) "il ne peut y avoir d'objection raisonnable contre le fait que la relation entre l'effet physiologique d'un composé et sa constitution chimique existe . . . ". Cependant, à cette époque, les structures moléculaires n'étaient pas encore connues.

Il a fallu attendre les années 60 avec les travaux de Free et Wilson (Free et Wilson, 1964) et les travaux de Hansh et Fujita (Hansch *et al.*, 1962) pour obtenir les premiers modèles de régression reliant les caractéristiques des structures chimiques à leurs propriétés physico-chimiques et/ou biologiques. Pour l'analyse de Hansch (Hansch C *et al.*, 1962), trois types de descripteurs sont fondamentaux dans cette approche afin de décrire les activités biologiques : les descripteurs hydrophobiques, stériques et électroniques. Dans Free et Wilson (Free et Wilson, 1964), des fragments moléculaires dérivés de l'analyse 2D des structures chimiques sont plutôt considérés. Par la suite comme décrits ci-dessous, de nombreux descripteurs moléculaires ont été définis avec des techniques d'analyses statistiques extrêmement variées.

III.1 Définition et formalisme:

La modélisation QSAR sur un ensemble de produits chimiques structurellement apparentés fait référence au développement d'une corrélation mathématique entre une réponse chimique et des attributs chimiques quantitatifs définissant les caractéristiques des molécules analysées. Par conséquent, une telle étude tente d'établir un formalisme mathématique entre le comportement d'une réaction chimique et un ensemble d'attributs chimiques quantitatifs qui peuvent être extraits des structures chimiques en utilisant des moyens expérimentaux ou théoriques appropriés.

Dès lors qu'une relation mathématique est développée, de telles études permettent de prédire le comportement moléculaire de nouveaux produits chimiques ou même de molécules hypothétiques. Par conséquent, le formalisme de base de la technique QSAR peut être représenté mathématiquement comme suit:

Activité biologique=f(attributs chimiques) (46)

Les attributs chimiques sont les informations fondamentales des produits chimiques qui contrôlent la réponse étudiée. Puisque le but était de développer une corrélation mathématique, ces caractéristiques ou attributs sont des informations chimiques quantitatives précises qui pourraient être dérivées en utilisant une analyse expérimentale ou un algorithme théorique approprié qui diagnostique la chimie des molécules. Parfois, des informations obtenues à la fois sur le plan théorique et sur le plan expérimental sont utilisées. Par conséquent, les attributs chimiques dans l'équation (46) sont souvent décrits en termes d'information dérivée directement de la structure chimique et des informations physicochimiques habituellement obtenues en utilisant des techniques expérimentales conduisant à l'expression suivante (Todeschini *et al.* 2009).

Réponse= f(structure chimique, propriété physicochimique) (47)

Ainsi, les mathématiques servent ici d'outil pour dériver une relation appropriée qui est ensuite exploitée selon l'exigence du chercheur (Tute, 1990).

Les données quantitatives sont obtenues à partir de deux composantes majeures, à savoir la réponse et les variables prédictives ou indépendantes (c'est-à-dire les variables X) définissant les attributs chimiques. Les données de réponse peuvent être une activité (QSAR) (par exemple, antipaludique, anti-oxydante, anti-arythmique, anti-VIH et anticancéreuse), des propriétés (QSPR) (par exemple, solubilité aqueuse, coefficient de partage n-octanol / eau, point de fusion, tension de surface, valeur de concentration micellaire critique et rétention chromatographique), ou toxicologique (QSTR) (toxicité aiguë / chronique spécifique à un organe ou une maladie, cancérogénicité, irritation de la peau, génotoxicité, hépatotoxicité et toxicité pour l'environnement en termes de décès d'organismes indicateurs spécifiques tels que *Tetrahymena*, bactéries et poissons).

La matrice de données comprenant des réponses et des descripteurs peut être soumise à un développement de modèle linéaire et non linéaire en combinaison avec un algorithme pour servir à la sélection de variables appropriées. La régression linéaire multiple (MLR) et

les moindres carrés partiels (PLS) sont les techniques représentatives pour le développement de modèles de corrélation linéaire tandis que l'algorithme génétique (GA), l'algorithme pas à pas, etc. peuvent servir à la sélection de variables. Les approches de modélisation non linéaire comprennent par exemple les réseaux de neurones artificiels (RNA), les machines à vecteurs support (SVM).

III.2 Importance des QSAR:

Bien que le développement de modèles QSAR / QSPR / QSTR prédictifs semble être une tâche relativement simple, il a d'énormes applications pour répondre aux besoins de la communauté scientifique. Il a toujours été curieux de voir comment il est possible que différents agents chimiques exercent un profil de réponse différent. Par conséquent, les caractéristiques chimiques semblent être très cruciales dans la détermination du comportement des produits chimiques (Roy *et al.* 2015). Les techniques QSAR peuvent offrir plusieurs avantages en termes de prédictivité du modèle et d'utilisation de ressources expérimentales limitées, en utilisant moins de temps de calcul. Ces caractéristiques encouragent l'utilisation des techniques QSAR et des techniques connexes dans des programmes de recherche coûteux tels que la découverte et le développement de médicaments qui peuvent fournir des informations précieuses en favorisant une stratégie rationnelle de conception. De plus, puisque la technique QSAR peut permettre la prédiction d'une réponse chimique d'un nombre relativement important de composés (dans le domaine chimique) en utilisant des données de réponse d'un nombre limité de produits chimiques, elle est largement utilisée dans l'analyse prédictive de la toxicologie pour l'évaluation des risques chimiques .

L'encodage des caractéristiques chimiques dans l'analyse QSAR est effectué en utilisant un algorithme mathématique approprié. L'objectif étant d'établir un diagnostic précis des caractéristiques structurelles chimiques suivi de la dérivation de nombres quantitatifs appelés aussi descripteurs. Ces descripteurs portent des informations structurelles explicites et servent à établir une corrélation avec une réponse d'intérêt. Par conséquent, dans une terminologie simple, les descripteurs fournissent la base de la représentation quantitative de la structure chimique, c'est-à-dire des nombres quantitatifs dérivés d'une opération mathématique appropriée d'une information chimique.

III.3 Descripteurs moléculaires:

III.3.1 Définition:

Un modèle QSAR peut être exprimé comme une équation mathématique simple qui peut corréler les propriétés (physico-chimiques / biologiques / toxicologiques) de molécules employant divers paramètres quantitatifs calculés ou expérimentalement appelés «descripteurs» Les descripteurs sont corrélés avec les propriétés expérimentales (réponse) en utilisant une variété d'outils chimiométriques afin d'obtenir un modèle QSAR statistiquement significatif. Les descripteurs moléculaires sont les «termes qui caractérisent l'information spécifique d'une molécule étudiée.» Ce sont les «valeurs numériques associées à la constitution chimique pour la corrélation de la structure chimique avec diverses propriétés physiques, réactivité chimique ou activité biologique». (Guha et Willighagen, 2012). En d'autres termes, la réponse d'un produit chimique peut être présentée mathématiquement comme la fonction des descripteurs (éq. 48).

Réponse (Activité/Propriété/ Toxicité)= f (Descripteurs) (48)

III.3.2 Types de descripteurs:

Les descripteurs peuvent être de différents types selon la méthode de calcul ou de détermination: physicochimique (hydrophobe, stérique ou électronique), structurale (fréquence d'apparition d'une sous-structure), topologique, électronique (calculs orbitales moléculaires), géométrique (surface moléculaire). Dans une perspective plus large, les descripteurs (spécifiquement, descripteurs physicochimiques) peuvent être classés en deux groupes principaux: (1) les constantes de substitution et (2) les descripteurs de la molécule entière (Todeschini et Consonni, 2000, Livingstone, 2000). Les constantes de substitution sont essentiellement des descripteurs physicochimiques qui sont conçus sur la base de facteurs qui régissent les propriétés physico-chimiques des entités chimiques. Les descripteurs de la molécule entière sont des expansions de l'approche constantes de substitution, mais beaucoup d'entre eux sont également dérivés d'approches expérimentales.

Les descripteurs peuvent également être classés en fonction des dimensions. Le tableau 7 donne une illustration utile des descripteurs moléculaires couramment utilisés en fonction des dimensions. Il est intéressant de noter que nous avons limité ici notre discussion des descripteurs 0D à 3D, bien que des descripteurs de dimensions plus élevée soient également disponibles.

Tableau 7. Différents descripteurs utilisés dans les études QSAR classés par dimension.

Dimension des descripteurs	**Paramètres**
Descripteurs 0D	Indices constitutionnels, propriété moléculaire, nombre d'atomes et de liaisons.
Descripteurs 1D	Nombre de fragment, empreintes moléculaires.
Descripteurs 2D	Paramètres topologiques, paramètres structuraux, paramètres physico-chimiques y compris les descripteurs thermodynamiques
Descripteurs 3D	Paramètres électroniques, paramètres spatiaux, paramètres d'analyse de forme moléculaire, paramètres d'analyse de champ moléculaire et paramètres d'analyse de surface de récepteur

A- Descripteurs 2D:

Topologique: Les descripteurs topologiques sont calculés sur la base de la représentation graphique des molécules et donc ils ne nécessitent aucune estimation de propriétés physico-chimiques et n'ont pas besoin des calculs rigoureux impliqués dans la dérivation des descripteurs chimiques quantiques. La représentation de la structure de la molécule dépend de sont graphe topologie 2D indiquant la position individuelle des atomes et les connexions entre eux. La formulation de ces descripteurs est basée sur la caractérisation de la structure chimique par la théorie des graphes. La détermination théorique des graphes de la structure moléculaire implique des sommets symbolisant les atomes et les liaisons covalentes représentant les bords (Roy et Das, 2014). Parmi les descripteurs topologiques les plus utilisés on trouve: l'indice de Balaban J, les indices de connectivité liaison / bord, les indices d'état électronique, les Indices de connectivités moléculaire, Indice de Wiener W, le groupe d'indices Zagreb, etc...

Paramètres structurels: Parmi lesquels : le poids moléculaire MW, nombre de centres chiraux, nombre groupes ou moitiés donneurs ou accepteur de liaison hydrogène etc...

Paramètres physico-chimiques: sont conçus sur la base de facteurs qui régissent les propriétés physiques et chimiques des entités chimiques. En raison du changement des propriétés physicochimiques, de l'absorption, de la distribution, du transport, du métabolisme et de l'élimination, le comportement des entités chimiques bioactives peut être modifié. Les facteurs physicochimiques importants affectant la bioactivité des médicaments et des produits chimiques comprennent l'hydrophobicité, le caractère électronique et le caractère stérique de

l'ensemble des molécules ainsi que les substituants présents dans les molécules (Taylor, 1991; Rekker, 1977; Hansch *et al.*, 1995). On peut citer comme exemples: le coefficient de partage octanol-eau logP, la constante d'hydrophobicité π, la constante de dissociation acide K_a, la réfractivité molaire MR etc...

D'autres types descripteur 2D existent comme: les descripteurs thermodynamiques.

B-Descripteurs 3D:

Paramètres électroniques: Les descripteurs électroniques sont définis en termes de charges atomiques et sont utilisés pour décrire les aspects électroniques à la fois de la molécule entière et de régions particulières, telles que les atomes, les liaisons et les fragments moléculaires. Les charges électriques dans la molécule sont la force motrice des interactions électrostatiques, et il est bien connu que les densités ou charges électroniques locales jouent un rôle fondamental dans de nombreuses réactions chimiques et propriétés physico-chimiques (Todeschini *et al.* 2009). Par exemple : la somme des polarisabilités atomiques, le moment dipolaire, l'énergie de la plus haute orbitale moléculaire occupée HOMO, l'énergie de la plus basse orbitale moléculaire occupée LUMO etc...

Paramètres spatiaux: qui comprennent une série de descripteurs calculés en fonction de l'arrangement spatial des molécules et de la surface qu'elles occupent parmi lesquels: le rayon de giration, la surface moléculaire, la densité, le volume moléculaire etc ...

III.4 Constitution des jeux de calibrage et de validation:

Généralement un seul ensemble de structures avec les activités biologiques ou propriétés mesurées associées est disponible pour une modélisation QSAR/QSPR. Cet ensemble doit donc être divisé en un jeu de calibrage et un jeu de validation afin d'évaluer le pouvoir de prévision du modèle. Il faut que le jeu de validation sont (Fortuné, 2006):

- compte au moins 5 composés ;
- ces composés couvrent la gamme des structures et des activités du jeu d'apprentissage (Golbraikh et Tropsha, 2002a; Golbraikh *et al.*, 2003);
- chaque composé du jeu de calibrage soit proche d'un composé du jeu de validation (Golbraikh *et al.*, 2003).

La méthode la plus simple consiste à faire appel au hasard en prélevant un nombre déterminé de composés de façon aléatoire dans l'ensemble de départ. Les composés restant

constituent le jeu de calibrage et l'opération est répétée pour obtenir différents couples de jeux de validation et calibrage (Globisch *et al.*, 2006).

On peut améliorer la représentativité du jeu de validation par une sélection plus rationnelle. Les méthodes de partage qui peuvent être utilisées sont nombreuses: algorithme de Kennard-Stone (Kennard et Stone, 1969), algorithme DUPLEX (Snee, 1977), sélection aléatoire, OPTISIM (Clark, 1997), Répartition uniforme des échantillons sur la variable dépendante, etc...) (Bouveresse, 2004). Dans ce travail l'algorithme de Kennard et Stone (aussi appelé CADEX) à été privilégié.

Algorithme de Kennard-Stone:

Une des alternatives à la sélection aléatoire est l'utilisation de l'algorithme de Kennard et Stone (Dantas Filho, 2004). L'algorithme maximise la distance euclidienne minimale entre les échantillons déjà sélectionnés et les échantillons restants. La procédure est décrite ci-dessous et est illustrée dans la figure 8.

a- sélection des échantillons les plus éloignés. Il s'agit ici des échantillons n°1 et 2 qui sont entourés sur la figure 8;

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

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

Cette procédure est répétée jusqu'à ce que le nombre d'échantillons spécifié par l'utilisateur soit atteint.

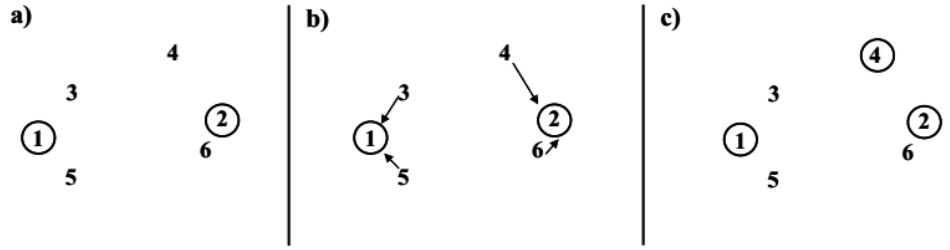


Figure 8: Répartition des échantillons avec l'algorithme de Kennard et Stone (de Groot, 1999)

Les avantages de cet algorithme sont que les échantillons d'étalonnage (calibrage) couvrent toujours complètement la région mesurée de l'espace des variables d'entrée et qu'aucun échantillon de validation ne tombe en dehors de cette région (Durand, 2007). L'algorithme CADEX a été considéré comme l'une des meilleures façons de construire des ensembles d'entraînement et de validation (Tropsha *et al.*, 2003, Wu *et al.*, 1996).

L'élaboration des modèles QSAR repose sur le calcul des structures (descripteurs), ce dernier est assuré en faisant appel aux outils de la modélisation moléculaire. Différentes approches sont envisageables dans le cadre des outils de modélisation moléculaire. Les méthodes quantiques en l'occurrence les méthodes *ab initio*, la théorie de la fonctionnelle de la densité et les méthodes semi-empiriques (Errahoui, 2015) sont capables de calculer plusieurs propriétés des systèmes. Il s'agit d'expliciter dans cette partie, les méthodes de chimie quantique utilisées seulement pour le calcul des structures moléculaires (descripteurs) nécessaires à la mise en place de modèles performants. Il ne s'agit pas d'une description exhaustive et pour plus de détail de nombreux ouvrages spécialisés sont disponibles. (McWeeny et Sutcliffe, 1969; Atkins, 1983; Szabo et Osllund, 1982; Rivail, 1994; Jensen, 2007)

La mécanique quantique est une théorie qui se fonde sur un ensemble d'axiomes, l'un d'eux stipule que tout état d'un système n'évoluant pas dans le temps constitué de N particules est complètement décrit par une fonction mathématique Ψ, appelée fonction d'onde, qui dépend des coordonnées de chacune des particules, la fonction d'onde ne possède aucune signification physique, en revanche, la quantité $|\Psi^2|$ permet de déterminer la probabilité de présence des particules dans un élément de volume. Un second axiome énonce que l'action d'un opérateur mathématique hermétique sur cette fonction permet d'atteindre la grandeur physique observable correspondante. Ainsi l'opérateur associé à l'énergie E est l'opérateur hamiltonien H:

$$\boldsymbol{H}\Psi = \boldsymbol{E}\Psi \tag{49}$$

La résolution exacte de l'équation (49) n'est possible que pour l'atome d'hydrogène et les systèmes hydrogénoïdes. Pour les systèmes poly-électroniques, il est nécessaire de faire appel aux méthodes d'approximation.

La première approximation en chimie quantique est de considérer l'équation de Schrödinger (49) (Schrödinger, 1926) non relativiste indépendante du temps où l'hamiltonien est défini par :

$$\boldsymbol{H} = -\frac{1}{2}\sum_i \Delta_i - \frac{1}{2}\sum_A \Delta_A - \sum_i \sum_A \frac{\boldsymbol{Z_A}}{\boldsymbol{r_{iA}}} + \sum_i \sum_{j>i} \frac{1}{\boldsymbol{r_{ij}}} + \sum_A \sum_{B>A} \frac{\boldsymbol{Z_A Z_B}}{\boldsymbol{r_{AB}}} \tag{50}$$

D'autres approximations sont adoptées et employées : l'approximation de Born-Oppenheimer (Born et Oppenheimer, 1927), l'approximation d'orbitales moléculaires basée sur le déterminant de Slater (Slater, 1929) et l'approximation C.L.O.A. (Bouakkadia, 2016)

Ayant utilisé les méthodes semi-empiriques pour l'optimisation de géométries des composés, un bref aperçu des éléments essentiels de ces méthodes, leurs avantages et défauts est donné.

Contrairement aux méthodes *ab initio*, les méthodes semi-empiriques utilisent des données ajustées sur des résultats expérimentaux afin de simplifier les calculs. La lenteur et la difficulté des calculs est en grande partie due aux intégrales biélectroniques qui apparaissent aux cours du processus de résolution.

Pour réduire ce temps de calcul, la base d'orbitales atomiques est d'abord réduite au minimum. Ensuite, les méthodes quantiques semi-empiriques font certaines approximations sur ces intégrales parmi lesquelles, certaines intégrales bi-électroniques sont négligées et d'autres sont paramétrées grâce à des données expérimentales. Plusieurs méthodes pour le traitement des intégrales bi-électroniques existent comme la méthode CNDO (Complete Neglect of Differential Overlap). De plus, les intégrales de recouvrement S_{ij} sont exprimées par :

$$\boldsymbol{S_{ij}} = \delta_{ij} \begin{cases} =1 & si \quad i=j \\ =0 & si \quad i\neq j \end{cases} \tag{51}$$

En principe, seuls les électrons de valence sont considérés diminuant ainsi le nombre de termes calculés. Le reste de l'atome est traité comme un « cœur » de charge « Z - le nombre d'électrons de cœur ». Les électrons de cœur sont pris en compte par une fonction de répulsion « cœur-cœur » en même temps que la répulsion nucléaire. Durant l'évolution des méthodes semi empiriques, cette fonction a été améliorée pour mieux rendre compte notamment, des liaisons hydrogènes. De nombreuses méthodes semi-empiriques basées sur ces approximations ont été développées notamment les méthodes MNDO, AM1, PM3 et PM6. L'avantage majeur de ces méthodes est la rapidité de calcul au sacrifice de la qualité de la description énergétique. L'incorporation des effets de dispersion et de liaison hydrogène passe par l'ajout de fonctions empiriques de type « champ de force ». En revanche, l'usage de paramètres basés sur des propriétés expérimentales permet de prendre en compte, bien que partiellement, des effets de corrélation électronique (absente dans la méthode de Hartree-Fock Hatree, 1928 ; Fock, 1930 ; Slater, 1928))

IV.1 MNDO:

La méthode MNDO (Modified Neglect of Differential Overlap) est une méthode introduite par Dewar et Thiel en 1977 (Dewar et Thiel, 1977a; 1977b). À l'origine, MNDO ne considérait que les atomes de carbone, d'hydrogène, d'oxygène et d'azote. Puis, par la suite, la paramétrisation s'est étendue à un plus grand nombre d'atomes. Un des désavantages de cette méthode est qu'elle décrit mal les liaisons hydrogènes (Stewart, 2007) qui sont pourtant essentielles.

IV.2 AM1:

Afin de corriger le problème de la représentation des liaisons hydrogènes, Dewar *et al.* ont développé la méthode AM1 (Austin Model 1) (Dewar *et al.*, 1985) en ajoutant des fonctions gaussiennes à la méthode MNDO pour représenter les interactions noyau-noyau.

IV.3 PM3:

Malgré les efforts effectués dans le développement de la méthode AM1, certains problèmes de paramétrisation persistaient. Stewart a donc proposé une nouvelle méthodologie nommée « Parametrized Model 3 » (PM3) (Stewart, 1989). Dans cette méthodologie, la paramétrisation atomique a été effectuée par groupe d'élément. Deux fonctions gaussiennes par atome sont utilisées pour le calcul de la répulsion cœur-cœur. De plus, des paramètres pour les éléments du groupe d sont prévus dans cette méthode.

IV.4 PM6:

Stewart *et al.* ont développé une nouvelle méthode s'appuyant sur la PM3 nommée PM6 dans laquelle a été incorporé un nouveau paramétrage cœur-cœur en visant l'amélioration des résultats pour les composés d'intérêt biologique (Stewart, 2007). Pour cela, ils ont modifié l'interaction cœur-cœur par une fonction de Voityuk (Voityuk et Rösch, 2000) qui permet de prendre en compte la répulsion de deux atomes non chargés grâce à l'incorporation d'un terme diatomique.

V.1 Régression linéaire multiple :

L'étude d'un phénomène peut, le plus souvent, être schématisé de la manière à ce qu'on s'intéresse à une grandeur y, que nous appellerons par la suite réponse ou variable expliquée, qui dépend d'un certain nombre de variables x_1; x_2; ….x_n que nous appellerons facteurs ou variables explicatives (Bouakkadia, 2016).

La régression est une des méthodes les plus connues et les plus appliquées en statistique pour l'analyse de données quantitatives. Elle est utilisée pour établir une liaison entre une variable quantitative et une ou plusieurs autres variables quantitatives, sous la forme d'un modèle. Si on s'intéresse à la relation entre deux variables, on parlera de régression simple en exprimant une variable en fonction de l'autre. Si la relation porte entre une variable et plusieurs autres variables, on parlera de régression multiple. La mise en œuvre d'une régression impose l'existence d'une relation de cause à effet entre les variables prises en compte dans le modèle (Chouquet, 2010).

La régression multi-linéaire (MLR, pour Multiple Linear Regression) (Lejeune, 2004) est la méthode la plus simple et la plus communément employée pour le développement de modèles prédictifs. Elle repose sur l'hypothèse qu'il existe une relation linéaire entre une variable dépendante y (ici, la propriété) et une série de n variables indépendantes xi (ici, les descripteurs). L'objectif est d'obtenir une équation de la forme suivante :

$$y = a_0 + a_1x_1 + a_2x_2 + \ldots + a_nx_n \qquad (52)$$

où a_i sont les coefficients de la régression.

La détermination de l'équation (52) se fait alors à partir d'une base de données de p échantillons pour laquelle à la fois les variables dépendantes et la variable indépendante sont connues. Il s'agit donc de considérer un système de p équations.

$$\begin{aligned} \hat{y}_1 &= a_0 + a_1x_{1,1} + a_2x_{2,1} + \ldots + a_nx_{n,1} + \varepsilon_1 \\ \hat{y}_2 &= a_0 + a_1x_{1,2} + a_2x_{2,2} + \ldots + a_nx_{n,2} + \varepsilon_2 \\ \hat{y}_p &= a_0 + a_1x_{1,p} + a_2x_{2,p} + \ldots + a_nx_{n,p} + \varepsilon_p \end{aligned} \qquad (53)$$

où les résidus ε_i représentent l'erreur du modèle, constituée par l'incertitude sur la variable dépendante y_i d'une part, sur les variables indépendantes x_i d'autre part, mais aussi par les informations contenues dans les variables indépendantes mais non exprimées via les variables dépendantes.

Ce système d'équations peut être écrit sous la forme matricielle suivante :

$$\begin{pmatrix} y_1 \\ . \\ . \\ . \\ y_p \end{pmatrix} = \begin{pmatrix} 1 & x_{1,1} & \dots & x_{n,1} \\ . & . & . & . \\ . & . & . & . \\ . & . & . & . \\ 1 & x_{1,p} & \dots & x_{1,p} \end{pmatrix} \begin{pmatrix} a_0 \\ . \\ . \\ . \\ a_p \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ . \\ . \\ . \\ \varepsilon_p \end{pmatrix} \qquad (54)$$

Soit de manière condensée :

$$\mathbf{Y} = \mathbf{X}\,\mathbf{A} + \boldsymbol{\varepsilon} \qquad (55)$$

La méthode consiste alors à choisir les coefficients du vecteur **A** en faisant en sorte de minimiser la somme des carrés des écarts entre les valeurs prédites et les valeurs réelles sur l'intégralité de la base de données et ceci sous couvert de certaines hypothèses de départ.

En premier lieu, les variables indépendantes x_i, comme leur nom l'indique, sont supposées indépendantes entre elles et leur incertitude est négligeable. Ensuite, les différents échantillons yi sont supposés indépendants entre eux et suivent une distribution normale. L'erreur ε est elle-même supposée suivre une distribution normale, centrée en 0. Enfin, par nature, la dépendance de y vis-à-vis des x_i est supposée linéaire.

La valeur prédite de la variable dépendante est alors :

$$\hat{y}_i = \hat{a}_0 + \hat{a}_1 x_{1,i} + \dots + \hat{a}_n x_{n,i} \qquad (56)$$

Les résidus peuvent donc être définis comme la différence entre les valeurs prédites et observées de y.

$$\varepsilon_i = y_i - \hat{y}_i \qquad (57)$$

Il s'agit alors de trouver les coefficients $\hat{a}_i$ afin de minimiser la somme des carrés de ces résidus pour l'intégralité de la base de données.

$$\min [\sum(\varepsilon_i)^2] = \min [\sum(y_i - \hat{y}_i)^2] = \min [\sum(y_i - \hat{a}_0 - \hat{a}_1 x1_{,i} - \dots - \hat{a}_n x_{n,i})^2]$$

$$= \min (Y-X\hat{A})^T (Y-X\hat{A}) \quad (58)$$

Les coefficients peuvent être obtenus à partir de l'équation matricielle suivante :

$$\hat{A} = (X^T X)^{-1} X^T Y \quad (59)$$

Bien entendu, la régression multilinéaire souffre de certains désavantages. Le principal découle de sa linéarité. Elle est donc défaillante pour la mise en évidence de dépendances non-linéaires. Cela dit, elle n'en reste pas moins une méthode simple et efficace dans la plupart des cas.

De plus, pour peu que les variables indépendantes soient choisies de manière raisonnée, les équations obtenues peuvent être interprétées d'un point de vue phénoménologique (Fayet, 2010).

V.2 Réseaux de neurones artificiels:

Depuis quelques années, les Réseaux de Neurones Artificiels (RNA) font l'objet d'un vif engouement de la part de chercheurs chimistes et biologistes (Douali, 2007). En effet, les RNA semblent apporter une solution satisfaisante au dépouillement des spectres RMN 1H, à l'interprétation des spectres infrarouge, à la prédiction des déplacements chimiques de ^{13}C, à l'analyse des spectres de masse, à la détermination de la structure secondaire des protéines, etc (McGregor *et al.*, 1989). Plusieurs travaux ont été réalisés afin d'établir des relations structure-activité biologique et ont montré que les RNA constituent un outil efficace pour la résolution des problèmes de ce type (Arakawa *et al.*, 1989).

Les réseaux de neurones ont été étudiés depuis les années 40. Les idées de base de cette technique viennent de la recherche cognitive, d'où vient le nom 'réseaux de neurones'. L'origine des réseaux de neurones peut être attribuée à McCulloh et Pitts en 1943 (McCulloh et Pitts, 1943) ; ils proposent un modèle mathématique décrivant le fonctionnement d'un neurone biologique. Dans les années 80, Hopfield suscite à nouveau l'intérêt des scientifiques en proposant des neurones associatifs (Hopfield, 1982).

Un réseau de neurones est composé d'unités de calculs « le neurone artificiel » (Figure 9) disposées en couches et reliées entre elles pour échanger de l'information (Hinton, 1992). Le réseau contient trois types de couches ; les couches d'entrée, les couches cachées et les couches de sortie. L'information circule des neurones d'entrée vers les neurones de sortie sans retour arrière possible via des fonctions de transfert ou d'activation. Les valeurs d'entrée sont multipliées par leur poids W_{ij} correspondant et additionnées pour obtenir la somme $S = \Sigma W_{ij}X_j$.

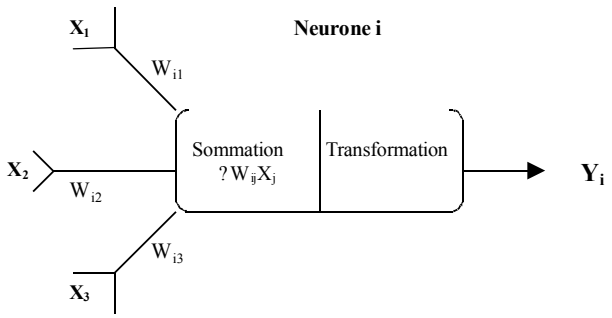


Figure 9: Le neurone artificiel générique ou formel.

Cette somme devient l'argument de la fonction d'activation, ces fonctions de transfert existent essentiellement sous trois formes : les fonctions linéaires, les fonctions seuils et les fonctions sigmoïdes. Ces dernières sont généralement les plus utilisées car elles représentent un bon compromis entre les fonctions seuils et linéaires (figure 10).

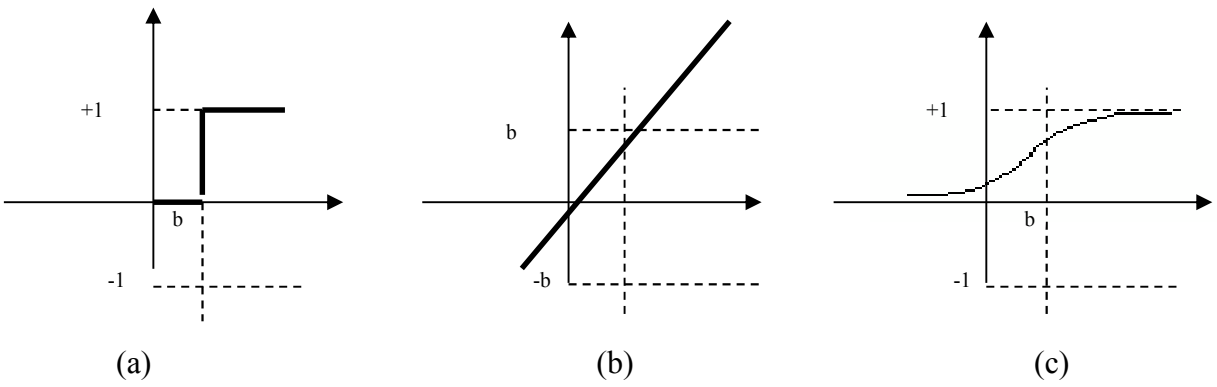


Figure 10 : Fonction de transfert (a) seuil, (b) linéaire et (c) sigmoïde du neurone.

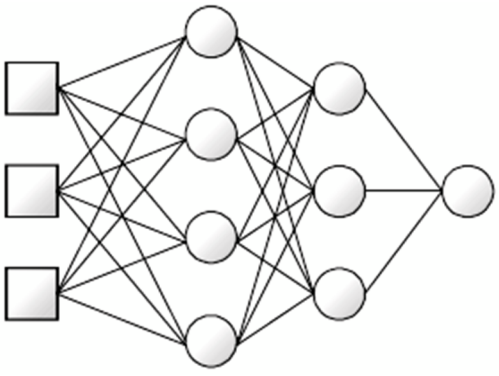
Le choix de la fonction d'activation dépend de l'application. S'il faut avoir des sorties binaires c'est la première fonction que l'on choisit habituellement.

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 (Rumelbart *et al.,* 1988).

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 (Rumelbart *et al.,* 1988), 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 multicouches ou perceptron multicouches PMC : sont les réseaux les plus puissants des réseaux de neurones qui utilisent l'apprentissage supervisé. Ces réseaux (figure 11) se composent des entrées, une couche de sortie et zéro ou plusieurs couches cachées (Rumelbart *et al.,*1988). 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 est utilisée.



Les entrées Couches cachées Couche de sortie

Figure 11: Structure générale du perceptron multicouche : schéma de principe

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.

Par analogie aux modèles biologiques, les connections à l'intérieur des RNA sont modifiables. Leur état évolue selon une règle d'adaptation ou d'apprentissage. L'apprentissage des RNA consiste à modifier les poids (W_{ij}) des connections jusqu'à obtenir une stabilisation du réseau où les poids ne se modifient plus que d'une façon infime. On peut distinguer deux types d'apprentissage : supervisé et non supervisé.

V.2.1 Apprentissage non supervisé:

Les RNA doués de ce type d'apprentissage présentent des capacités d'auto organisation et des fonctionnements en mémoires associatives. Au cours de ce processus, on ne présente au réseau que des exemples en entrée et on le laisse s'auto-organiser uniquement au moyen de lois locales qui régissent l'évolution des poids. Le réseau est ainsi amené à un état stable, après un certain nombre d'itérations. Les réseaux de Hopfield et de Kohonen (Kohonen, 1995) présentent de célèbres exemples utilisant l'apprentissage non supervisé.

V.2.2 Apprentissage supervisé:

L'apprentissage supervisé consiste à présenter au réseau les exemples ainsi que les réponses désirées et à modifier les poids en fonction des réponses fournies par le réseau. Ce type d'apprentissage est généralement utilisé par les réseaux multicouches (Gasteiger *et al.*, 1993; Burns et Whitesides, 1993). L'ajustement des poids se fait souvent selon les règles de l'algorithme de rétropropagation du gradient de l'erreur qui est une méthode pour calculer le gradient de l'erreur pour chaque neurone du réseau, de la dernière couche vers la première. Dans le cas des réseaux de neurones, les poids synaptiques qui contribuent à engendrer une erreur importante se verront modifiés de manière plus significative que les poids qui ont engendré une erreur marginale.

V.3 Sélection de sous - ensemble de variables par algorithme génétique (GA - VSS):

Les algorithmes génétiques (GA) (Goldberg, 1989; Alliot *et al.*, 2002) initiés dans les années 1970 par John Holland, sont des algorithmes d'optimisation stochastiques itérés s'appuyant sur des techniques dérivées de la génétique et des mécanismes d'évolution de la nature : croisement, mutation, sélection. Les algorithmes génétiques fournissent des solutions

aux problèmes n'ayant pas de solutions calculables en temps raisonnable de façon analytique ou algorithmique.

Selon cette méthode, des milliers de solutions (génotypes) plus au moins bonnes sont créées au hasard puis sont soumises à un procédé d'évaluation de la pertinence de la solution mimant l'évolution des espèces : les plus "adaptés", c'est- à- dire les solutions au problème qui sont les plus optimales survivent davantage, que celles qui le sont moins et la population évolue par générations successives en croisant les meilleures solutions entre elles et les faisant muter, puis en relançant ce procédé un certain nombre de fois afin d'essayer de tendre vers la solution optimale.

Les algorithmes génétiques constituent une méthode de choix pour la sélection de sous- ensembles de variables explicatives ou en anglais: Genetic Algorithm for Variable Subsets Selection ou GA-VSS.

Dans le présent travail, la sélection des descripteurs a été réalisée par algorithme génétique dans le logiciel MobyDigs (Todeschini *et al.,* 2004).

V.4 Evaluation d'un modèle QSAR/ QSPR:

La modélisation vise à fournir un modèle qui soit non seulement ajusté aux données d'apprentissage, mais aussi capable de prédire la valeur de la sortie sur de nouveaux exemples, c'est-à-dire généralisable. Pour minimiser l'influence des erreurs de détermination des valeurs explicatives (ou régresseurs) sur la précision des résultats de la régression 4 à 5 données (variables dépendantes, ou encore observations) doivent, au minimum, être associées à chaque variable explicative. Le nombre de degrés de liberté final (n-p-1) doit être (Tomassone *et al.,* 1983) tel que :

$$n- p- 1 \geq 10 \qquad (60)$$

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

Pour les modèles à plus de deux descripteurs, de faibles coefficients de corrélation croisés n'assurent pas forcément l'orthogonalité des descripteurs. Une indépendance globale acceptable des descripteurs sera vérifiée lorsque les facteurs d'inflation de la variance (FIV) calculés pour chacun d'eux obéissent à la condition FIV< 5 (Tomassone *et al.,*1983).

Deux paramètres statistiques sont couramment utilisés pour l'évaluation de la qualité du modèle :

- Le coefficient de détermination multiple :

$$R^2 = 1 - \frac{SCE}{SCT} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \quad (61)$$

où $\bar{y}$ est la valeur moyenne des valeurs observées.

- La racine de l'écart quadratique moyen de prédiction :

$$EQMP = \sqrt{\frac{1}{n}\sum_{1}^{n}(y_i - \hat{y}_{(i)})^2} = \sqrt{\frac{PRESS}{n}} \quad (62)$$

Il est intéressant de considérer, également, la racine de l'écart quadratique moyen calculé sur les ensembles de calibrage (EQMC), et sur l'ensemble de validation externe ($EQMP_{ext}$) :

$$EQMC = \sqrt{\frac{\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2}{n}} \quad (63)$$

$$EQMP_{ext} = \sqrt{\frac{\sum_{i=1}^{n_{ext}}\left(y_i - \hat{y}_{(i)}\right)^2}{n_{ext}}} \quad (64)$$

La validation croisée par « leave – one – out » (LOO) (Wehrens *et al.,* 2000) consiste à recalculer le modèle sur (n-1) observations, et à utiliser le modèle ainsi obtenu pour calculer la grandeur d'intérêt du composé écarté, notée $\hat{y}_{(i)}$. On répète le procédé pour chacune des grandeurs d'intérêt. La somme des carrés des erreurs de prédiction, désignée par le symbole PRESS (éq. 62), est une mesure de la dispersion des estimations. On l'utilise pour définir le coefficient de prédiction (Wehrens *et al.,* 2000):

$$Q^2_{LOO} = \frac{SCT - PRESS}{SCT} \quad (65)$$

Contrairement à R^2 qui augmente avec le nombre de paramètres du modèle, le facteur Q^2_{LOO} affiche une courbe avec maximum (ou avec palier) obtenu pour un certain nombre de descripteurs, puis décroît de façon monotone. Ce fait confère une grande importance au coefficient Q^2_{LOO}. Une valeur de $Q^2_{LOO} > 0,5$ est considérée comme satisfaisante, et une valeur supérieure à 0,9 est excellente (Eriksson *et al.,* 2003).

Si de petites valeurs de Q^2_{LOO} indiquent des modèles peu robustes, caractérisés par de faibles capacités prédictives internes, le contraire n'est pas nécessairement vrai. En fait, si une forte valeur de Q^2_{LOO} est une condition nécessaire de robustesse et d'une possible capacité prédictive élevée d'un modèle, cette condition seule n'est pas suffisante, et peut conduire à une surestimation de la capacité prédictive du modèle lorsqu'il est appliqué à des composés réellement externes.

Evidemment, on peut être amené à écarter 2, 3 ou un plus grand nombre d'éléments à la fois, ce qui conduit aux procédures LMO (leave – many – out).

La validation interne peut être également réalisée en utilisant la technique du bootstrap (bootstrapping): Q^2_{boot}. Elle consiste à simuler m échantillons de même taille n que l'échantillon initial. Ils sont obtenus par tirage au hasard avec remise parmi les n individus observés au départ, ceux-ci ayant tous la même probabilité 1/n d'être choisis (Wehrens *et al.,* 2000; Draper et Smith, 1998). Contrairement aux validations croisées par LOO et LMO, les méthodes de bootstrap sont plus efficaces et plus stables.

Il est intéressant, pour juger de la qualité du modèle, de considérer le coefficient de prédiction externe calculé comme suit :

$$Q^2_{EXT} = 1 - \frac{\sum_{i=1}^{next} (\hat{y}_{i/i} - y_i)^2 / n_{ext}}{\sum_{i=1}^{ntr} (y_i - \overline{y}_{tr})^2 / n_{tr}} \quad (66)$$

Il a été démontré sur plusieurs jeux de données qu'il n'existe aucune corrélation entre Q^2_{LOO} et Q^2_{EXT} (Golbraikh *et al.*, 2003, Peterson *et al.*2006, Golbraikh et Tropsha, 2002b). La validation par un jeu de Test externe est donc une étape essentielle de la validation d'un modèle QSAR

Une validation externe supplémentaire selon (Golbraikh et Tropsha, 2002b) est appliquée uniquement à l'ensemble de Test. Selon les critères recommandés, un modèle QSPR prédictif, doit satisfaire aux conditions suivantes:

1) $R^2_{CV\ EXT} > 0{,}5$ (67-a)

2) $R^2 > 0{,}6$ (67-b)

3) $(R^2 - R_0^2)/R^2 < 0{,}1$ et $0{,}85 < k < 1{,}15$ (67-c)

$(R^2 - R'^2_0)/R^2 < 0{,}1$ et $0{,}85 < k' < 1{,}15$ (67-d)

4) $\left|R_0^2 - R'^2_0\right| < 0{,}3$ (67-e)

où:

$$R = \frac{\sum (y_i - \overline{y})(\tilde{y}_i - \overline{\tilde{y}})}{\sqrt{\sum (y_i - \overline{y})^2 \sum (\tilde{y}_i - \overline{\tilde{y}})^2}} \qquad (68\text{-a})$$

$$R_0^2 = 1 - \frac{\sum (y_i - y_i^{r_0})^2}{\sum (y_i - \overline{y})^2} \qquad (68\text{-b})$$

$$R'^2_0 = 1 - \frac{\sum (\tilde{y}_i - \tilde{y}_i^{r_0})^2}{\sum (\tilde{y}_i - \overline{\tilde{y}})^2} \qquad (68\text{-c})$$

$$k = \frac{\sum (y_i \tilde{y}_i)}{\sum (\tilde{y}_i)^2} \qquad (68\text{-d})$$

$$k' = \frac{\sum (y_i \tilde{y}_i)}{\sum (y_i)^2} \qquad (68\text{-e})$$

où R est le coefficient de corrélation entre les valeurs calculées et expérimentales dans l'ensemble de validation; R_0^2 (valeurs calculées par rapport aux observées) et R'^2_0 (valeurs observées par rapport aux calculées) sont les coefficients de détermination; k et k' sont les pentes des droites de régressions passant par l'origine pour les valeurs calculées par rapport aux valeurs observées et observées par rapport aux calculées, respectivement. $y_i^{r_0}$ et $\tilde{y}_i^{r_0}$ sont tels que définis respectivement par : $y_i^{r_0} = k\,\tilde{y}$ et, $\tilde{y}_i^{r_0} = k'\,y$; et les sommations sont sur tous les échantillons dans l'ensemble de validation.

La raison d'utiliser et d'exiger des valeurs de k qui sont proches de 1 est que lorsque sont comparées les propriétés réelles aux prédites, un ajustement précis est nécessaire, non seulement une corrélation.

Une fois tous les Tests statistiques susmentionnés effectués et vérifiés, il est essentiel de déterminer quand on peut utiliser le modèle développé. Le domaine d'application (AD) (Tropsha *et al.,* 2003; Shen *et al.,* 2004) est une région théorique dans l'espace définie par les descripteurs du modèle et la réponse modélisée, pour lequel un modèle QSPR donné devrait faire des prédictions fiables. Dans ce travail, l'AD structurel a été vérifié par l'approche des leviers (h_i) (Weisberg, 2005). Les leviers notés h_i sont calculés comme suit:

$$h_i = x_i \left(X^T X \right) x_i^T \tag{69}$$

Où x_i est le vecteur ligne des descripteurs du composé i et **X** la matrice des valeurs des descripteurs de l'ensemble de calibrage; l'indice T désigne le vecteur (ou la matrice) transposé (e).

Un avertissement sur l'effet de levier important d'un échantillon est, en général, donné pour un h_i¨>= h*= 3(p + 1)/n où n est le nombre total d'observations dans l'ensemble de calibrage et p le nombre de descripteurs impliqués dans la corrélation.

La présence de valeurs aberrantes en réponse (valeurs aberrantes en **Y**) et les composés structurellement influents (valeurs aberrantes en **X**) a été vérifiée par le diagramme de Williams (SCAN- Software for Chemometric Analysis - 1995), le tracé des résidus standardisés de prédiction (éq. 70) en fonction des valeurs des leviers.

$$e_{i\,STD} = \frac{e_{(i)}}{s \cdot \sqrt{1 - h_i}} \tag{70}$$

PARTIE B: APPLICATIONS

Dans cette partie nous allons tenter de modéliser le facteur acentrique de Pitzer (ω) et le volume critique (V_C) de divers composés organiques et les modèles obtenus seront comparés à ceux de la littérature et plus spécialement aux modèles issus de la méthode de contribution de groupes.

I.1 Collecte de données:

Les valeurs du facteur acentrique (ω) et du volume critique (V_C) ont été prélevées de la 5ème édition du livres de Poling B. E. et coauteurs: **THE PROPERTIES OF GASES AND LIQUIDS (**Poling *et al*, 2001) en prenant soin de ne prendre que les valeurs expérimentalement mesurées pour V_C et les valeurs ω calculées avec des équations de la pression de vapeur où des extrapolations de moins de 10 Kelvin sont nécessaires. Nous obtenons donc **196** points de données pour le volume critique et **158** pour le facteur acentrique pour un total de **265** composés différents.

I.2 Optimisation des géométries moléculaires:

Nous avons utilisé le logiciel de modélisation moléculaire HyperChem 6.03 (HyperChem^{TM} Release 6.03 for Windows, 2000) pour représenter les **265** molécules puis, à l'aide de la méthode semi-empirique PM3 (Dewar *et al.,* 1985), on a obtenu les géométries finales. Tous les calculs ont été menés dans le cadre du formalisme RHF (Levine, 2000) sans interaction de configuration. Les structures moléculaires ont été prèoptimisées à l'aide de l'algorithme Polak- Ribiere avec pour critère d'arrêt une racine du carré moyen du gradient égale à 0,001kcal/mol.

I.3 Calcul des descripteurs:

Les géométries ainsi optimisées ont été transférées dans le logiciel Dragon version 5.3 (Todeschini *et al.,* 2006) pour le calcul des 1664 descripteurs disponibles appartenant à différentes classes. Les descripteurs d'un même groupe, à valeur constante (écarts types inférieurs à 0,0001) ont été exclus. Pour un seuil de corrélation de $r \geq 0,95$ entre deux descripteurs ; celui qui présente le plus de corrélations avec les autres variables, est exclu.

I.4 Sous - ensembles de calibrages et validations:

En vu de valider les modèles QSPR, les ensembles de données du facteur acentrique et du volume critique ont été séparés à l'aide de l'algorithme CADEX (Kennard et Stone, 1969) en lui donnant comme entrées les deux matrices $\boldsymbol{X}$ des deux réponses à modéliser. Ainsi les données ont été partagées en **110** composés destinés à la recherche du modèles pour ω (**137** pour V_C) et **48** composées (c.-à-d. 30% des données disponibles) pour leurs validations (**59** pour V_C).

I.5 Procédure pour l'obtention et l'évaluation des modèles QSPR:

L'analyse de régression linéaire multiple (MLR) basée sur les moindres carrés ordinaires (MCO) et la sélection des variables ont été effectuées par le logiciel MobyDigs (Todeschni, 2009) et l'algorithme génétique (GA-VSS) pour la sélection des sous-ensembles de variables explicatives (Leardi *et al.,* 1992).

La RLM a été utilisée comme technique linéaire, alors que les réseaux de neurones artificiels (RNA) ont été employés comme technique non linéaire pour la construction des modèles QSPR dans le cas de ω.

L'application ANN (pour: Artificial Neural Network) du logiciel Molegro Data Modeller (Molegro; 2009) a été employée pour la modélisation non linéaire par RNA: obtention de l'architecture optimale du réseau et prédiction du facteur acentrique.

Dans les deux méthodes de régressions les modèles ont été justifiés par le R^2, le R^2 ajusté, les valeurs de la validation croisée de Q^2 par leave-one-out (LOO), les valeurs de ratio F et l'erreur standard s. Le $Q^2_{\text{bootstrap}}$ a été utilisé seulement dans la RLM.

Les modèles proposés ont également été vérifiés pour la fiabilité et la robustesse par un teste de permutation: les nouveaux modèles sont recalculés pour une réponse enregistrée de façon aléatoire (Y- scrambling) en utilisant la même matrice des variables indépendantes d'origine. Après avoir répété ce teste plusieurs fois (100 fois dans ce travail), il est prévu d'obtenir de nouveaux modèles qui ont des R^2 et Q^2 nettement inférieur que le modèle original. Si cette condition n'est pas vérifiée les modèles originaux ne sont pas acceptables, car ils étaient dus à des corrélations hasardeuses ou à des redondances structurelles dans les ensembles de calibrages.

L'obtention d'un modèle robuste ne donne pas des informations réelles sur son pouvoir de prédiction. Ceci est évalué en prédisant les composés inclus dans l'ensemble de validation.

L'analyse des résidus présente un intérêt à plusieurs égards. Elle permet en effet de vérifier, a posteriori, la validité des modèles utilisés. Cette analyse à été faite en utilisant le diagramme de Williams pour en même temps définir les domaines d'applications des modèles obtenus par régression linéaire.

II.1 Modèle par régression linéaire multiple:

II.1.1 Taille du modèle:

L'application du GA-VSS a conduit à plusieurs modèles sur la base de différents ensembles de descripteurs moléculaires pour les différentes dimensions que nous avons imposés allant de 1 à 10. La figure 1 représente les variations de R^2 et Q^2 (valeurs dans le tableau 1) en fonction du nombre de descripteurs (p) impliqués dans le modèle.

Tableau 1. Valeur de R^2 et Q^2 pour chaque taille du modèle.

p	1	2	3	4	5	6	7	8	9	10
R^2	66,12	72,39	82,65	88,06	91,89	92,93	94,17	94,75	95,60	96,08
Q^2	65,27	70,49	79,96	86,08	90,38	91,47	92,60	93,31	94,43	95,07

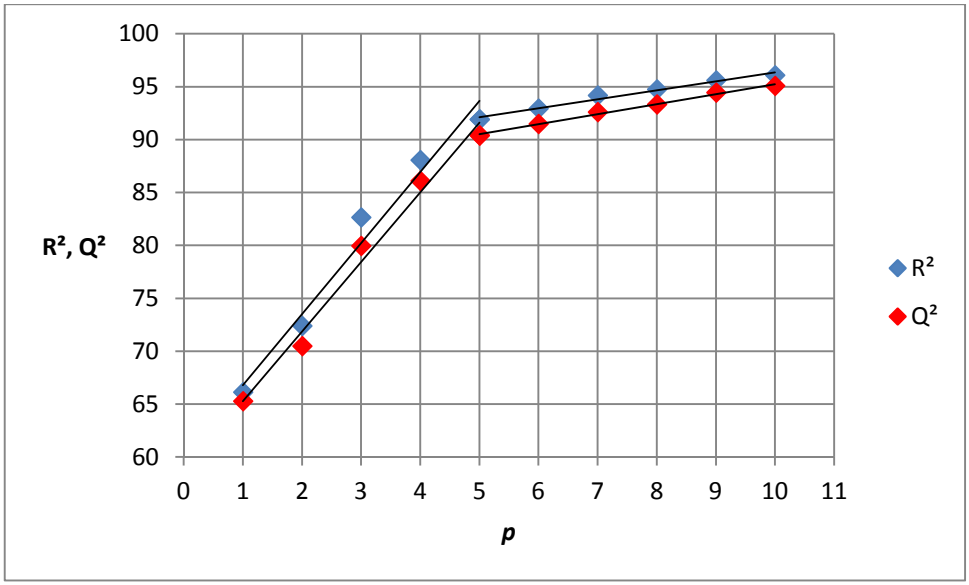


Figure 1: Variation de Q^2 et R^2 en fonction de la taille des modèles (cas de ω)

La taille optimale du modèle du facteur acentrique est clairement identifiée à l'aide de cette procédure et c'est 5 descripteurs.

II.1.2 Choix du modèle:

En utilisant le logiciel Minitab (MINITAB, 2000) pour l'analyse de régression nous obtenons l'équation du modèle et les diagnostiques suivants:

ω = - 0,0013 + 0.00333 Mol. Wt. - 0,0688 MATS2m - 0,0148 ESpm05u + 1,79 R1u+

- 3,17 R1p+ (1)

Régresseurs	Coéf.	ES Coéf.	t	P	VIF
Constante	-0,00128	0,02028	-0,06	0,950	
Mol. Wt.	0,0033261	0,0001038	32,04	0,000	1,276
MATS2m	-0,068838	0,009819	-7,01	0,000	1,252
ESpm05u	-0,014802	0,001930	-7,67	0,000	1,050
R1u+	1,7908	0,1033	17,34	0,000	2,157
R1p+	-3,1703	0,2598	-12,20	0,000	1,867

Les valeurs des FIV sont toutes, comme exigées, inférieures à 5 ce qui exclu la multi-colinéarité des descripteurs sélectionnés qui sont aussi significatifs comme le prouvent les valeurs des probabilités P, associées au Test de Student reportés dans l'analyse, toutes inférieures à α=0,05. On trouvera plus d'informations concernant ces descripteurs dans le guide d'utilisation du logiciel Dragon (Todeschini *et al.,* 2005) et les références afférentes.

II.1.3 Qualité statistique:

Le tableau 2 condense les statistiques jugeant de la qualité du modèle

Tableau 2. Statistiques relatives au modèle QSPR du facteur acentrique.

R^2	Q^2	Q^2_{BOOT}	R^2_{adj}	$Q^2_{L10\%O}$	$Q^2_{L20\%O}$	*EQMC*	*EQMP*	*F*	*S*
91,89	90,38	89,45	91,5	90,29	90,16	0,048	0,052	235,8038	0,0494

Les paramètres statistiques montrent que les cinq descripteurs permettent de corréler les facteurs acentriques des 110 composés. En effet, la valeur du coefficient de détermination signifie que 91,89 % de la variabilité ω, peut être expliquée. La grande valeur du F de Fisher indique que le modèle est très significatif. La validation interne du modèle est vérifiée par Q^2 , $Q^2_{L10\%O}$ et $Q^2_{L20\%O}$ qui ont de grandes valeurs. Avec ces dernières; la valeur de Q^2_{BOOT} montre que le modèle est stable car toutes les valeurs sont très proches entre elles et similaires à R^2 et R^2_{adj}. Les écarts quadratiques moyens sont porches et faibles.

Le tableau 3 réunit les valeurs expérimentales ω_{EXP}, calculées ω_{CAL} et prédites ω_{PRED} (de l'ensemble de calibrage) du facteur acentrique ainsi que les valeurs des leviers h_i et des résidus standardisés de prédictions $e_{i\ STD}$.

Tableau 3. Valeurs de ω_{EXP} , ω_{CAL} , ω_{PRED} , h_i et $e_{i\ STD}$ pour l'ensemble de calibrage.

N°	Composé	ω_{EXP} ,	ω_{CAL}	ω_{PRED}	h_i	$e_{i\ STD}$
1	Chlorotrifluoromethane	0,175	0,2574	0,2668	0,102	1,9587
2	Trichloromonofluoromethane	0,195	0,3224	0,3332	0,078	2,9114
3	Tetrafluoromethane	0,177	0,1930	0,1948	0,099	0,3795
4	Fluoroform	0,267	0,1889	0,1845	0,054	-1,7176
5	Difluoromethane	0,278	0,1627	0,1533	0,075	-2,6237
6	Chloromethane	0,151	0,1288	0,1268	0,085	-0,5123
7	Methyl fluoride	0,204	0,1298	0,1209	0,108	-1,7810
8	Methane	0,011	0,0521	0,0595	0,153	1,0672
9	methanol	0,565	0,6354	0,6585	0,248	2,1817
10	Methanethiol	0,150	0,0872	0,0237	0,503	-3,6227
11	Pentafluoroethyl chloride	0,251	0,2616	0,2647	0,221	0,3132
12	Acetylene	0,189	0,1715	0,1705	0,050	-0,3835
13	1,1,1-tri?uoroethane	0,259	0,2969	0,2987	0,047	0,8236
14	ethanoic acid	0,445	0,4985	0,5044	0,099	1,2668
15	Ethane	0,099	0,1897	0,1941	0,046	1,9699
16	1,2-Propadiene	0,122	0,2027	0,2087	0,070	1,8201
17	Propene	0,142	0,1893	0,1921	0,057	1,0447
18	Cyclopropane	0,130	0,1999	0,2024	0,035	1,4905
19	Propylamine	0,283	0,3723	0,3766	0,046	1,9395
20	1,3-Butadiene	0,195	0,2038	0,2042	0,046	0,1902
21	Cyclobutane	0,185	0,2044	0,2053	0,045	0,4208
22	cis-2-butene	0,203	0,2237	0,2247	0,046	0,4490
23	Isopropyl Alcohol	0,665	0,5459	0,5291	0,124	-2,9379
24	Ethyl Acetate	0,361	0,3061	0,3039	0,038	-1,1784
25	propyl methanoate	0,320	0,3262	0,3264	0,027	0,1312
26	2-chlorobutane	0,267	0,2829	0,2841	0,069	0,3589
27	Butane	0,200	0,2107	0,2112	0,044	0,2318
28	Isobutane	0,186	0,1669	0,1662	0,033	-0,4064
29	Diethyl sulfide	0,295	0,2874	0,2862	0,138	-0,1927
30	1-pentene	0,237	0,2707	0,2722	0,040	0,7260
31	2-methyltetrahydrofuran	0,292	0,3417	0,3440	0,044	1,0772
32	1-Propanol	0,629	0,5752	0,5687	0,107	-1,2898
33	pentane	0,252	0,2534	0,2534	0,037	0,0295
34	2-methylbutane	0,229	0,2028	0,2021	0,026	-0,5518
35	neopentane	0,197	0,1692	0,1681	0,036	-0,5955
36	3-methyl-1-butanol	0,559	0,5305	0,5284	0,071	-0,6428
37	ethyl propyl ether	0,328	0,4078	0,4123	0,054	1,7538
38	benzene	0,210	0,2418	0,2430	0,037	0,6809
39	cyclohexane	0,211	0,2717	0,2740	0,036	1,2985
40	1-hexene	0,281	0,2905	0,2909	0,035	0,2038
41	propyl propanoate	0,373	0,3979	0,3985	0,022	0,5220

Tableau 3 suite

N°	Composé	ω_{EXP} ,	ω_{CAL}	ω_{PRED}	h_i	$e_{i\ STD}$
42	3-methylbutyl methanoate	0,400	0,3960	0,3959	0,013	-0,0826
43	hexane	0,300	0,2791	0,2784	0,034	-0,4453
44	2-methylpentane	0,278	0,2433	0,2425	0,021	-0,7254
45	2,2-dimethylbutane	0,233	0,2123	0,2116	0,031	-0,4394
46	1-hexanol	0,573	0,4683	0,4645	0,035	-2,2340
47	4-methyl-2-pentanol	0,552	0,5168	0,5146	0,059	-0,7806
48	toluene	0,264	0,3221	0,325	0,048	1,2654
49	4-methylphenol	0,510	0,4285	0,4245	0,047	-1,7727
50	butyl-2-propenoate	0,312	0,3766	0,3785	0,029	1,3649
51	methylcyclohexane	0,235	0,2735	0,2743	0,019	0,8022
52	ethylcyclopentane	0,270	0,2699	0,2699	0,019	-0,0028
53	propyl butanoate	0,399	0,4307	0,4315	0,024	0,6653
54	heptane	0,350	0,3445	0,3443	0,029	-0,1167
55	2-methylhexane	0,331	0,2891	0,2884	0,017	-0,8700
56	Ethanol	0,649	0,6115	0,6041	0,164	-0,9933
57	3,3-dimethylpentane	0,269	0,2574	0,2571	0,028	-0,2449
58	1-heptanol	0,588	0,5039	0,5006	0,037	-1,8021
59	1,3-dimethylbenzene	0,327	0,3293	0,3294	0,028	0,0487
60	cyclooctane	0,254	0,3409	0,3439	0,033	1,8490
61	1-octene	0,393	0,3943	0,3944	0,029	0,0282
62	octane	0,399	0,3742	0,3735	0,028	-0,524
63	2-methylheptane	0,378	0,3318	0,3311	0,015	-0,9568
64	4-methylheptane	0,371	0,3255	0,3248	0,016	-0,9425
65	2,4-dimethylhexane	0,344	0,3174	0,3169	0,018	-0,5538
66	3-ethyl-2-methylpentane	0,331	0,3167	0,3164	0,019	-0,2975
67	3-ethyl-3-methylpentane	0,305	0,2996	0,2995	0,026	-0,1132
68	2,2,4-trimethylpentane	0,304	0,2975	0,2973	0,026	-0,1365
69	2,3,3-trimethylpentane	0,291	0,2893	0,2893	0,029	-0,0357
70	2,2,3,3-tetramethylbutane	0,248	0,2727	0,2736	0,035	0,5282
71	1-octanol	0,594	0,5807	0,5801	0,044	-0,2880
72	propylbenzene	0,345	0,3500	0,3501	0,02	0,1033
73	1-ethyl-4-methylbenzene	0,364	0,3631	0,3631	0,023	-0,0181
74	1,2,3-trimethylbenzene	0,367	0,3363	0,3356	0,022	-0,6427
75	1,3,5-trimethylbenzene	0,399	0,3586	0,3576	0,025	-0,849
76	3-methylbutyl butanoate	0,583	0,5122	0,5098	0,032	-1,5052
77	nonane	0,445	0,4370	0,4367	0,028	-0,1696
78	2,2-dimethylheptane	0,383	0,3359	0,3348	0,023	-0,9878
79	2,2,3,3-tetramethylpentane	0,304	0,3262	0,3269	0,032	0,4712
80	2,2,3,4-tetramethylpentane	0,301	0,3323	0,3332	0,029	0,6620
81	2,2,4,4-tetramethylpentane	0,314	0,3264	0,3268	0,031	0,2622
82	naphthalene	0,304	0,2947	0,2945	0,022	-0,1948
83	butylbenzene	0,393	0,3317	0,3304	0,021	-1,2804
84	2-methylpropylbenzene	0,383	0,3615	0,3610	0,021	-0,4496

Tableau 3 suite et fin

N°	Composé	ω_{EXP} ,	ω_{CAL}	ω_{PRED}	h_i	$e_{i\ STD}$
85	1,4-diethylbenzene	0,403	0,3494	0,3485	0,018	-1,1135
86	1-(1-methylethyl)-4-methylbenzene	0,376	0,3916	0,3919	0,022	0,3258
87	1,2,4,5-tetramethylbenzene	0,423	0,3668	0,3655	0,023	-1,1766
88	trans-bicyclo[4,4,0]decane	0,303	0,3471	0,3481	0,021	0,9213
89	decane	0,49	0,4595	0,4586	0,03	-0,6459
90	3,3,5-trimethylheptane	0,383	0,3788	0,3787	0,027	-0,0875
91	2,2,3,3-tetramethylhexane	0,366	0,3670	0,3671	0,033	0,0220
92	2,2,5,5-tetramethylhexane	0,377	0,3678	0,3675	0,033	-0,1962
93	1-decanol	0,661	0,7165	0,7208	0,072	1,2562
94	1-methylnaphthalene	0,348	0,4461	0,4502	0,041	2,1120
95	2-methylnaphthalene	0,374	0,4655	0,4711	0,058	2,0243
96	undecane	0,537	0,5103	0,5094	0,034	-0,5687
97	1,1'-biphenyl	0,404	0,4347	0,4355	0,025	0,6455
98	dodecane	0,576	0,5538	0,5529	0,039	-0,4767
99	diphenylmethane	0,481	0,4391	0,4378	0,031	-0,8876
100	tridecane	0,618	0,5991	0,5982	0,045	-0,4105
101	phenanthrene	0,479	0,4420	0,4407	0,034	-0,7893
102	anthracene	0,501	0,4364	0,4341	0,034	-1,3771
103	tetradecane	0,644	0,626	0,625	0,054	-0,3949
104	pentadecane	0,685	0,6935	0,6941	0,065	0,1903
105	hexadecane	0,718	0,7368	0,7384	0,076	0,4292
106	2,2,4,4,6,8,8-heptamethylnonane	0,548	0,6333	0,6402	0,074	1,9381
107	heptadecane	0,753	0,7379	0,7364	0,089	-0,3512
108	octadecane	0,800	0,8048	0,8054	0,104	0,1149
109	nonadecane	0,845	0,8533	0,8544	0,120	0,2027
110	eicosane	0,865	0,8954	0,9002	0,138	0,7681

II.1.4 Validation externe du modèle

L'application du modèle obtenu (éq. 1) aux 48 composés de l'ensemble de validation conduit aux résultats réunit dans le tableau 4.

Tableau 4. Valeurs de ω_{EXP} , ω_{PRED} , h_i et $e_{i\ STD}$ pour l'ensemble de validations.

N°	Composé	ω_{EXP} ,	ω_{PRED}	h_i	$e_{i\ STD}$
111	Methylamine	0,283	0,3884	0,069	2,2104
112	1,1-difluoroethane	0,276	0,2815	0,044	0,1128
113	Acetone	0,307	0,2821	0,059	-0,5190
114	Propane	0,152	0,1737	0,051	0,4505
115	1-Butene	0,194	0,2271	0,047	0,6865
116	trans-2-butene	0,218	0,2197	0,045	0,0350
117	2-methylpropene	0,199	0,1711	0,036	-0,5757

Tableau 4 suite et fin

N°	Composé	ω_{EXP} ,	ω_{PRED}	h_i	$e_{i\ STD}$
118	1-butanol	0,590	0,5673	0,081	-0,4792
119	2-methyl-1-propanol	0,590	0,4440	0,052	-3,0338
120	2-methyl-2-propanol	0,613	0,5406	0,129	-1,5686
121	diethyl ether	0,281	0,3847	0,064	2,1694
122	1-Butanamine	0,338	0,4077	0,038	1,4370
123	1-pentyne	0,394	0,3213	0,085	-1,5384
124	2-methyl-2-butene	0,339	0,2061	0,028	-2,7269
125	3-methyl-1-butene	0,211	0,2140	0,030	0,0613
126	2-pentanone	0,346	0,3337	0,041	-0,2540
127	3-pentanone	0,342	0,3376	0,041	-0,0901
128	ethyl propanoate	0,390	0,3691	0,024	-0,4278
129	1-pentanol	0,579	0,6161	0,085	0,7847
130	methylcyclopentane	0,227	0,2317	0,023	0,0952
131	4-methyl-2-pentanone	0,351	0,3434	0,040	-0,1575
132	ethyl butanoate	0,463	0,3757	0,023	-1,7860
133	butyl ethanoate	0,407	0,4104	0,021	0,0692
134	2-methylpropyl ethanoate	0,456	0,4013	0,023	-1,1199
135	3-methylpentane	0,273	0,2283	0,022	-0,9138
136	2,3-dimethylbutane	0,248	0,2213	0,026	-0,5480
137	2-methyl-1-pentanol	0,498	0,5100	0,048	0,2499
138	3-methylhexane	0,323	0,2855	0,018	-0,7664
139	2,2-dimethylpentane	0,287	0,2545	0,027	-0,6657
140	2,3-dimethylpentane	0,297	0,2726	0,021	-0,4996
141	2,4-dimethylpentane	0,304	0,2734	0,020	-0,6255
142	2,2,3-trimethylbutane	0,250	0,2499	0,029	-0,0023
143	1,4-dimethylbenzene	0,322	0,3223	0,025	0,0059
144	3-methylheptane	0,371	0,3277	0,016	-0,8827
145	3-ethylhexane	0,362	0,3286	0,016	-0,6804
146	2,2-dimethylhexane	0,339	0,2859	0,024	-1,0873
147	2,3-dimethylhexane	0,347	0,3042	0,019	-0,874
148	2,5-dimethylhexane	0,357	0,3223	0,017	-0,7092
149	3,3-dimethylhexane	0,320	0,3010	0,025	-0,3888
150	3,4-dimethylhexane	0,338	0,3071	0,020	-0,6320
151	2,2,3-trimethylpentane	0,298	0,2927	0,028	-0,1086
152	2,3,4-trimethylpentane	0,316	0,3072	0,021	-0,1802
153	1-methylethylbenzene	0,326	0,3322	0,022	0,1277
154	1,2,4-trimethylbenzene	0,377	0,3507	0,023	-0,5385
155	cis-bicyclo[4,4,0]decane	0,276	0,3529	0,020	1,5718
156	2-Butanone	0,322	0,2907	0,048	-0,6495
157	propyl acetate	0,389	0,3714	0,023	-0,3609
158	3-ethylpentane	0,311	0,2836	0,018	-0,5588

Les statistiques relatives à l'ensemble de validation sont : $Q^2_{EXT} = 91,35\%$ très proches de R^2 et Q^2 ce qui démontre la capacité de notre modèle à prédire des composés qui n'ont pas servis à son calibrage. La valeur de $EQMP_{EXT} = 0,05$ est presque la même que $EQMC$ et $EQMP$.

Les statistiques de la validation externe fournies par MobyDigs ne sont pas seules capables de prouver la validité du modèle. Une autre méthode plus rigoureuse a été appliquée, c'est celle de Golbraikh et Tropsha dont les résultats sont :

1) $R^2_{CV\,EXT} = 0,7778 > 0,5$
2) $R^2 = 0,7876 > 0,6$
3) $\left(R^2 - R_0^2\right)/R^2 = -0,2403 < 0,1$ et $0,85 < k = 0,9424 < 1,15$

   ou

   $\left(R^2 - R_0'^2\right)/R^2 = -0,2177 < 0,1$ et $0,85 < k' = 0,9424 < 1,15$
4) $\left|R_0^2 - R_0'^2\right| = 0,0178 < 0,3$

Tous les critères sont remplis et nous pouvons maintenant conclure définitivement de la validité de notre modèle proposé.

II.1.5 Qualité de l'ajustement:

La qualité de l'ajustement a été vérifiée par le graphe des valeurs calculées et prédites du facteur acentrique en fonction des celles expérimentales. Le graphe présenté dans la figure 2, fait ressortir une faible dispersion autour de la première bissectrice ce qui prouve le bon ajustement du modèle et pour les deux sous-ensembles.

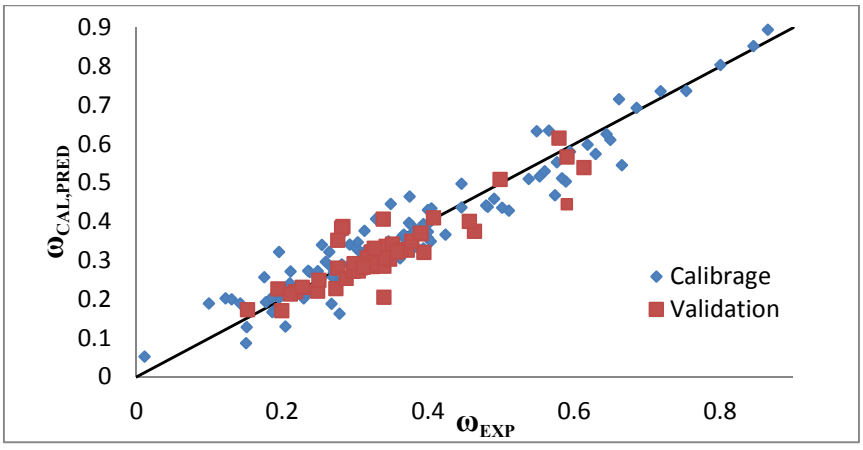


Figure 2: **Qualité de l'ajustement** ***(cas de ω)***

II.1.6 Domaine d'application:

Comme indiqué précédemment, l'approche des leviers est l'un des meilleurs outils de définition du domaine d'application et la figure 3 en est la représentation.

Sur le diagramme de Williams (fig. 3) 3 composés de l'ensemble de calibrage sont signalés influents (avec des $h_i > h^* = 0{,}164$) et 2 composés du même ensemble et un de validation sont proches de la limites des $e_{i\,STD} = \pm 3$. Deux composés sont aberrant en Y et ils sont de calibrage. Ces remarques sont résumées dans le tableau 5.

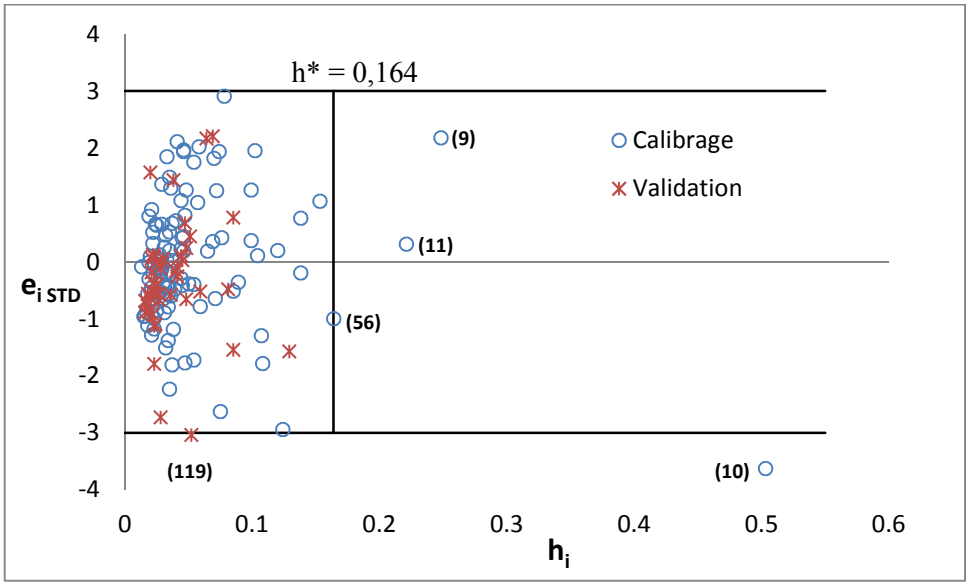


Figure 3: **Diagramme de Williams** ***(cas de ω)***

Tableau 5. Observations aberrantes signalées dans le modèle:

Observations aberrantes	
En X	En Y
(9) Methanol	(119) 2-methyl-1-propanol *
(10) Methanethiol	
(11) Pentafluoroethyl chloride	
(56) Ethanol	

* *Composé de validation*

II.1.7 Test de randomisation:

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

$$y = (y_1, y_2, \ldots, y_{109}, y_{110}) \ \underline{RND} \ y_{RND} = (y_8, y_5, \ldots y_{27}, y_3) \qquad (2)$$

et utilisées comme sources d'observations pour des modèles QSPR dans les conditions optimales établies.

La figure 4 qui représente le graphique des coefficients statistiques Q^2 et R^2 permet de comparer les résultats obtenus pour les modèles randomisés au modèle réel de départ. Il est clair que les statistiques obtenues pour les vecteurs modifiés de ω sont plus petites que celles du modèle QSPR réel (le point isolé sur la figure), ce qui permet d'affirmer que le modèle proposé n'est pas aléatoire.

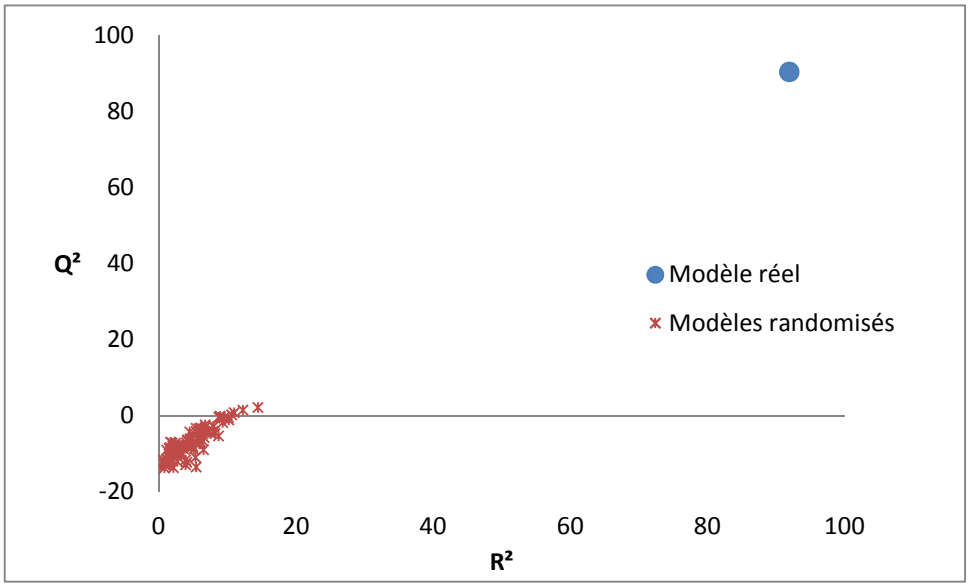


Figure 4: **Test de randomisation** ***(cas de ω).***

II.1.8 Contributions relatives des descripteurs:

En se basant sur une procédure décrite dans la littérature (Zheng *et al.,* 2006; Guha et Jurs, 2005), les contributions relatives des descripteurs du modèle ont été déterminées et sont représentées dans la figure 5. Elles diminuent selon l'ordre suivant: (1) Mol. Wt. (36,84%)> (4) R1u+ (18,83%) > (5) R1p+ (15,96%) > (3) ESpm05u (15,01%)> (2) MATS2m (13,36%).

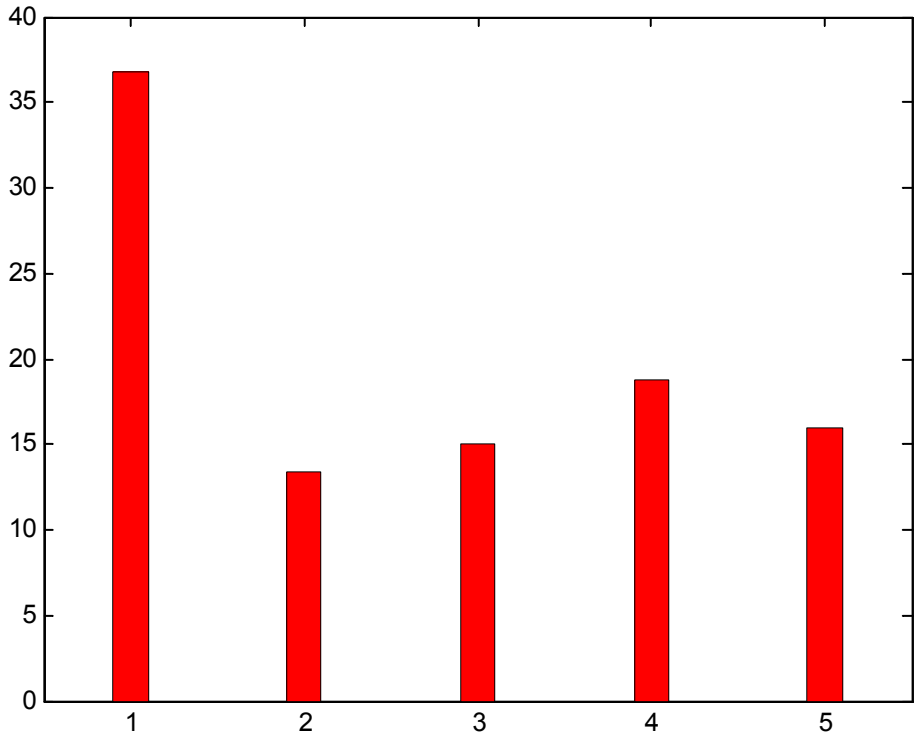


Figure 5: **Contributions relatives des descripteurs** ***(cas de ω).***

Des pourcentages mentionnés auparavant et à partir de la figure 5 nous remarquons que la masse moléculaire est le descripteur de plus important (36,84 %). Les 4 autres ont presque la même contribution.

II.1.9 Définition et interprétation des descripteurs:

Les descripteurs du modèle QSPR facteur acentrique et leurs classes et types sont dans le tableau 6.

Tableau 6: Types et classes des descripteurs.

Descripteur	Type	Classe
Mol. Wt.: masse moléculaire	2D	Descripteurs constitutionnels
MATS2m : autocorrélation de Moran de distance topologique égale à 2/pondérée par les masses atomiques		Indices d'autocorrelation 2D
ESpm05u: moment spectral 05 de matrice de contiguïté de bord		Indices de contiguïté de bord
R1u+: autocorrélation maximale R de distance topologique égale à 1 / non pondérée	3D	Descripteurs GETAWAY
R1p+: autocorrélation maximale R de distance topologique égale à 1/ pondérée polarisabilités atomiques		

Les autocorrélations 2D sont des descripteurs moléculaires qui décrivent comment une propriété considérée est distribuée le long d'une structure moléculaire topologique et les autocorrélations MATSkw de Moran (Moran, 1950), w étant la propriété atomique employée pour pondérer le graphe moléculaire et k la distance topologique, sont calculés en appliquant le coefficient de Moran au graphe moléculaire :

$$MATSkw = \frac{\frac{1}{\Delta}\sum_{i=1}^{nSK}\sum_{j=1}^{nSK}\delta_{ij}\left(w_i - \bar{w}\right)\left(w_j - \bar{w}\right)}{\frac{1}{nSK}\sum_{i=1}^{nSK}\left(w_i - \bar{w}\right)^2} \qquad (3)$$

où w_i est n'importe quelle propriété atomique, $\bar{w}$ est sa valeur moyenne dans la molécule, nSK est le nombre d'atomes (hydrogènes exclus), δ_{ij} est le delta de Kronecker ($\delta_{ij} = 1$ si $d_{ij} = k$, zéro autrement, d_{ij} étant la distance topologique entre deux atomes considérés). Δ est la somme des deltas de Kronecker, c.-à-d. le nombre de paires d'atome à la distance égale à k. L'autocorrélation spatiale positive correspond aux valeurs positives du coefficient tandis que l'autocorrélation spatiale négative produit des valeurs négatives.

Les indices de contiguïté de bord sont des descripteurs moléculaires calculés à partir de la matrice de contiguïté de bord d'une molécule. La matrice de contiguîté de bord est

dérivée du graphe moléculaire (hydrogène exclu) et code la connectivité entre les bords de graphe. C'est une matrice symétrique carrée de dimension BxB, où B est le nombre de liaisons entre les paires d'atomes non-hydrogène. Les entrées de la matrice égalent un si les liaisons considérés sont adjacentes et zéro autrement.

Les moments spectraux de la matrice de contiguïté de bord (ESpmku, ESpmkx, ESpmkd, ESpmkr) sont calculés en additionnant les éléments diagonaux de la k^{ème} puissance de la matrice de contiguîté de bord. L'ordre k du moment spectral est donné par l'ordre de puissance de la matrice (Estrada, 1996). Le moment spectral d'ordre k peut être exprimé comme combinaison linéaire des comptes de différents fragments structuraux (sous-graphes) dans le graphe. Par exemple, le moment spectral d'ordre zéro correspond au nombre de bords dans le graphe. Plusieurs relations entre les moments spectraux et les comptes de fragment ont été dérivées par Estrada (Estrada, 1998; Markovic et Gutman I, 1999)

Proposés par Consonni *et al.* (Consonni *et al.*, 2002a; 2002b); les descripteurs GETAWAY (GEometry, Topology, and Atom-Weights AssemblY)) en tant que descripteurs de structure chimique dérivés d'une nouvelle représentation de la structure moléculaire, la Matrice d'Influence Moléculaire (MIM), dénotés par **H.** Parmi ses descripteurs, les indices maximaux R qui sont des descripteurs moléculaires pour des corrélations structure-propriété, mais ils peuvent également être employés en tant que profils moléculaires appropriés à l'analyse et à l'étude de la similarité/ diversité. Ces descripteurs, comme basés sur l'autocorrélation spatiale, codent l'information sur les fragments structuraux et semblent donc être particulièrement appropriés à décrire les différences dans des séries de congénères molécules. Les descripteurs GETAWAY sont des descripteurs géométriques codant l'information sur la position effective des substituant et des fragments dans l'espace moléculaire. D'ailleurs, ils sont indépendants de l'alignement de la molécule et, par extension, expliquent également l'information sur la taille et la forme moléculaires aussi bien que pour les propriétés atomiques spécifiques.

Les deux descripteurs GETAWAY du modèle se calculent comme suit:

$$\boldsymbol{Rkw+} = \max_{ij}\left(\frac{\sqrt{\boldsymbol{h}_{ii}\cdot\boldsymbol{h}_{jj}}}{\boldsymbol{r}_{ij}}\boldsymbol{w}_i\cdot\boldsymbol{w}_j\cdot\delta\left(\boldsymbol{k},\boldsymbol{d}_{ij}\right)\right)\boldsymbol{i}\neq\boldsymbol{j}\text{ et }\boldsymbol{k}=1,2,...,8 \qquad (4)$$

ici d_{ij} est la distance topologique entre les atomes i et j; $\boldsymbol{w}_i$ est un poids physico-chimique atomique; δ (k ; d_{ij}) est la fonction delta de Dirac (δ=1 si d_{ij} = k, zéro autrement); h_{ii} et h_{jj} sont

les leviers (éléments diagonaux de la matrice **H**) des deux atomes considérés, et r_{ij} est leur distance interatomique.

II.2 Modèle par réseau de neurones artificiels:

Le modèle obtenu par régression linéaire est de bonne qualité mais dans cette partie nous allons essayer de l'améliorer en utilisant les réseaux de neurones artificiels. Les entrées que nous allons utiliser sont les descripteurs choisis dans le modèle linéaire (c.-à-d.: Mol. Wt.; R1u; R1p+ ; ESpm05u et MATS2m) comme s'est l'usage dans de nombreux travaux (Fernandez M. *et al.*, 2006; Manallack D. T. *et al.*, 1994; Dastmalchi S. *et al.*, 2012).

Pour trouver l'architecture du modèle; la première étape est de trouver le nombre optimal de neurones. Nous avons fais varié le nombre de neurones de 1 à 10 et tracé le graphiques (Fig. 6) de la modification des écarts quadratiques moyen (***EQM***) de calcul, de validation interne par LOO et de validation externe (respectivement: ***EQMC*** ; ***EQMP*** et $\boldsymbol{EQMP_{EXT}}$). Le nombre d'itérations à été fixé à 100 qui est la valeur minimale permise dans le logiciel Molegro Data modeller.

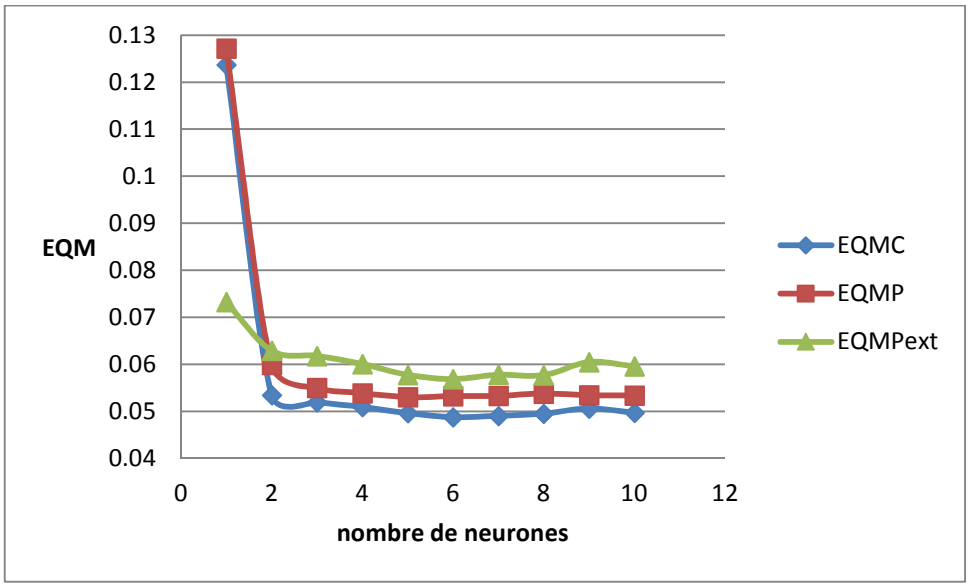


Figure 6: Variation des EQM en fonction de nombre de neurones.

La diminution des écarts en passant de 1 à 2 neurones est très significative alors qu'à partir de 3 ces erreurs commencent à se stabiliser ce qui nous fait dire que 2 est le nombre de neurones à choisir. Le tableau regroupant les valeurs de ces EQM est en annexe.

La deuxième étape est de choisir le nombre d'itérations maximum à utiliser pour l'entrainement du réseau donc son apprentissage afin de régler les poids du réseau à 5 entrées et 2 neurones.

De la même manière qu'à été trouvé le nombre de neurones, nous avons fais varier le nombre d'itérations de 100 à 2100 par pas de 10 itérations. Puisqu'il y a 200 modèles à examiner nous avons choisi ici pour des raisons pratiques de tracer seulement la variation pour les itérations de 1640 à 1910 (Fig. 7). Les valeurs des EQM sont aussi en annexe.

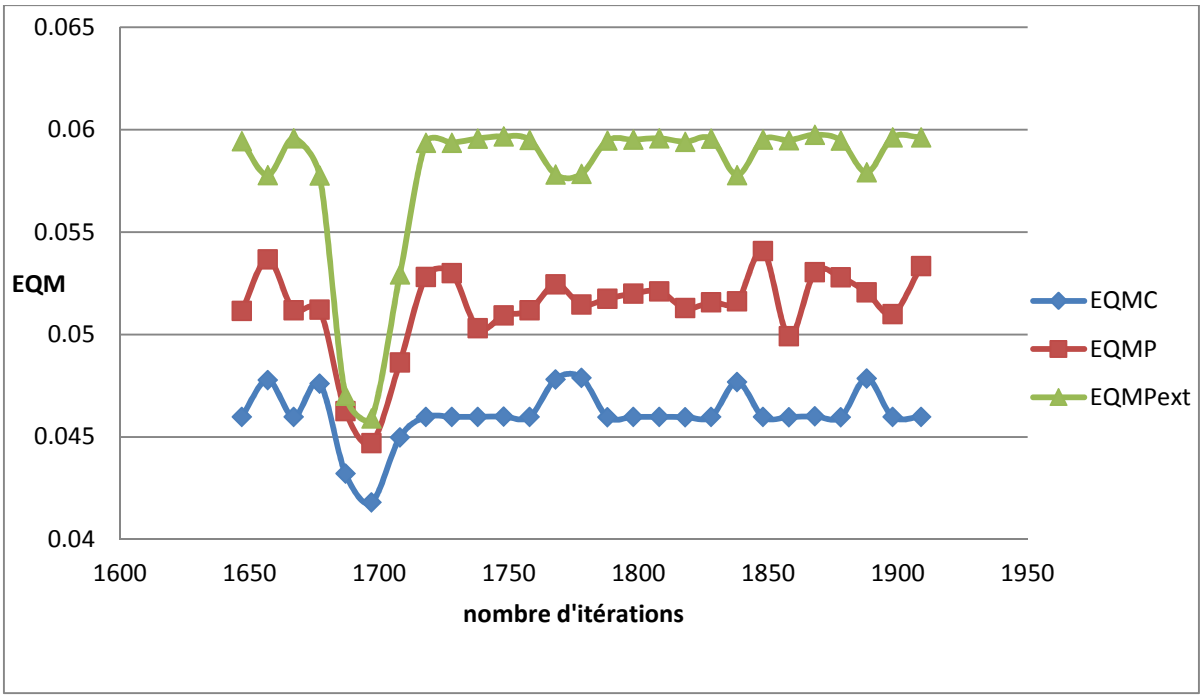


Figure 7:Variation des EQM en fonction de nombre d'itérations.

Le nombre d'itérations nécessaire à l'entrainement du réseau est 1697 pour lequel nous obtenons les écarts les plus faibles. Ainsi le réseau optimal pour prédire le facteur acentrique est à 5 entées (les descripteurs précédemment choisis par GA-VSS), une couche cachée comptant 2 neurones et enfin une sortie qui est le facteur acentrique. Le modèle a pour équation:

1) Les descripteurs sont modifiés et donnés en entrées (E_i: i=0 à 4) comme suit:

   E_0 = "Mol. Wt."*0.00375219 - 0.0601964 (5a)

   E_1 = MATS2m*0.5 + 0.5 (5b)

   E_2 = ESpm05u*0.147232 + 0 (5c)

   E_3 = "R1u+"*2.34192 + 0 (5d)

   E_4 = "R1p+"*4.329 + 0 (5e)

2) Dans les 2 neurones de la couche cachée (CC_0 et CC_1) les entrées sont aussi modifiées:

CC_0 = sigmoid(-2.75335*E_0 + 0.855124*E_1 - 0.0859207*E_2 - 1.56641*E_3 + 3.03706*E_4) (5f)

CC_1 = sigmoid(-0.246086*E_0 + 1.58412*E_1 - 1.64019*E_2 + 2.64191*E_3 + 1.8422*E_4) (5g)

3) la sortie qui est la valeur prédite du facteur acentrique est calculée par la dernière équation:

Sortie = (sigmoid(-7.49517*CC_0 + 3.69615*CC_1) + 0.0128806) / 1.17096 (5h)

Dans les équations (5a à 5h), sigmoid symbolise la fonction d'activation sigmoïde (Yonaba *et al.* 2010) utilisée dans le logiciel Molegro Data Modeller.

Les valeurs expérimentales calculées et prédites par LOO ($\omega_{PRED/LOO}$ pour l'ensemble de calibrage) sont réunies dans le tableau 7.

Tableau 7. Facteurs acentriques calculés et prédits par me modèle RNA des 158 composés modélisés.

N°	Composé	ω_{EXP}	$\omega_{PRED/LOO}$	$\omega_{CAL/PRED}$	ID	Composé	ω_{EXP}	$\omega_{PRED/LOO}$	$\omega_{CAL/PRED}$
1	Chlorotrifluoromethane	0,175	0,169524	0,178738	80	2,2,3,4-tetramethylpentane	0,301	0,317812	0,319238
2	Trichloromonofluoromethane	0,195	0,238505	0,236252	81	2,2,4,4-tetramethylpentane	0,314	0,307207	0,311751
3	Tetrafluoromethane	0,177	0,137775	0,156875	82	naphthalene	0,304	0,282624	0,279461
4	Fluoroform	0,267	0,182164	0,276235	83	butylbenzene	0,393	0,318455	0,316423
5	Difluoromethane	0,278	0,192799	0,179400	84	2-methylpropylbenzene	0,383	0,350974	0,351512
6	Chloromethane	0,151	0,159242	0,158579	85	1,4-diethylbenzene	0,403	0,345982	0,34523
7	Methyl fluoride	0,204	0,213179	0,225070	86	1-(1-methylethyl)-4-methylbenzene	0,376	0,390226	0,391194
8	Methane	0,011	0,013447	0,013519	87	1,2,4,5-tetramethylbenzene	0,423	0,364115	0,364121
9	methanol	0,565	0,542122	0,523303	88	trans-bicyclo[4,4,0]decane	0,303	0,343589	0,345600
10	Methanethiol	0,150	0,162564	0,156517	89	decane	0,490	0,465237	0,463622
11	Pentafluoroethyl chloride	0,251	0,266901	0,254232	90	3,3,5-trimethylheptane	0,383	0,372233	0,374496
12	Acetylene	0,189	0,180027	0,154533	91	2,2,3,3-tetramethylhexane	0,366	0,348586	0,348317
13	1,1,1-tri?uoroethane	0,259	0,257041	0,258988	92	2,2,5,5-tetramethylhexane	0,377	0,349358	0,348009
14	ethanoic acid	0,445	0,496050	0,504784	93	1-decanol	0,661	0,710320	0,713681
15	Ethane	0,099	0,090948	0,097280	94	1-methylnaphthalene	0,348	0,435260	0,442296
16	1,2-Propadiene	0,122	0,136692	0,118352	95	2-methylnaphthalene	0,374	0,437162	0,465262
17	Propene	0,142	0,145543	0,146800	96	undecane	0,537	0,528595	0,52876
18	Cyclopropane	0,130	0,193336	0,178826	97	1,1'-biphenyl	0,404	0,438624	0,460833
19	Propylamine	0,283	0,374769	0,382838	98	dodecane	0,576	0,574630	0,576536
20	1,3-Butadiene	0,195	0,172237	0,172014	99	diphenylmethane	0,481	0,442315	0,440119
21	Cyclobutane	0,185	0,180025	0,180618	100	tridecane	0,618	0,617342	0,619814
22	cis-2-butene	0,203	0,177047	0,175573	101	phenanthrene	0,479	0,440838	0,442525
23	Isopropyl Alcohol	0,665	0,663341	0,644938	102	anthracene	0,501	0,433684	0,425138
24	Ethyl Acetate	0,361	0,287372	0,276781	103	tetradecane	0,644	0,643102	0,647853
25	propyl methanoate	0,320	0,284833	0,288609	104	pentadecane	0,685	0,693818	0,693626
26	2-chlorobutane	0,267	0,256545	0,257297	105	hexadecane	0,718	0,715834	0,715957

Tableau 7 Suite

N°	Composé	ω_{EXP}	$\omega_{PRED/LOO}$	$\omega_{CAL/PRED}$	ID	Composé	ω_{EXP}	$\omega_{PRED/LOO}$	$\omega_{CAL/PRED}$
27	Butane	0,200	0,185694	0,185252	106	2,2,4,4,6,8,8-heptamethylnonane	0,548	0,584754	0,621968
28	Isobutane	0,186	0,168186	0,167096	107	heptadecane	0,753	0,719415	0,716425
29	Diethyl sulfide	0,295	0,258426	0,262625	108	octadecane	0,800	0,748565	0,745040
30	1-pentene	0,237	0,211308	0,212876	109	nonadecane	0,845	0,764519	0,758712
31	2-methyltetrahydrofuran	0,292	0,322513	0,325126	110	eicosane	0,865	0,775630	0,768090
32	1-Propanol	0,629	0,568253	0,546598	111	Methylamine	0,283	/	0,243853
33	pentane	0,252	0,221616	0,221380	112	1,1-difuoroethane	0,276	/	0,265551
34	2-methylbutane	0,229	0,197606	0,196047	113	Acetone	0,307	/	0,262943
35	neopentane	0,197	0,172090	0,170925	114	Propane	0,152	/	0,153787
36	3-methyl-1-butanol	0,559	0,554000	0,553157	115	1-Butene	0,194	/	0,174233
37	ethyl propyl ether	0,328	0,428838	0,433853	116	trans-2-butene	0,218	/	0,236103
38	benzene	0,210	0,214822	0,215794	117	2-methylpropene	0,199	/	0,163784
39	cyclohexane	0,211	0,254428	0,255102	118	1-butanol	0,59	/	0,575979
40	1-hexene	0,281	0,234127	0,232816	119	2-methyl-1-propanol	0,59	/	0,454152
41	propyl propanoate	0,373	0,397954	0,399309	120	2-methyl-2-propanol	0,613	/	0,552673
42	3-methylbutyl methanoate	0,400	0,397386	0,397785	121	diethyl ether	0,281	/	0,298521
43	hexane	0,300	0,253292	0,251275	122	1-Butanamine	0,338	/	0,417586
44	2-methylpentane	0,278	0,234251	0,232110	123	1-pentyne	0,394	/	0,283189
45	2,2-dimethylbutane	0,233	0,204880	0,204075	124	2-methyl-2-butene	0,339	/	0,354304
46	1-hexanol	0,573	0,492660	0,487620	125	3-methyl-1-butene	0,211	/	0,194654
47	4-methyl-2-pentanol	0,552	0,532720	0,530441	126	2-pentanone	0,346	/	0,317961
48	toluene	0,264	0,258592	0,292779	127	3-pentanone	0,342	/	0,320386
49	4-methylphenol	0,510	0,421292	0,415433	128	ethyl propanoate	0,390	/	0,364372
50	butyl-2-propenoate	0,312	0,369383	0,375315	129	1-pentanol	0,579	/	0,620770
51	methylcyclohexane	0,235	0,268963	0,269655	130	methylcyclopentane	0,227	/	0,225228
52	ethylcyclopentane	0,270	0,262932	0,261971	131	4-methyl-2-pentanone	0,351	/	0,318260
53	propyl butanoate	0,399	0,433776	0,435402	132	ethyl butanoate	0,463	/	0,370718

Tableau 7 Suite et fin

ID	Composé	ω_{EXP}	$\omega_{PRED/LOO}$	$\omega_{CAL/PRED}$	ID	Composé	ω_{EXP}	$\omega_{PRED/LOO}$	$\omega_{CAL/PRED}$
54	heptane	0,350	0,314631	0,314134	133	butyl ethanoate	0,407	/	0,415931
55	2-methylhexane	0,331	0,281163	0,279033	134	2-methylpropyl ethanoate	0,456	/	0,399078
56	Ethanol	0,649	0,617962	0,613334	135	3-methylpentane	0,273	/	0,220348
57	3,3-dimethylpentane	0,269	0,244780	0,245010	136	2,3-dimethylbutane	0,248	/	0,218102
58	1-heptanol	0,588	0,535527	0,526969	137	2-methyl-1-pentanol	0,498	/	0,531977
59	1,3-dimethylbenzene	0,327	0,302680	0,301745	138	3-methylhexane	0,323	/	0,278299
60	cyclooctane	0,254	0,336097	0,342199	139	2,2-dimethylpentane	0,287	/	0,242914
61	1-octene	0,393	0,349480	0,361296	140	2,3-dimethylpentane	0,297	/	0,263594
62	octane	0,399	0,356409	0,355111	141	2,4-dimethylpentane	0,304	/	0,267874
63	2-methylheptane	0,378	0,329063	0,329121	142	2,2,3-trimethylbutane	0,250	/	0,236594
64	4-methylheptane	0,371	0,324383	0,324404	143	1,4-dimethylbenzene	0,322	/	0,300352
65	2,4-dimethylhexane	0,344	0,313326	0,312153	144	3-methylheptane	0,371	/	0,326061
66	3-ethyl-2-methylpentane	0,331	0,311555	0,310834	145	3-ethylhexane	0,362	/	0,326752
67	3-ethyl-3-methylpentane	0,305	0,286987	0,287012	146	2,2-dimethylhexane	0,339	/	0,272865
68	2,2,4-trimethylpentane	0,304	0,285299	0,284899	147	2,3-dimethylhexane	0,347	/	0,299344
69	2,3,3-trimethylpentane	0,291	0,273897	0,274951	148	2,5-dimethylhexane	0,357	/	0,319645
70	2,2,3,3-tetramethylbutane	0,248	0,248781	0,251419	149	3,3-dimethylhexane	0,320	/	0,289248
71	1-octanol	0,594	0,611511	0,613330	150	3,4-dimethylhexane	0,338	/	0,302980
72	propylbenzene	0,345	0,331387	0,332480	151	2,2,3-trimethylpentane	0,298	/	0,278390
73	1-ethyl-4-methylbenzene	0,364	0,349526	0,349473	152	2,3,4-trimethylpentane	0,316	/	0,299985
74	1,2,3-trimethylbenzene	0,367	0,328184	0,327116	153	1-methylethylbenzene	0,326	/	0,315295
75	1,3,5-trimethylbenzene	0,399	0,34495	0,344740	154	1,2,4-trimethylbenzene	0,377	/	0,341373
76	3-methylbutyl butanoate	0,583	0,513224	0,505530	155	cis-bicyclo[4,4,0]decane	0,276	/	0,350663
77	nonane	0,445	0,428462	0,427639	156	2-Butanone	0,322	/	0,270593
78	2,2-dimethylheptane	0,383	0,327256	0,327138	157	propyl acetate	0,389	/	0,368855
79	2,2,3,3-tetramethylpentane	0,304	0,305409	0,308971	158	3-ethylpentane	0,311	/	0,274897

Les statistiques du tableau 8 démontrant la qualité du modèle RNA et l'amélioration qu'il apporte par rapport au modèle de la régression linéaire.

Tableau 8. Statistiques du modèle RNA.

R^2	Q^2_{LOO}	Q^2_{EXT}	*EQMC*	*EQMP*	$EQMP_{EXT}$
93,88	92,99	92,61	0,0418	0,0447	0,0459

Les valeurs expérimentales, calculées et prédites du facteur acentrique ont servi à tracer la droite d'ajustement du modèle RNA (Fig. 8) où l'on remarque une faible dispersion autour de la première bissectrice

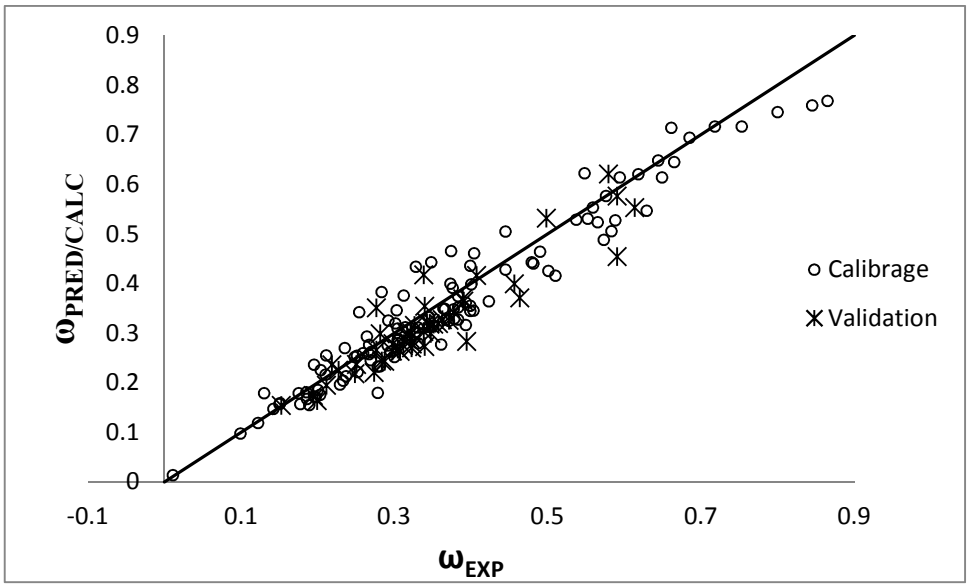


Figure 8: Qualité de l'ajustement du modèle RNA.

II.3 Relation structure/ facteur acentrique d'un ensemble hétérogène d'alcools et de phénols :

L'objectif de cette partie vise à utiliser la méthodologie RSP (pour Relation Structure/ Propriété), dans l'approche algorithme génétique/ régression linéaire multiple (AG/RLM), pour relier les facteurs acentriques, compris entre 0,433 et 0,665, d'un ensemble hétérogène d'alcools et de phénols, à des descripteurs moléculaires reflétant certaines particularités des molécules prises en compte. L'interprétation de ces descripteurs permettrait d'avoir un aperçu sur les facteurs vraisemblablement liés aux facteurs acentriques des alcools et phénols considérés.

II.3.1 Données:

Les facteurs acentriques d'un ensemble hétérogène d'alcools et de phénols ont été prélevés dans la littérature **(**Poling *et al*, 2001). Ces données (tab. 9) se rapportent à 13 alcanols à chaînes ouvertes (linéaires ou ramifiées) ou fermées, et 5 dérivés phénoliques ; on y relève plusieurs isomères (chaîne, position).

Tableau 9. Valeurs des facteurs acentriques et des descripteurs moléculaires sélectionnés.

N°	Nom	ω_i	L2p	HATS3v	h_i	e_{istd}
1	o-Cresol	0,433	1,726	0,118	0,265	0,2882
2	Phenol	0,438	1,481	0,115	0,186	1,3342
3	m-Cresol	0,454	1,516	0,123	0,178	0,5987
4	Pentafluorophenol	0,502	1,539	0,148	0,150	-0,9329
5	p-Cresol	0,505	1,297	0,137	0,101	-0,2194
6	Cyclohexanol	0,528	1,479	0,218	0,230	1,8875
7	Methanol	0,556	0,239	0,092	0,267	1,0597
8	Heptan-1-ol	0,560	0,565	0,107	0,146	-0,5199
9	Hexan-1-ol	0,560	0,541	0,119	0,123	0,2334
10	Butan-2-ol	0,577	0,703	0,177	0,086	1,2230
11	Pentan-1-ol	0,579	0,54	0,134	0,102	-0,0220
12	Octan-1-ol	0,587	0,563	0,100	0,166	-2,4464
13	2-Methylpropan-1-ol	0,592	1,220	0,192	0,112	-1,6401
14	Butan-1-ol	0,593	0,496	0,160	0,105	0,7535
15	2-Methylpropan-2-ol	0,612	1,218	0,237	0,262	-0,6252
16	Propan-1-ol	0,623	0,487	0,189	0,147	0,7080
17	Ethan-1-ol	0,644	0,377	0,203	0,214	1,0411
18	Propan-2-ol	0,665	0,792	0,212	0,161	-2,2746

II.3.2 Sélection du modèle:

La sélection des descripteurs a été réalisée par algorithme génétique dans le logiciel MobyDigs, en maximisant le coefficient de prédiction Q^2_{LOO}. L'optimisation par algorithme génétique (GA-VSS) conduit à de nombreux modèles de différentes dimensions. Parmi les modèles sélectionnés nous avons retenu le plus simple à deux variables explicatives (de coefficient de corrélation r= 0,092 pour une valeur de p= 0,716) qui sont des descripteurs moléculaires géométriques : l'autocorrélation à levier pondéré de distance topologique 3/ pondérée par les volumes atomiques de van der Waals v (HATS3v), et la seconde composante de l'indice de taille WHIM dirigé/ pondérée par les polarisabilités p (L2p).

Les descripteurs moléculaires à invariant holistique pondéré (WHIM) (Todeschini *et al.*, 1994; Todeschini et Gramatica, 1997), permettent de saisir dans le détail les informations relatives à la taille, la forme, la symétrie et la distribution des atomes d'une molécule par rapport à des cadres de références fixes. Le calcul des descripteurs WHIM repose sur l'analyse en composantes principales de la matrice de covariance des coordonnées atomiques pondérées, dont les éléments sont définis par :

$$s_{jk} = \sum_{i=1}^{n} w_i \left(q_{ij} - \overline{q}_j\right)\left(q_{ij} - \overline{q}_k\right) \Big/ \sum_{i=1}^{n} w_i$$

(6)

où n représente le nombre d'atomes de la molécule, w_i, le poid du i^{ème} atome, q_{ij} la j^{ème} coordonnée cartésienne de l'atome i (j=1,2,3) alors que $\overline{q}_j$ est la moyenne de cette j^{ème} coordonnée.

Six modèles de pondération, rapportés à l'échelle de l'atome de carbone, sont proposés, et selon le mode adopté on obtient différentes matrices de covariance et différents axes principaux (c'est-à-dire des composantes $\mathbf{t_m}$, m=1, 2,3) pour la molécule.

On distingue les descripteurs WHIM dirigés, calculés individuellement selon les directions des composantes principales, et les descripteurs WHIM non dirigés, ou globaux, calculés pour la molécule entière à partir des combinaisons des premiers.

Les indices de taille WHIM dirigés, Lkw , sont définis par les valeurs propres $\lambda_k \left(k = 1,2,3\right)$ de la matrice de covariance des coordonnées atomiques pondérées de la molécule. Chaque vecteur propre mesure la dispersion (variance pondérée) des atomes projetés sur l'axe principal considéré, renseignant ainsi sur la dimension de la molécule selon cette direction principale.

Le descripteur moléculaire L2p qui est lié à la dimension des molécules sur le deuxième axe principal, met également en évidence le rôle de la polarisabilité.

Les descripteurs "Assemblage de géométrie, topologie et poids atomiques" GETAWAY (pour GEometry, Topology, and Atom Weights AssempblY) (Consonni *et al.*, 2002a; 2002b) sont basés sur les formules d'autocorrélation spatiales, en pondérant les atomes dans les molécules par des propriétés physico-chimiques, et par les informations 3D contenues dans les éléments des matrices influence moléculaires **H** et influence/distances **R** qui en est déduite (par minimisation des interactions entre paires d'atomes trop éloignés). La matrice **H,** elle-même, est définie à partir de la matrice moléculaire **M** des coordonnées cartésiennes x, y, z des atomes (y compris les hydrogènes) prises par rapport au barycentre de la molécule, considérée dans la conformation choisie. Les éléments diagonaux h_{ii} de **H** (ou leviers) renseignent sur "l'influence" de chaque atome de la molécule quant à déterminer la forme globale de celle-ci; en fait, les atomes périphériques possèdent toujours de plus grands h_{ii} que les atomes voisins du barycentre de la molécule. De plus, l'ampleur du levier maximal d'une molécule dépend de sa grosseur et de sa forme. Notons enfin, ce qui peut être déduit de la géométrie moléculaire, que les valeurs des leviers sont sensibles à des changements conformationnels significatifs, et aux longueurs de liaison qui tiennent compte des types d'atomes et de la multiplicité des liaisons.

Les descripteurs d'autocorrélation à levier pondéré de distance topologique k (=3, dans notre cas), sont calculés à partir de l'équation :

$$HATSkw = \sum_{n=1}^{n_{AT}\text{-}1} \sum_{j>i} \left(w_i h_{ii}\right)\left(w_j h_{jj}\right)\delta\left(k; d_{ij}\right)$$
$$k = 0,1,2,...,8 \qquad (7)$$

n_{AT} est le nombre d'atomes de la molécule; d_{ij} est la distance topologique entre les atomes i et j, c'est-à-dire le nombre de liaisons du chemin le plus court reliant ces deux atomes; w_i est une pondération atomique physico-chimique (volume de van der Waals dans le cas présent); $\delta\left(k; d_{ij}\right)$ est une fonction delta de Dirac ($\delta = 1$ si $d_{ij} = k$, sinon zéro).

Ils apportent une information sur la position effective, dans l'espace moléculaire, des substituants et des fragments de la molécule. De plus, ils renseignent, jusqu'à un certain point, sur la dimension et la forme moléculaire, ainsi que sur les propriétés atomiques spécifiques.

Les diagrammes de probabilités établis à partir des données du tableau 10 montrent que les variables considérées se distribuent selon la loi normale, puisque les R obtenus sont systématiquement supérieurs aux R critiques (R_C) donnés par les tables pour les niveaux $\alpha=1\%$ et $\alpha=5\%$, pour n=18 individus (tab.10).

Tableau 10. Vérification de la loi de Laplace-Gauss pour n=18 individus.

	ω	HATS3v	L2p
R (%)	97,84	97,42	94,99
R_C (%)	94,55 (pour $\alpha = 5\%$) et 92,18 (pour $\alpha = 1\%$)		

Le modèle basé sur les descripteurs sélectionnés a pour équation :

$$\omega_{CAL} = 0{,}510(\pm 0{,}022) + 0{,}938(\pm 0{,}124)\,\text{HATS3v} - 0{,}107(\pm 0{,}011)\,\text{L2p} \qquad (8)$$

Il vérifie les hypothèses d'un modèle statistique linéaire à effets fixes. En effet la figure 9 reproduit la distribution des résidus normalisés RESN (Rapport : résidus ordinaires/ racine du carré moyen des écarts) en fonction des valeurs ajustées AJUST, qui semble aléatoire (sans tendance particulière), ce qui montre la constance des variances σ^2 , c'est-à-dire leur indépendance des régresseurs et de la variable dépendante ajustée.

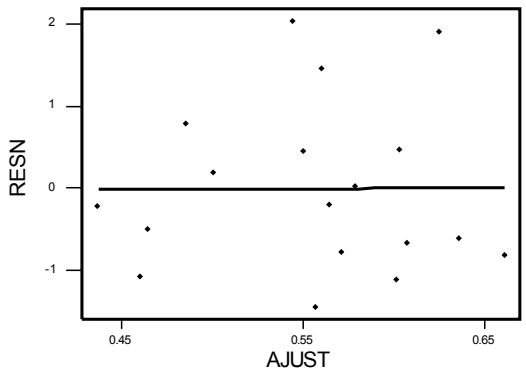


Figure 9: Résidus normalisés en fonction des facteurs acentriques ajustés.

La quasi-linéarité ($R = 0,9675; R_C = 0,9455$) du diagramme des scores normaux (fig. 10) est un indice de normalité. La statistique de Durbin-Watson (Durbin et Watson, 1971), d=1,85, est plus grande que la valeur supérieure donnée par les tables pour 2 régresseurs, et pour tout risque raisonnable α, ce qui établit l'indépendance des résidus.

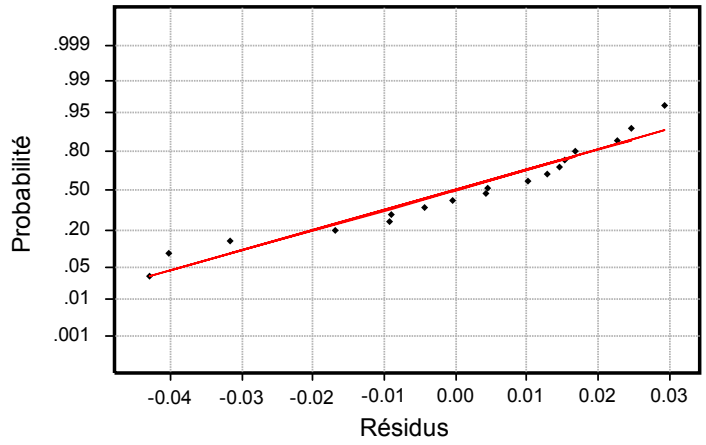


Figure 10 : Diagramme des scores normaux

II.3.3 Qualité du modèle:

Les diagnostiques statistiques du modèle sont rapportées ci-après :

$$R^2(\%) = 89,83\ ;\ Q^2(\%) = 85,35\ ;\ R^2_{adj} = 88,47\ ;\ EQMC = 0,020\ ;\ EQMP = 0,025\ ;$$

$$F = 66,24\ ;\ SE = 0,023$$

Les valeurs de R^2 et de R^2_{adj} montrent la qualité de l'ajustement, alors que la petite différence entre R^2 et Q^2 renseigne sur la robustesse du modèle qui, en outre, est hautement significatif (grande valeur du paramètre de Fisher F). De plus, la similitude de $EQMC$ et $EQMP$ signifie que la capacité de prédiction interne du modèle n'est pas trop dissemblable de son pouvoir d'ajustement. Le modèle permet de reproduire les facteurs acentriques observés avec une précision moyenne inférieure à 3 % ; il possède une robustesse et une capacité prédictive satisfaisantes.

II.3.4 Test de randomisation:

Les modèles RSP, à cause (souvent) de leur complexité et de la sophistication des outils de chimiométrie employés, peuvent constituer une source de corrélation fortuite. Dans le but d'établir que le modèle obtenu n'est pas dû au hasard, nous avons appliqué le Test de randomisation de y . Ce Test consiste à générer un vecteur "facteur acentrique" par permutation aléatoire des composantes du vecteur réel. On calcule alors sur le vecteur obtenu (considéré comme vecteur expérimental réel) un modèle RSP, selon la méthode habituelle. Ce procédé est répété plusieurs fois (100 dans notre cas).

La figure 11 qui représente le graphe des coefficients statistiques Q^2 et R^2 permet de comparer les résultats obtenus pour les modèles randomisés (carrés pleins) au modèle de départ (astérisque). Il est clair que les statistiques obtenues pour les vecteurs modifiés des facteurs acentriques sont plus petites (la majorité des valeurs de Q^2 sont même négatives) que celles du modèle RSP réel, ce qui permet d'assurer qu'une relation structure/facteur acentrique réelle a été établie.

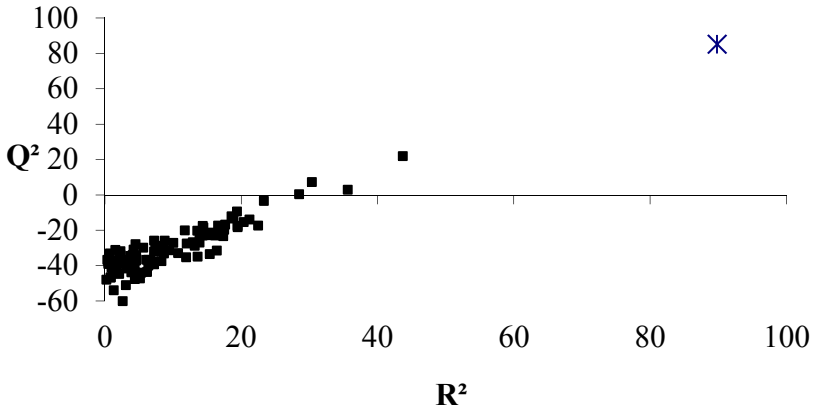


Figure 11 : Test de randomisation associé au modèle RSP

I.3.5 Détection des observations aberrantes:

Pour détecter les observations aberrantes nous avons utilisé les résidus de prédiction standardisés (Draper et Smith, 1998) :

Les valeurs absolues des résidus de prédiction standardisés (tableau 9, dernière colonne) étant toutes inférieures à 3 unités d'écart type $\left(\left|e_{istd}\right| < 3\sigma\right)$ aucune donnée aberrante n'est ainsi détectée pour le modèle.

Les leviers h_i , éléments diagonaux de la matrice **H** de passage du vecteur y au vecteur $\hat{y}$, permettent de juger de l'influence d'une observation i dans la détermination de l'équation de régression (lorsque h_i est supérieur à la valeur critique $3(2+1)/n$). Toutes les valeurs reproduites dans la colonne h_i (Tab. 9) étant inférieures à $3\times 3/18 = 0,5$, aucune observation n'est influente.

II.3.6 Comparaison avec d'autres modèles de la littérature:

Du fait de la différence entre les bases de données modélisées (du point de vue source et nombre de données) d'une part et des complexités des méthodes utilisées d'autre part, la comparaison des erreurs de calcul a été privilégiée. Les erreurs de calcul du facteur acentrique par les méthodes de contribution de groupes (MCG) **(**Poling *et al*, 2001) se distribuent entre 0,04 et 0,07 en unité log. Elles sont supérieures à celles des travaux cités dans le tableau 11 où AAD est la moyenne des valeurs absolues des déviations, et AAD% la valeur relative.

Tableau 11. Comparaison avec les travaux antérieurs.

	n	Méthode	AAD (AAD%)	SE	EQMC
(Carande *et al.*, 2015)	614	QSPR (RVS)[b]	0,0310 (6,9)	0,023	0,048
(Wang *et al.*, 2012)	477(48)[a]	MCG	0,0613 (10,39)	-	-
(Mokshina *et al.*, 2014)	331	QSPR (RF)[c]	0,0140 (-)	0,027	-
Notre travail	18	SPR (AG/RLM)	0,0172 (3,05)	0,023	0,020

[a] 48 alcools pour lesquels les résultats sont rapportés [b] Régression par vecteurs supports

[c] Random Forest pour la sélection des descripteurs.

Si le modèle de Mokshina *et al.* est le meilleur, il ne reste pas moins difficile à mettre en œuvre, la méthode de calcul des descripteurs n'étant pas automatisée contrairement au modèle bilinéaire que nous présentons dont les variables explicatives sont calculables rapidement par les logiciels disponibles.

Les **196** données précédemment séparés en **139** composés de calibrage et **59** de validation ont servis dans cette section à trouver un modèle QSPR pour le volume critique.

III.1 Taille du modèle:

L'ensemble de calibrage a été utilisé pour déterminer en premier la taille du modèle et la sélection des descripteurs les plus pertinents. En procédant sur cet ensemble par la méthode (GA-VSS) en maximisant le coefficient de prédiction interne Q^2_{Loo} pour des dimensions allant de 1 à 10 et pour chaque taille le modèle plus performant à été sélectionné et ses statistique (R^2 et Q^2_{Loo}) ont servi à tracer le graphique de leurs variations; réunies dans le tableau 12; en fonction de la dimension (fig. 12).

Tableau 12. Valeur de R^2 et Q^2 pour chaque taille du modèle (cas de V_C).

p	1	2	3	4	5	6	7	8	9	10
R^2	80,37	93,30	96,88	98,19	98,49	98,84	99,00	99,13	99,20	99,31
Q^2	79,27	92,52	96,23	97,91	98,34	98,65	98,77	98,92	99,05	99,19

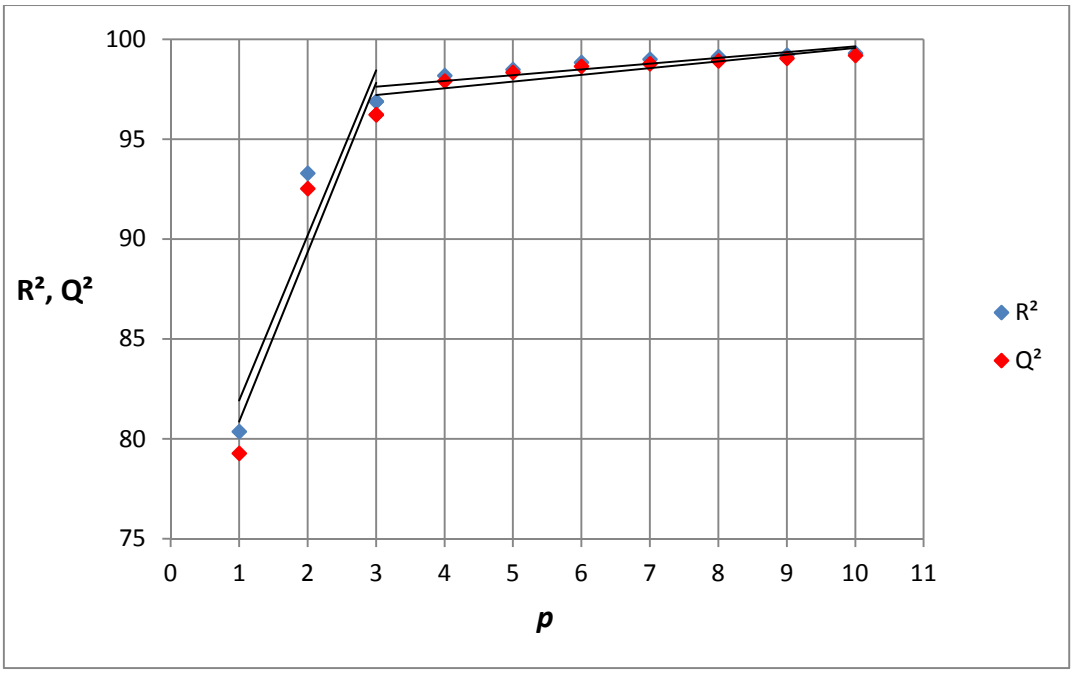


Figure 12: Variation de Q^2 et R^2 en fonction de la taille du modèle (cas de V_C)

A partir de cette figure nous remarquons que les performances d'ajustement (R^2) et prédiction interne (Q^2) sont très similaires. Le nombre de descripteurs à inclure dans le modèle est 3 car

à partir de cette taille, l'ajout d'un nouveau descripteur (augmentation de la taille) n'apporte pas d'amélioration significative. Une amélioration est dite significative (Xu *et al.*, 2011) si l'augmentation de la taille du modèle s'accompagne d'une amélioration du coefficient de détermination (R^2) d'au moins 2 (ou 0,02 si R^2 n'est pas exprimé en pourcentage) ce qui est le cas ici.

III.2 Choix du modèle:

En utilisant le logiciel Minitab pour l'analyse de régression nous obtenant l'équation du modèle basé sur les 3 descripteurs choisis par algorithmes génériques et les diagnostiques suivants:

$$V_C = 44{,}2 + 1{,}26\,Mol.\,Wt. - 243\,Mor18v - 405\,Mor17p \quad (9)$$

Régresseurs	Coéf.	ES Coéf.	t	P	FIV
Constante	44,168	6,497	6,80	0,000	
Mol. Wt.	1,26175	0,05218	24,18	0,000	1,202
Mor18v	-242,78	14,80	-16,40	0,000	1,035
Mor17p	-405,072	9,599	-42,20	0,000	1,165

Là aussi les valeurs des FIV sont toutes, comme exigé, inférieures à 5 ce qui exclu la muti-colinéarité des descripteurs sélectionnés qui sont aussi significatif comme le prouvent les test de Student reportés dans l'analyse. En plus de l'inexistence de muti-colinéarité il est aussi utile de vérifier la corrélation des descripteurs avec le volume critique et s'assurer que les descripteurs ne sont pas corrélés deux à deux entre eux. Ceci est confirmé par la matrice de corrélation ci-dessous:

	Vc	Mol. Wt.	Mor18v
Mol. Wt.	0,711		
	0,000		
Mor18v	-0,336	-0,176	
	0,000	0,040	
Mor17p	-0,852	-0,373	0,012
	0,000	0,000	0,887

III.3 Qualité statistique

Les statistiques du modèle calculées par le logiciel MobyDigs pour le modèle basé sur les trois descripteurs et reliés entre eux par RLM sont reportées dans le tableau 13.

Tableau 13. Statistiques relatives au modèle QSPR du volume critique.

R^2	Q^2	Q^2_{BOOT}	R^2_{adj}	$Q^2_{L10\%O}$	$Q^2_{L20\%O}$	$EQMC$	$EQMP$	F	S
96,88	96,23	94,92	96,81	96,24	95,59	35,084	38,557	1376	35,608

Selon ces statistiques, nous pouvons dire que notre modèle:

- A un bon pouvoir d'ajustement confirmé par la valeur de R^2;
- a une différence entre R^2 et R^2_{adj} est très faible ce qui prouve que les trois descripteurs sont nécessaire et qu'il n'y a pas de sur-paramétrisation;
- est validé en interne car Q^2 est presque identique à R^2 ;
- est stable car toutes les valeurs de Q^2_{BOOT} , $Q^2_{L10\%O}$ et $Q^2_{L20\%O}$ sont proches entre elles et proches de R^2 et Q^2;
- est significatif (Valeur de F très élevée);
- a des $EQMC$, $EQMP$ et S de proches valeurs.

Les statistiques relatives à l'ensemble de validation externe sont: $Q^2_{EXT} = 94,30\%$ proche de R^2 et Q^2. La valeur de $EQMP_{EXT} = 47,448$ est supérieure à $EQMC$ et $EQMP$.

En prédiction, les performances ne sont pas aussi bonnes qu'en ajustement mais la capacité de notre modèle à prédire des nouveaux composés reste acceptable. Cette anomalie doit être expliquée.

Le diagramme de Williams (fig. 13) est la définition du domaine d'application du modèle QSPR. Les observations atypiques révélées par le diagramme sont dans le tableau 14.

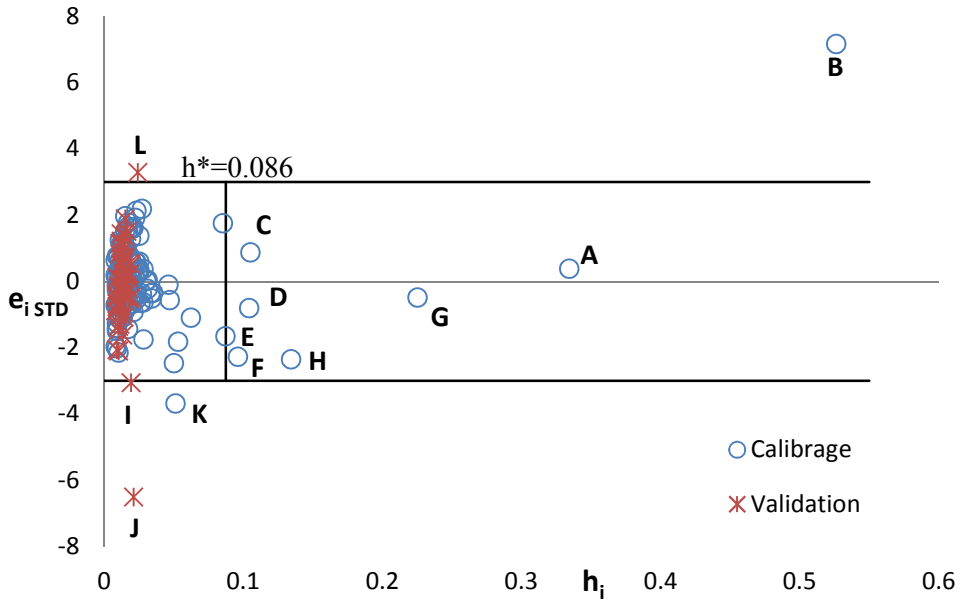


Figure 13: **Diagramme de Williams** ***(cas de V_C)***

Tableau 14. Observations aberrantes signalées.

Observations aberrantes	
En X	En Y
A octadecafluorooctane	K pentadecane
B docosafluorodecane	
C phenanthrene	I 2,4-dimethylphenol*
D anthracene	J 3,5-dimethylphenol*
E hexadecane	L cis-bicyclo[4.4.0]decane*
F heptadecane	
G 1,4-diphenylbenzene	
H octadecane	

** Composés de validation*

Nous remarquons que les composés C, D, E, F, et H sont structurellement influents car caractérisés par le plus grand nombre de carbones d'une part et pour les uns (C, D, G) ayants 3 cycles benzéniques d'autre part.

Le docosafluorodecane (B) compte 22 atomes de fluore qui remplace les 22 hydrogène de l'alcane normale $C_{10}H_{22}$. Il est aberrant en X avec un bras de levier $h_i = 0,526$ et en Y où

$e_{i\,STD} = 7,1694$ ce qui est anormal car généralement les composés influents dans un modèle QSPR sont bien prédits. Le octadecafluorooctane (A) est aussi un composé influent mais bien prédit, il a aussi 18 atomes de fluore au lieu de 18 hydrogènes (similaire au composé B). Nous avons donc les deux composés les plus influents de l'ensemble de calibrage avec une particularité (tous les hydrogènes remplacés par des fluores) qu'ils sont les seuls à avoir.
Les composés aberrants en Y (I, J) de l'ensemble de validation sont prédits avec des résidus élevés ($e_i = -107,56$ **et** $-228,70$ respectivement) et sont les deux seuls phénols bisubstitués.
Il est donc nécessaire de remanier la base de données en enlevants ces composés atypiques (A, B, I et J) et reconstruire le modèle sur 135 composés et le valider en utilisant les 57 restants.
L'équation du modèle basé sur les 3 descripteurs devient:

$$V_C = 14,9 + 1,88\,Mol.\,Wt. - 136\,Mor18v - 349\,Mor17p \qquad (10)$$

Régresseurs	Coéf.	ES Coéf.	t	P	FIV
Constante	14,867	7,692	1,93	0,055	
Mol. Wt.	1,8830	0,1173	16,06	0,000	3,612
Mor18v	-163,20	19,03	-8,58	0,000	2,017
Mor17p	-349,39	12,99	-26,89	0,000	2,630

Par rapport au premier modèle les FIV sont plus élevés mais acceptables.

Le tableau 15 comporte les valeurs expérimentales $V_{C\,EXP}$, calculées $V_{C\,CAL}$ et prédites $V_{C\,PRED}$ du **volume critique** ainsi que les valeurs des leviers h_i et des résidus standardisés de prédictions $e_{i\,STD}$ des deux sous-ensembles.

Tableau 15. Valeurs de $V_{C\,EXP}$, $V_{C\,CAL}$, $V_{C\,PRED}$, h_i et $e_{i\,STD}$ des 192 composés.

N°	Composé	$V_{C\,EXP}$	$V_{C\,CAL}$	$V_{C\,PRED}$	h_i	$e_{i\,STD}$
1	Chlorotrifluoromethane	180,3	231,8574	234,3270	0,046	1,7375
2	Trichloromonofluoromethane	248	281,5919	284,9141	0,090	1,2157
3	Carbon Tetrachloride	276	255,4751	251,5595	0,160	-0,8379
4	Tetrafluoromethane	140,7	181,5915	183,0471	0,034	1,3539
5	Trichloromethane	240	220,2741	218,6460	0,076	-0,6980
6	Fluoroform	133	151,6267	152,0580	0,023	0,6056
7	Difluoromethane	121	126,2265	126,3330	0,020	0,1692
8	Chloromethane	143	136,0065	135,8692	0,019	-0,2262
9	Methyl fluoride	113,3	104,4838	104,2370	0,027	-0,2887
10	nitromethane	173	148,8766	148,3269	0,022	-0,7839

Tableau 15 suite

N°	Composé	$V_{C\,EXP}$	$V_{C\,CAL}$	$V_{C\,PRED}$	h_i	$e_{i\,STD}$
11	Methane	98,6	83,3982	82,6407	0,047	-0,5137
12	methanol	118	105,3941	105,0154	0,029	-0,4140
13	Methanethiol	145	122,4910	122,021	0,020	-0,7294
14	Pentafluoroethyl chloride	256	309,3644	319,4874	0,159	2,1755
15	tetrafluoroethene	172	190,2350	191,2283	0,052	0,6203
16	Acetylene	112,2	77,0512	75,9057	0,032	-1,1587
17	1,1,1-tri?uoroethane	193,6	203,2185	203,4439	0,023	0,3129
18	Ethylene	131,1	92,4186	91,0379	0,034	-1,2809
19	1,1-dichloroethane	236	247,5488	247,8614	0,026	0,3777
20	1,2-dichloroethane	220	265,7814	266,8987	0,024	1,4913
21	Methyl formate	172	152,0336	151,6517	0,019	-0,6454
22	Ethyl Chloride	199	198,9423	198,9415	0,014	-0,0019
23	Ethane	145,5	139,4194	139,18	0,038	-0,2024
24	Ethanethiol	207	184,8219	184,4853	0,015	-0,7127
25	Dimethyl sulfide	201	189,0422	188,8634	0,015	-0,3841
26	Propyne	163,5	128,8614	128,0172	0,024	-1,1283
27	Cyclopropane	162,8	217,1498	219,0583	0,034	1,7982
28	1,2-dichloropropane	287,66	319,2246	320,0636	0,026	1,0315
29	methyl ethanoate	228	205,0234	204,6451	0,016	-0,7398
30	Ethyl formate	229	208,2116	207,8573	0,017	-0,6699
31	n-Propyl chloride	254	235,7932	235,5665	0,012	-0,5827
32	Propane	200	189,8964	189,5992	0,029	-0,3315
33	1-Propanol	219	211,4487	211,3296	0,016	-0,2429
34	N,N-dimethylmethanamine	254	220,5445	219,9602	0,017	-1,0787
35	Furan	218	250,4983	250,9811	0,015	1,0438
36	Thiophene	219	229,9394	230,1033	0,015	0,3514
37	1-Butyne	208	179,9866	179,4854	0,018	-0,9038
38	1,3-Butadiene	221	160,0256	158,7171	0,021	-1,9776
39	Cyclobutane	218	256,2946	257,4052	0,028	1,2558
40	2-methylpropene	238,8	201,7902	200,9343	0,023	-1,2033
41	1,4-dioxane	238	247,6783	247,8301	0,015	0,3112
42	Ethyl Acetate	286	272,1608	272,0032	0,011	-0,4422
43	propyl methanoate	285	266,9820	266,7198	0,014	-0,5785
44	2-chlorobutane	312	308,3713	308,3342	0,010	-0,1158
45	Butane	255	246,7272	246,5472	0,021	-0,2684
46	2-methyl-2-propanol	275	261,1740	261,0112	0,012	-0,4421
47	diethyl ether	280	266,2333	266,0754	0,011	-0,4400
48	1,2-dimethoxyethane	270,64	289,0067	289,2041	0,011	0,5863
49	1,2-Butanediol	303,05	286,6300	286,4455	0,011	-0,5246
50	Diethyl sulfide	318	323,9011	323,9546	0,009	0,1879
51	pyridine	254	243,2517	243,1066	0,013	-0,3445
52	Cyclopentene	245	241,2165	241,1584	0,015	-0,1216
53	cyclopentane	260	251,3066	251,1916	0,013	-0,2786
54	1-pentene	298,4	247,4175	246,6413	0,015	-1,6384
55	2-methyl-2-butene	292	267,8246	267,3801	0,018	-0,7806
56	3-methyl-2-butanone	310	302,8232	302,7526	0,010	-0,2288

Tableau 15 suite

N°	Composé	$V_{C\,EXP}$	$V_{C\,CAL}$	$V_{C\,PRED}$	h_i	$e_{i\,STD}$
57	2-methyltetrahydrofuran	267	289,5579	289,8906	0,015	0,7244
58	pentanoic acid	336,2	313,9444	313,6121	0,015	-0,7149
59	ethyl propanoate	345	308,9863	308,4625	0,014	-1,1562
60	pentane	311	302,3737	302,2317	0,016	-0,2777
61	neopentane	303,2	294,0734	293,9365	0,015	-0,2932
62	1-pentanol	326	321,9465	321,9077	0,009	-0,1292
63	benzene	256	262,8416	263,0160	0,025	0,2232
64	phenol	229	280,8334	281,8526	0,019	1,6767
65	cyclohexene	296,88	289,8690	289,7579	0,016	-0,2255
66	4-methyl-3-penten-2-one	353,43	308,0800	307,6172	0,010	-1,4466
67	cyclohexane	308	324,2595	324,4620	0,012	0,5204
68	1-hexene	355,1	305,3167	304,8335	0,010	-1,5869
69	ethyl butanoate	421	367,0350	366,2238	0,015	-1,7338
70	3-methylbutyl methanoate	411,4	357,0889	356,1628	0,017	-1,7501
71	hexane	368	377,4688	377,6249	0,016	0,3049
72	2-methylpentane	366,7	389,9962	390,4399	0,019	0,7529
73	2,2-dimethylbutane	359,1	363,7645	363,8314	0,014	0,1497
74	1-hexanol	381	387,5667	387,6230	0,008	0,2090
75	4-methylphenol	277	331,2437	332,3480	0,020	1,7565
76	butyl-2-propenoate	427,54	376,9180	375,4801	0,028	-1,6586
77	cycloheptane	359	411,7831	412,718	0,017	1,7025
78	methylcyclohexane	368	384,2574	384,4118	0,009	0,5180
79	cis-1,3-dimethylcyclopentane	363,3	342,4823	342,2974	0,009	-0,6628
80	heptanoic acid	429,7	425,7754	425,6960	0,020	-0,1271
81	2-methylhexane	421	442,6155	442,9789	0,017	0,6963
82	3-ethylpentane	415,8	441,4547	441,8777	0,016	0,8260
83	2,2,3-trimethylbutane	397,6	426,6746	427,1267	0,015	0,9348
84	1-heptanol	435	435,7777	435,7858	0,010	0,0248
85	1,4-dimethylbenzene	378	391,7494	392,0797	0,023	0,4476
86	cyclooctane	410	468,4594	469,4562	0,017	1,8838
87	1-octene	468	437,8122	437,5320	0,009	-0,9616
88	octane	492	487,2127	487,1410	0,015	-0,1538
89	2-methylheptane	488,2	488,9251	488,9360	0,015	0,0233
90	4-methylheptane	476	503,4845	503,9741	0,018	0,8866
91	2,4-dimethylhexane	472	512,4837	513,3224	0,020	1,3116
92	3-ethyl-2-methylpentane	445,3	480,2847	480,7522	0,013	1,1212
93	3-ethyl-3-methylpentane	455,1	509,3253	510,3897	0,019	1,7540
94	2,2,4-trimethylpentane	469,7	447,226	446,9746	0,011	-0,7179
95	2,3,3-trimethylpentane	455,1	492,2788	492,8587	0,015	1,1955
96	2,2,3,3-tetramethylbutane	482	466,2126	466,0240	0,012	-0,5049
97	1-octanol	490	512,8082	513,1199	0,013	0,7313
98	propylbenzene	440	438,7096	438,6808	0,022	-0,0419
99	1-ethyl-4-methylbenzene	440	447,3110	447,5228	0,028	0,2397
100	1,2,3-trimethylbenzene	435	407,4693	406,3977	0,037	-0,9159
101	1,2,4-trimethylbenzene	435	424,4238	424,0333	0,036	-0,3508
102	1,3,5-trimethylbenzene	430	471,0189	472,1925	0,028	1,3444

Tableau 15 suite

N°	Composé	$V_{C\,EXP}$	$V_{C\,CAL}$	$V_{C\,PRED}$	h_i	$e_{i\,STD}$
103	1-nonene	526	499,858	499,5732	0,011	-0,8348
104	nonane	555	557,4072	557,4533	0,019	0,0778
105	2,2,5-trimethylhexane	519	568,914	569,9997	0,021	1,6196
106	2,2,3,3-tetramethylpentane	478	541,995	543,0027	0,016	2,0582
107	2,2,3,4-tetramethylpentane	490	526,9781	527,5981	0,016	1,1911
108	2,2,4,4-tetramethylpentane	504	524,2105	524,7882	0,028	0,6624
109	2,3,3,4-tetramethylpentane	493	539,9721	540,8955	0,019	1,5194
110	1-nonanol	544	574,2517	575,3984	0,037	1,0050
111	naphthalene	407	398,7234	398,3165	0,047	-0,2794
112	1,2,3,4-tetrahydronaphthalene	408	474,941	476,9469	0,029	2,1983
113	butylbenzene	497	476,6752	476,2726	0,019	-0,6576
114	1,4-diethylbenzene	480,5	485,3249	485,4513	0,026	0,1576
115	1-(1-methylethyl)-4-methylbenzene	497	507,4307	507,7347	0,028	0,3421
116	trans-bicyclo[4,4,0]decane	480	513,7723	514,1662	0,012	1,0796
117	1-decene	584	547,5952	547,1181	0,013	-1,1663
118	decane	624	607,5141	607,1675	0,021	-0,5344
119	1-decanol	600	633,0199	633,9074	0,026	1,0795
120	undecane	689	670,3759	669,8338	0,028	-0,6108
121	1,1'-biphenyl	497	536,6937	540,6951	0,092	1,4403
122	1,3,5-triethylbenzene	624,14	621,7989	621,7252	0,031	-0,0771
123	1,3-dimethyltricyclo[3,3,1,13,7]decane	571,45	641,4576	643,0276	0,022	2,2738
124	dodecane	754	741,5938	741,1370	0,036	-0,4115
125	1-dodecanol	718	730,1789	730,7080	0,042	0,4078
126	diphenylmethane	563	533,1555	531,1589	0,063	-1,0333
127	tridecane	823	807,9242	807,1765	0,047	-0,5093
128	phenanthrene	554	577,4655	580,2336	0,106	0,8714
129	nthracene	554	534,7895	532,5401	0,105	-0,7126
130	etradecane	894	846,4208	843,5536	0,057	-1,6319
131	pentadecane	966	864,2231	857,5733	0,061	-3,5159
132	hexadecane	1034	995,6112	991,7675	0,091	-1,3916
133	heptadecane	1103	1047,911	1041,7141	0,101	-2,0308
134	1,4-diphenylbenzene	729	720,1616	717,5815	0,226	-0,4077
135	octadecane	1189	1140,503	1132,3397	0,144	-1,9241
136*	1,1-di?uoroethene	153,5	/	142,8163	0,020	-0,3391
137	1,1-difluoroethane	181	/	180,6824	0,016	-0,0101
138	ethanoic acid	171	/	155,2747	0,018	-0,4984
139	Ethanol	167	/	161,9946	0,021	-0,1590
140	Dimethyl ether	170	/	161,2291	0,022	-0,2786
141	Dimethylamine	180	/	157,8642	0,023	-0,7036
142	Propene	184,6	/	147,2791	0,028	-1,1890
143	Acetone	209	/	189,5946	0,016	-0,6146
144	Isopropyl Alcohol	220	/	209,0236	0,015	-0,3475
145	methyl ethyl ether	221	/	214,0116	0,016	-0,2213
146	2-Propanamine	221	/	207,6378	0,016	-0,4232
147	1-Butene	240,8	/	193,4715	0,020	-1,5022
148	trans-2-butene	237,7	/	206,0954	0,021	-1,0036

Tableau 15 suite et fin

N°	Composé	$V_{C\,EXP}$	$V_{C\,CAL}$	$V_{C\,PRED}$	h_i	$e_{i\,STD}$
149	cis-2-butene	233,8	/	214,7152	0,023	-0,6066
150	2-Butanone	267	/	240,0946	0,012	-0,8504
151	Tetrahydrofuran	224	/	233,0792	0,013	0,2872
152	Butanoic acid	292	/	256,0176	0,014	-1,1383
153	Isobutane	262,7	/	239,436	0,020	-0,7383
154	1-butanol	275	/	263,2773	0,011	-0,3704
155	2-methyl-1-propanol	273	/	259,0800	0,011	-0,4398
156	cyclopentanone	268	/	242,0445	0,018	-0,8229
157	cis-2-pentene	302,1	/	262,6734	0,016	-1,2486
158	3-methyl-1-butene	304,9	/	236,5681	0,016	-2,1645
159	2-pentanone	301	/	305,1034	0,010	0,1296
160	3-pentanone	336	/	290,2475	0,010	-1,4444
161	propyl acetate	345	/	307,9335	0,016	-1,1741
162	2-methylbutane	308,3	/	313,9472	0,018	0,1790
163	2-pentanol	329	/	330,5226	0,009	0,0481
164	ethyl propyl ether	339	/	314,4048	0,009	-0,7761
165	methylcyclopentane	319	/	292,5846	0,010	-0,8342
166	cyclohexanol	333,88	/	379,1388	0,015	1,4329
167	4-methyl-2-pentanone	340,6	/	344,8160	0,012	0,1332
168	hexanoic acid	377,2	/	367,2165	0,017	-0,3163
169	butyl ethanoate	412,8	/	359,0610	0,020	-1,7054
170	2-methylpropyl ethanoate	413	/	358,9415	0,022	-1,7171
171	3-methylpentane	366,7	/	342,3714	0,011	-0,7686
172	2,3-dimethylbutane	357,6	/	385,3140	0,018	0,8784
173	toluene	316	/	322,0327	0,023	0,1917
174	ethylcyclopentane	375	/	358,8461	0,012	-0,5105
175	trans-1,3-dimethylcyclopentane	363,3	/	337,1956	0,009	-0,8238
176	3-methylhexane	404	/	434,2371	0,015	0,9571
177	2,2-dimethylpentane	415,8	/	437,9838	0,015	0,7024
178	2,3-dimethylpentane	393	/	434,1888	0,015	1,3037
179	2,4-dimethylpentane	417,5	/	396,0825	0,009	-0,6760
180	3,3-dimethylpentane	414,1	/	432,3844	0,015	0,5789
181	1,3-dimethylbenzene	375	/	380,9667	0,024	0,1897
182	3-methylheptane	471,1	/	485,5255	0,014	0,4564
183	3-ethylhexane	460,5	/	506,3463	0,018	1,4537
184	2,2-dimethylhexane	478	/	487,3254	0,014	0,2951
185	2,3-dimethylhexane	468,2	/	489,4377	0,015	0,6722
186	2,5-dimethylhexane	482	/	499,9424	0,017	0,5685
187	3,3-dimethylhexane	442,8	/	501,5974	0,017	1,8632
188	3,4-dimethylhexane	458,8	/	499,3149	0,017	1,2835
189	2,2,3-trimethylpentane	436	/	474,9979	0,012	1,2328
190	2,3,4-trimethylpentane	456,2	/	476,7334	0,013	0,6494
191	1-methylethylbenzene	434,7	/	435,3536	0,020	0,0207
192	cis-bicyclo[4,4,0]decane	480	/	597,3680	0,024	3,7327

***** Ensemble de validation de 136 à 192.

III.4 Qualité du nouveau modèle:

Le modèle obtenu sur la base remanié a les statistiques réunis dans le tableau 16.

Tableau 16. Statistiques du nouveau modèle QSPR du volume critique.

n_{TR}	R^2	Q^2	Q^2_{BOOT}	R^2_{adj}	$Q^2_{L10\%O}$	$Q^2_{L20\%O}$	$EQMC$	$EQMP$	F	S
135	97,51	97,31	97,18	97,45	97,47	97,28	31,355	32,578	1707,708	31,8302
				n_{EXT}	Q^2_{EXT}	$EQMP_{EXT}$				
				57	97,35	32,336				

Une nette amélioration des statistiques de l'ensemble de calibrage est observée en comparant les tableaux 13 et 16. Pour l'ensemble de validation aussi les paramètres sont meilleurs, le Q^2_{EXT} passant même de 94,3 à 97,35% et le $EQMP_{EXT}$ a baissé de près de 15. Ceci prouve l'influence négatives des deux composés A et B sur l'ajustement du modèle et l'incapacité de ce dernier à prédire les composés éliminés (I et J) pour les raisons précédemment évoquées.

Les métriques de Golbraikh et Tropsha concernant seulement l'ensemble de validation sont une autre preuve de la validité du modèle QSPR pour le volume critique :

1) $R^2_{CV\,EXT} = 0,9125 > 0,5$
2) $R^2 = 0,9407 > 0,6$
3) $\left(R^2 - R_0^2\right)/R^2 = -0,0620 < 0,1$ et $0,85 < k = 0,9895 < 1,15$

   ou

   $\left(R^2 - R_0'^2\right)/R^2 = -0,0630 < 0,1$ et $0,85 < k' = 1,0016 < 1,15$
4) $\left|R_0^2 - R_0'^2\right| = 9,3596.10^{-4} < 0,3$

III.5 Qualité de l'ajustement:

Le graphe des valeurs calculées et prédites du volume critique en fonction des celles expérimentales est présenté dans la figure 14 et fait ressortir une faible dispersion autour de la première bissectrice caractéristique du bon ajustement du modèle pour les deux sous-ensembles.

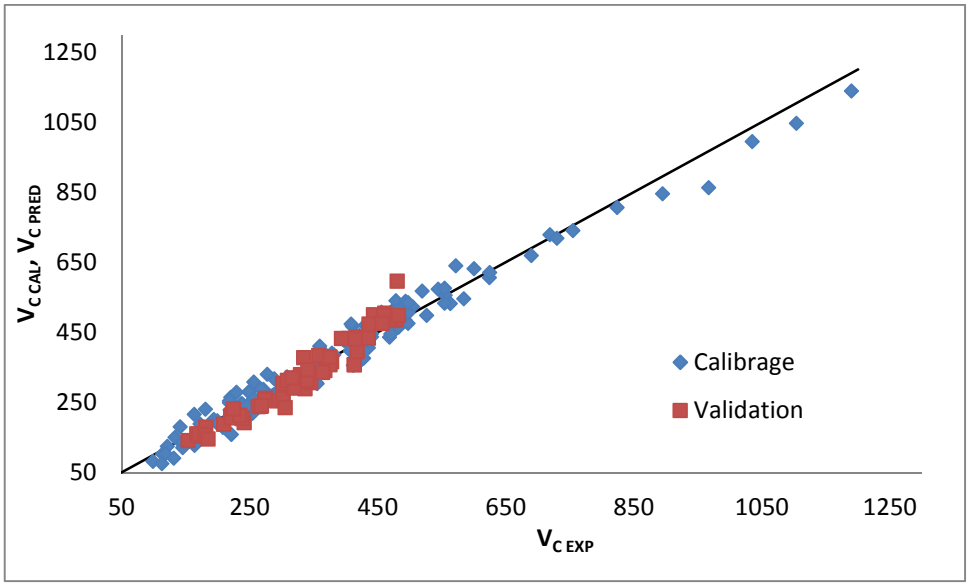


Figure 14: **Qualité de l'ajustement** ***(cas de V_C)***

III.6 Domaine d'application:

La figure 15 définit le domaine d'application du modèle par l'approche des leviers et où les composés aberrants en Y sont le pentadecane et le cis-bicyclo[4,4,0]decane déjà signalés dans la figure 13. Les composés aberrants en X ou structurellement influent (dont certains ont été remarqués dans le premier modèle) sont réunis dans le tableau 17 avec une description qui peut en être la cause.

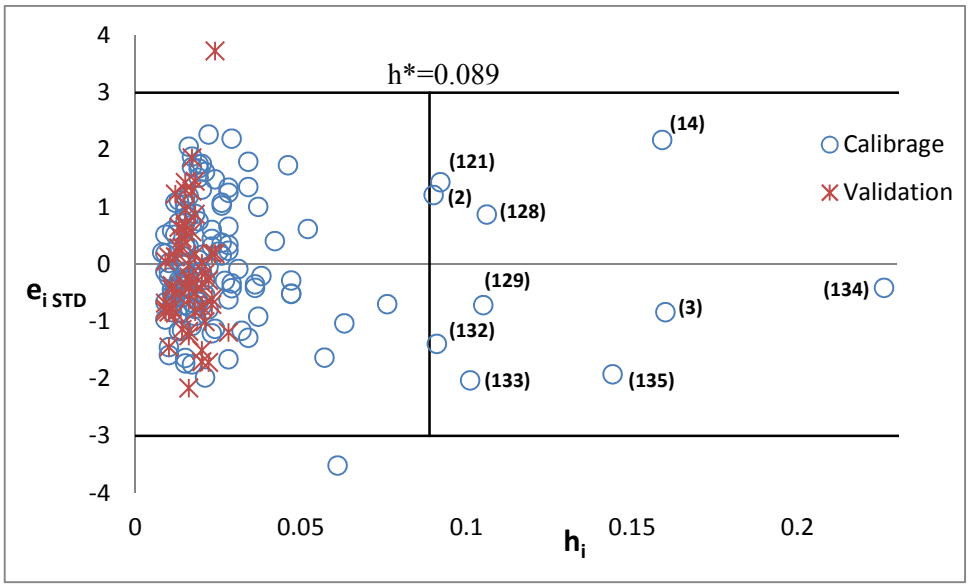


Figure 15: Diagramme de Williams du nouveau modèle

Tableau 17. Observations influentes du nouveau modèle.

Composé	Description
(2) Trichloromonofluoromethane	Présence du Fluore
(14) Pentafluoroethyl chloride	
(3) Carbon Tetrachloride	Un seul atome de carbone
(121) 1,1'-biphenyl	Plus d'un cycle benzénique
(128) phenanthrene	
(129) anthracene	
(134) 1,4-diphenylbenzene	
(132) hexadecane	16, 17 et 18 atomes de carbones
(133) heptadecane	
(135) octadecane	

III.7 Test de randomisation:

La même technique utilisée pour le facteur acentrique a été utilisée pour le volume critique. Là aussi les vecteurs réponses randomisés n'ont pas de statistiques d'un modèle QSAR acceptables (représentés par des astérisques rouge figure 16) ce qui prouve que le modèle original n'est pas le fruit d'une corrélation chasseuse.

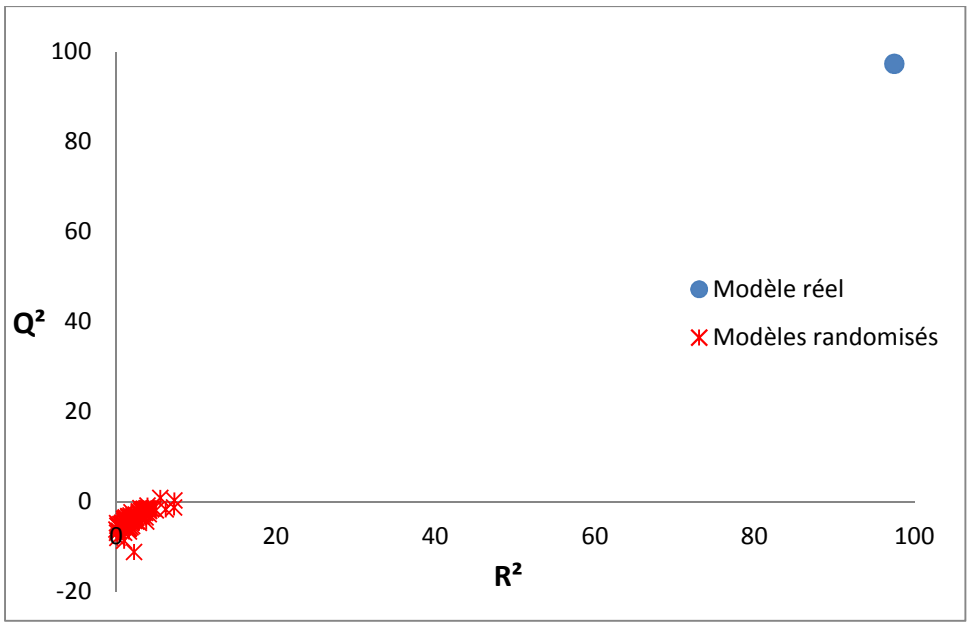


Figure 16: **Test de randomisation** ***(cas de V_C).***

III.8 Contributions relatives des descripteurs:

Les contributions relatives des descripteurs du modèle ont été déterminées et sont représenté dans la figure 17. Le descripteur le plus important est le Mor17p (3) avec une contribution de 43,45 % suivis de la masse moléculaire (Mol. Wt. (1)) avec 32,04 % suivi par le troisième descripteur Mor18v (2) (24,51 %).

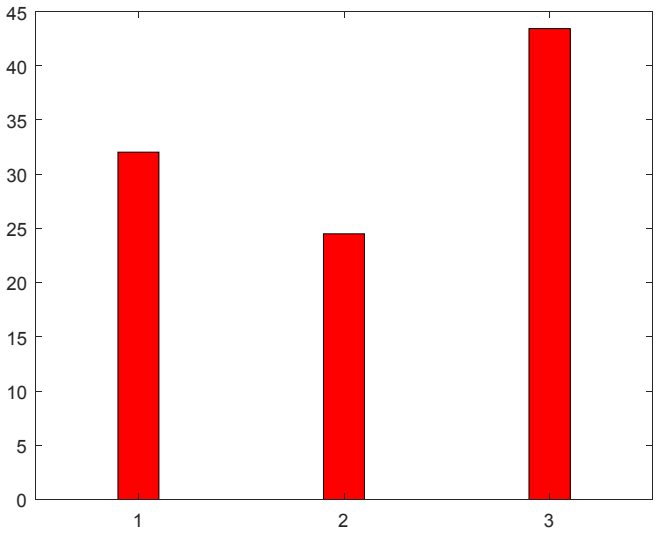


Figure 17: **Contributions relatives des descripteurs *(cas de V_C).***

III.10 Définition et interprétation des descripteurs:

Il est nécessaire d'essayer de trouver une interprétation phénoménologique de chaque descripteur c.-à-d. sa relation avec la propriété modélisé et cette relation peut être déduite après la définition du dit descripteur. Les trois descripteurs du modèle QSPR du volume critique et leurs classes et types sont dans le tableau suivant:

Tableau 18. Types et classes des descripteurs (cas de V_c).

Descripteur	Type	Classe
Mol. Wt.: masse moléculaire	2D	Descripteurs constitutionnels
Mor18v: signal 3D-MoRSE - 18/ pondéré par les volumes atomiques de van der Waals.	3D	Descripteur 3D-MoRSE
Mor17p: signal 3D-MoRSE - 17/ pondéré par les volumes atomiques de van der Waals.		

Les descripteurs 3D-MoRSE (pour: *3D-Molecule Representation of Structures based on Electron diffraction*; représentation 3D des structures moléculaires basée sur la diffraction d'électron) sont basés sur l'idée d'obtenir des informations à partir des coordonnées atomiques 3D par la transformation utilisée dans les études de diffraction d'électrons pour la préparation de courbes de diffusion théoriques (Schuur *et al.*, 1996). Ils sont calculés par la formule:

$$Morsw = \sum_{i=1}^{nAT-1} \sum_{j=i+1}^{nAT} w_i w_j \frac{\sin\left(s \cdot r_{ij}\right)}{s \cdot r_{ij}} \qquad (11)$$

où $Morsw$ est l'intensité des électrons diffusés, w est une propriété atomique, r_{ij} sont les distances interatomiques et nAT est le nombre d'atomes. Le terme s représente la diffusion dans diverses directions par une collection d'atomes nAT.

32 descripteurs 3D-MoRSE sont calculés pour cinq propriétés atomiques w : le cas non pondéré (u), la masse atomique (m), le volume de van der Waals (v), l'électronégativité atomique de Sanderson (e) et la polarisabilité atomique (p).

On remarque que lorsque la masse moléculaire augmente le volume critique augmente. Cette augmentation est accentuée par les valeurs de Mor18v et Mor17p (valeurs négatives) car multipliées par des coefficients négatifs (éq. 7 et 8).

III.11 Comparaison avec la méthode de contribution de groupe:

Deux méthodes de contribution de groupes ont été utilisées pour la prédiction du volume critique des **196** composés dont on a valeurs expérimentales. Il s'agit des méthodes de Joback et Reid (**J&R87**) et Lydersen (**Ly55**) (Voir: Sec II.2 et II.3 de la PARTIE A) pour lesquelles nous avons utilisé le logiciel Cranium Reader 2.0 (Cranium, 2011). Les valeurs prédites de V_C par ces deux méthodes sont en annexe.

Il est à signaler que le V_C du méthane n'a pas pu être prédit par les deux méthodes car il n'y as pas de fractionnement possible prévu. La méthode de Lydersen ne comporte pas de groupe N cyclique d'où le V_C prédit manquant de la pyridine.

Le tableau 19 résume les statistiques de notre modèle QSPR et celles des MCG utilisées.

Tableau 19. Comparatif des statistiques des MCG et du modèle QSPR.

	n	AAD	AAD %	SD	EQMC
QSPR	192	25,4137	8,1283	31,8302	31,355
J&R87	191	13,4542	3,4085	27,1676	26,8817
Ly55	190	15,2718	3,9109	30,3572	30,0359

Du point de vue statistique **J&R87** est, parmi les 3 méthodes de prédiction du volume critique, la meilleure. Mais notre modèle QSPR arrive à prédire tout les composés de la base modélisée et il en sera de même pour n'importe quel composé dont on voudra prédire le volume critique à l'avenir avec les restrictions suivantes:

- Phénols bisubstitués
- Fluorure d'alcanes où tous les hydrogènes sont remplacés par des atomes de fluore.

CONCLUSION GENERALE

L'objectif de cette thèse fut d'élaborer des modèles QSPR du facteur acentrique et du volume critique des composés chimiques. Quelques 265 substances dont les valeurs expérimentales pour V_C et ω ont été prélevées dans la littérature et utilisées pour paramétrer les modèles QSPR.

Après l'optimisation des géométries des molécules; les descripteurs moléculaires ont été générés et parmi eux les sous-ensembles les plus pertinents de variables explicatives sélectionnés par la méthode GA-VSS et reliés aux variables dépendantes (V_C et ω) par régression linéaire multiple pour les deux propriétés et pour la dernière, le sous-ensembles de descripteurs a été aussi la base d'une relation non linéaire par réseaux de neurones artificiels. Ces calculs ont eu pour outils des logiciels spécialisés disponibles dans notre laboratoire: *HyperChem, DRAGON, MobyDigs, Minitab et Molegro Data Modeller*. Enfin, *Cranium Reader* a été utilisé pour avoirs les prédictions par les méthodes de contribution de groupe en vue d'une comparaison.

Pour le facteur acentrique le modèle QSPR linéaire nécessitant cinq descripteurs est d'assez bonne qualité ($R^2, Q^2_{LOO}, Q^2_{EXT}, Q^2_{BOOT}$ *et* Q^2_{LMO} > ou au moins très proche de 90%) mais a été amélioré par RNA. De plus une relation structure/ facteur acentrique d'un ensemble d'alcools et de phénols a été établie en utilisant deux descripteurs et ses résultats comparés aux travaux similaires et récents. Cette comparaison est à l'avantage de notre modèle RSP car, en plus d'être simple, le calcul de ses descripteurs (au nombre de 2) par les logiciels disponibles est automatisé et rapide.

Concernant le volume critique trois descripteurs ont été suffisants au calibrage du modèle en usant de 137 points de données. Le modèle QSPR est stable robuste et arrive à prédire les 59 composés de validation sans trop d'erreurs sauf pour deux composés. En comparant les performances de notre modèle à deux méthodes de contributions de groupe qui sont celle de Lydersen et de Joback & Reid on remarque que la dernière méthode est meilleure mais sans pouvoir prédire un des composés de la base de données ce qui est un problème inhérent à toutes les MCG même les plus récentes. Malheureusement, ces dernières ne font pas toujours l'objet d'une automatisation sous forme de logiciel et donc inaccessibles aux chercheurs ou aux chimistes. Nous pouvons donc dire, qu'avec seulement deux descripteurs à calculer, vu que la masse moléculaire est disponible pour n'importe quelle

substance, par le biais de plateformes rassemblant des dizaines de milliers de composés dont les géométries ont été optimisées et des versions en ligne des logiciels de calcul de descripteurs, notre modèle est aisément applicable afin de prédire les volumes critiques de composés dont les valeur sont manquantes.

Les anomalies constatées pour notre modèle du volume critique nous ont conduit à restreindre son application ce qui constitue un frein à sa généralisation.

Ce problème pourrait être résolu en adoptant d'autres méthodes de régressions qui ne peuvent être que non linéaires comme les machines à vecteurs supports (SVM) ou les réseaux de neurones artificiels à fonction de base radiale (RBF-ANN). Un couplage des algorithmes génétiques avec les SVM (GA/SVM) ou les RNA (GA/RNA) serait une alternative intéressante au GA-VSS basée sur la régression linéaire multiple pour le choix d'un sous-ensemble pertinent de descripteur à inclure dans le modèle.

Enfin, avoir accès aux bases de données thermodynamiques et en extraire les propriétés critiques et les valeurs du facteur acentrique pour une gamme plus large de composés pallierait au problème remarqué de la singularité de certains types de molécules.

REFERENCES BIBLIOGRAPHIQUES

Abildskov J (1994). Development of a new group contribution method. Danish Technical University.

Alliot J M, Schiex T, Brisset P et Garcia F, (2002). intelligence artificielle et informatique théorique, 2e édition, Éditions Cepadues.

Ambrose D et Ghiassee N B (1987). Vapour pressures and critical temperatures and critical pressures of some alkanoic acids: C_1 to C_{10}. J Chem Thermodynamics, 19 (5): 505 -519.

Ambrose D et Patel N C (1984). The correlation and estimation of vapour pressures IV. Extrapolation of vapour pressures and estimation of critical pressures by the principle of corresponding states using two reference fluids with non-spherical molecules. J. Chem. Thermodynamics, 16 (5)**:** 459 - 468.

Ambrose D et Young C L (1995). Vapor-liquid critical properties of elements and compounds. 1. An introductory survey. J Chem Eng Data, 40 (2): 345 - 357.

Andrews T (1869). On the continuity of the gaseous and liquid states of matter. (cité par Ambrose et Young (1995)). Proc Roy Soc (London), 18: 42 - 45.

Anselme M J et Teja A S (1990). The critical properties of rapidly reacting substances. AIChE Symp Ser, 279**:** 115, 128-132.

Antoine C (1888). Vapor pressure: a new relationship between pressure and temperature. C R Hebd Seances Acad Sci,107: 681-684.

Arakawa M, Hasegawa K et Funatsu K, (2006). QSAR study of anti-HIV HEPT analogues based on multi-objective genetic programming and counter-propagation neural network. Chem Intel Labo Sys, 83 (2): 91–98.

Atkins P W (1983). Molecular Quantum Mechanics, Oxford University Press, Oxford.

Blake J (1886). On the connection between chemical constitution and physiological action. Nature, 34: 594-595.

Born M R et Oppenheimer R (1927). Zur Ouantentheorie der Molekeln. Ann Phys, 389 (20): 457-484.

BOUAKKADIA Amel. (2016) Modélisation de quelques propriétés (cteh, s, pv, koc(w)) contrôlant l'évolution dans l'environnement d'une série d'herbicides. Thèse de doctorat de l'Université Badji Mokhtar-Annaba.

Bouveresse D J R, Maalouly J et Jaillais B (2004). Sélection d'échantillons représentatifs par des méthodes chimiométriques : Application à des modèles d'étalonnage. Spectra analyse, 33 (237): 23-27.

Bruhat G (1933). Thermodynamique à l'usage de l'enseignement supérieur scientifique et technique, 2ème édition. Masson. Paris.

Burns J A et Whitesides G M (1993). Feed-forward neural networks in chemistry: mathematical systems for classification and pattern recognition. Chem Rev, 93 (8): 2583–2601.

Carande W H, Kazakov A, Muzny C et Frenkel M (2015). Quantitative Structure-Property Relationship Predictions of Critical Properties and Acentric Factors for Pure Compounds. J Chem Eng Data, 60 (5): 1377–1387.

Chouquet C, (2010) "Modèles Linéaires", Laboratoire de Statistique et Probabilités-Université Paul Sabatier-Toulouse.

Clark R D (1997). OptiSim: An extended dissimilarity selection method for finding diverse representative subsets. J Chem Inf Comput Sci, 37 (6): 1181-1188.

Claudel B (1996). Propriétés thermodynamiques des fluides. Techniques de l'ingénieur, Thermodynamique et énergétique. url = "http://www.techniques-ingenieur.fr/base-documentaire/energies-th4/thermodynamique-et-energetique-42216210/proprietes-thermodynamiques-des-fluides-b8020/" (accédé le 06/07 2015)

Consonni V, Todeschini R et Pavan M (2002a). Structure/response correlations and similarity/diversity analysis by GETAWAY descriptors. 1. Theory of the novel 3D molecular descriptors. J Chem Inf Comput Sci, 42 (3): 682–692.

Consonni V, Todeschini R, Pavan M et Gramatica P, (2002b). Structure/response correlations and similarity/diversity analysis by GETAWAY descriptors. 2. Application of the novel 3D molecular descriptors to QSAR/QSPR studies. J Chem Inf Comput Sci, 42 (3): 693–705.

Constantinou L (1993). Estimation of properties of acyclic organic compounds through conjugation. Ph.D. Thesis, University of Maryland, College Park, MD.

Constantinou L et Gani R (1994). New group contribution method for estimating properties of pure compounds. AIChE J, 40 (10): 1697–1710.

Constantinou L, Gani R et O'Connell J P (1995). Estimation of the acentric factor and the liquid molar volume at 298 K using a new group contribution method. Fluid Phase Equil, 103 (1): 11-22.

Constantinou L, Prickett S E et Mavrovouniotis M L (1993). Estimation of thermodynamical and physical properties of acyclic hydrocarbons using the abc approach and conjugation operators. Ind Eng Chem Res, 32 (8): 1734 - 1746.

Constantinou L, Prickett S E et Mavrovouniotis M L (1994). Estimation of properties of acyclic organic compounds using conjugation operators. Ind Eng Chem Res, 33 (2), 395-402.

Cranium Reader, (2011) Molecular Knowledge Systems,Inc

Dantas Filho H A, Galvao R K H, Araujo M C U; da Silva E C, Saldanha T C B, José G E, Pasquini C, Raimundo I M et Rohwedder J J R **(**2004). A strategy for selecting calibration samples for multivariate modelling. Chemom Int Lab Syst, 72 (1): 83-91.

Dastmalchi S, Hamzeh-Mivehroud M et Asadpour-Zeynali K (2012**).** Comparison of Different 2D and 3D-QSAR Methods on Activity Prediction of Histamine H3 Receptor Antagonists. Iran J Pharm Res, 11 (1): 97–108.

de Groot P J, Postma G J, Melssen W J et Buydens L M C (1999). Selecting a representative training set for the classification of demolition waste using remote NIR sensing. Anal Chim Acta, 392 (1): 67 - 75.

De La Tour C (1822). Cité par Ambrose et Young (1995). Ann Chim Physik, 21(2): 127-132.

De La Tour C (1822). Cité par Ambrose et Young (1995). Ann Chim Physik, 22(2): 410-415.

Dewar M J S et Thiel W (1977). Ground states of molecules. 38. The MNDO method. Approximations and parameters. J Am Chem Soc JACS, 99(15): 4899-4907.

Dewar M J S et Thiel W (1977). Ground states of molecules. 39. MNDO results for molecules containing hydrogen, carbon, nitrogen, and oxygen. J Am Chem Soc JACS, 99 (15): 4907–4917.

Dewar M J S, Zoebisch E G, Healy E F et Stewart J J P (1985). Development and use of quantum mechanical molecular models. 76. AM1: a new general purpose quantum mechanical molecular model. J Am Chem Soc JACS, 107 (13): 3902 – 3909.

Douali L, Schmitzer A R, Villemin D, Jarid A et Cherqaoui D (2007). Neural networks and their applications in chemistry and biology. Phys Chem News, 34: 131 – 144.

Draper N R et Smith H (1998). Applied Regression Analysis. Third Edition, Wiley Series in Probability and Statistics, New York.

DURAND Alexandra. (2007) Méthodes de Sélection de Variables Appliquées en Spectroscopie Proche Infrarouge pour l'Analyse et la Classification de Textiles. Thèse de Doctorat. Université des Sciences et Technologies de Lille.

Durbin J et Watson G S (1971). Testing for serial correlation in least squares regression III. Biometrika, 58 (1): 1-19.

Dymond J H et Smith E B (1980). The virial coefficients of pure gases and mixtures: A Critical Compilation. Clarendon Press, Wotton-under-Edge Gloucestershire.

ERRAHOUI Khadidja. (2015) Etude des relations quantitatives structure–toxicité des composés chimiques à l'aide des descripteurs moléculaires. Thèse de doctorat de l'Université Abou Bekr Belkaïd de Tlemcen.

Estrada E (1996). Spectral moments of the edge adjacency matrix in molecular graphs. 1. definition and applications to the prediction of physical properties of alkanes. J Chem Inf Comput Sci, 36 (4): 844-849.

Estrada E (1998). spectral moments of the edge adjacency matrix in molecular graphs. 3. molecules containing cycles. J Chem Inf Comput Sci, 38 (1): 23-27.

FAYET Guillaume. (2010) Développement de modèles QSPR pour la prédiction des propriétés d'explosibilité des composes nitroaromatiques, Thèse de doctorat de l'université Pierre et Marie Curie-Paris.

Fernandez M, Caballero J et Tundidor-Camba A (2006). Linear and nonlinear QSAR study of N-hydroxy-2-[(phenylsulfonyl)amino] acetamide derivatives as matrix metalloproteinase inhibitors. Bioorg Med Chem, 14 (12): 4137 – 4150.

Fock V (1930). Approximation method for solving the quantum mechanical multibody problem. Z Angew Phys, 61: 126 - 148.

FORTUNE Antoine. (2006). Techniques de modélisation moléculaire appliquées a l'etude et a l'optimisation de molécules immunogènes et de modulateurs de la chimiorésistance. Thèse de Doctorat. Université Joseph Fourier - Grenoble I.

Foussard J N, Julien E et Mathé S (2010). Thermodynamique: bases et applications. Dunod. Paris.

Free S M et Wilson J W (1964). A mathematical contribution to structure-activity studies. J Med Chem, 7 (4): 395-399.

Gasteiger J et Zupan J (1993). Neural networks in chemistry. Angewandte Chemie International Edition in English, 32 (4): 503-527.

Gharagheizi F, Eslamimanesh A, Mohammadi A H et Richon D (2011). Determination of critical properties and acentric factors of pure compounds using the artificial neural network group contribution algorithm. J Chem Eng Data, 56 (5): 2460–2476.

Gharagheizi F, Eslamimanesh A, Sattari M, Mohammadi A et Richon D (2015). Computation of the second virial coefficient of chemical compounds using a corresponding states based method. In: Advances in Chemistry Research. Taylor J C (Eds), Nova Science Publishers Inc, 24: 91-112.

Gharagheizi F. et Mehrpooya M. (2008) Prediction of some important physical properties of sulfur compounds using quantitative structure–properties relationships. Mol Divers, 12 (3-4): 143–155.

Globisch C, Pajeva I K et Wiese M (2006). Structure-activity relationships of a series of tariquidar analogs as multidrug resistance modulators. Bioorg Med Chem, 14 (5): 1588 - 1598.

Golbraikh A et Tropsha A (2002a). Predictive QSAR modeling based on diversity sampling of experimental datasets for the training and test set selection. J Comput Aided Mol Des, 16 (5-6): 357 - 369.

Golbraikh A et Tropsha A (2002b). Beware of q^2!. J Mol Graph Model, 20 (4): 269 - 276,

Golbraikh A, Shen M, Xiao Z, Xiao Y D, Lee K H et Tropsha A (2003). Rational selection of training and test sets for the development of validated QSAR models. J Comput Aided Mol Des, 17 (2-4): 241 – 253.

Goldberg D (1989). Genetic Algorithms in Search, Optimization and machine learning. Addison-Wesley Longman Publishing Co., Inc. Boston.

GUENDOUZI Abdelkrim (2015). Élaboration des modèles QSPR prédictifs des propriétés physico-chimiques à l'aide des descripteurs moléculaires. Thèse de doctorat de l'Université Abou Bekr Belkaïd de Tlemcen

Guha R et Willighagen E (2012). A survey of quantitative descriptions of molecular structure. Curr Top Med Chem, 12 (18):1946 – 1956

Guldberg C M (1890). Über die gesetze der molekularvolumina und der siedepunkte. ZPhysik Chem, 5 (1): 374 - 382.

Hansch C, Leo A et Hoekman D (1995). Exploring QSAR vol 2: hydrophobic, electronic and steric constants. ACS, Washington DC.

Hansch C, Maloney P P, Fujita T et Muir R M (1962). Correlation of biological activity of phenoxyacetic acids with hammett substituent constants and partition coefficients. Nature, 194 (4824):178-180

Hatree D R (1928). The wave mechanics of an atom with a non-Coulomb central field. Part I. Theory and methods. Math Proc Camb Philos Soc, 24 (1): 89-110.

Hopfield J J (1982). Neural networks and physical systems with emergent collective computational abilities. Proc Natl Acad Sci USA, 79 (8): 2554 - 2588.

Hougen O A, Watson K M et Ragatz R A (1962). Chemical process principles, Part II: Thermodynamics. John Wiley & Sons. New Jersey.

James A M et Lord M P (1992). Macmillan's chemical and physical data. Macmillan. London.

Jensen F (2007). Introduction to computational chemistry, John Wiley & Sons Ltd, Chichester.

Joback K G et Reid R C (1987). Estimation of pure-component properties from group-contributions. Chem Eng Comm, 57 (1-6): 233–243.

Kennard R W et Stone L A (1969). Computer aided design of experiments. Technometrics, 11 (1): 137 - 148.

Kobe K A et Lynn R E (1953). The critical properties of elements and compounds. Chem Rev, 52 (1): 117-236.

Kohonen T (1995). Self-Organizing Maps. In Springer Series in Information Sciences (Vol. 30). Berlin, Heidelberg, Springer. New York.

Liu L et Chen S (1996). Correlation of the acentric factor for hydrocarbons. Ind Eng Chem Res, 35 (7): 2484-2486.

Livingstone D J (2000). The characterization of chemical structures using molecular properties. A survey. J Chem Inf Comput Sci, 40 (5):195–209

Lydersen A L (1955). Estimation of critical properties of organic compounds. University of Wisconsin College Engineering, Eng. Exp. Stn. Rep. 3, Madison, Wisconsin.

Manallack D T, Ellis D D et Livingstone D J **(**1994**).** Analysis of linear and nonlinear QSAR data using neural networks**.** J Med Chem, 37 (22): 3758–3767.

Markovic S et Gutman I (1999). spectral moments of the edge adjacency matrix in molecular graphs. benzenoid hydrocarbons. J Chem Inf Comput Sci, 39 (2): 289-293.

Marrero J et Gani R (2001). Group-contribution based estimation of pure component properties. Fluid Phase Equilib, 183–184:183–208.

Marrero-Morejon J et Pardillo-Fontdevilla E (1999). Estimation of pure compound properties using group-interaction contributions. AIChE J, 45(3): 615–621.

Mason E A et Spurling T H (1968). The Virial Equation of State. Pergamon Press. Oxford.

McCulloh W S et Pitts W (1943). A logical calculus of the ideas immanent in neural nets. Bull Math Biophys, 5: 133-137.

McGregor M J, Flores T P et Sternberg M J (1989). Prediction of β-turns in proteins using neural networks. Protein Eng Sel, 2 (7): 521 – 526.

McWeeny R et Sutcliffe B T (1969). Methods of molecular quantum mechanics. Academic Press, London.

MINITAB, Release 13,31, Statistical software, (2000).Moran P A P (1950). Notes on continuous stochastic phenomena, Biometrika, 37 (1/2): 17-23.

Mokshina E G, Kuz'min V E et Nedostup V I (2014). QSPR modeling of critical parameters of organic compounds belonging to different classes in terms of the simplex representation of molecular structure. Russ J Organ Chem, 50 (3): 314–321.

Molegro Data modeller 2.1.0, (2009) Molegro ApS.

NANNOOLAL Y (2006). Development and critical evaluation of group contribution methods for the estimation of critical properties, liquid vapour pressure and liquid viscosity of organic compounds. Ph.D Thesis, University of Kwazulu-Natal.

Nannoolal Y, Rarey J et Ramjugernath D (2007). Estimation of pure component properties. Part 2: Estimation of critical property data by group contribution. Fluid Phase Equilib, 252 (1-2): 1 - 27.

Nelson L C et Obert E F (1954). Generalized P-V-T properties of gases. Trans ASME, 76: 1057 - 1066.

Pardillo-Fontdevila E et González-Rubio R (1997). A group-interaction contribution approach. A New strategy for the estimation of physical-chemical properties of branched isomers. Chem Eng Comm, 163: 245.

Peng D Y et Robinson D B (1976). A new two-constant equation of state. Ind Eng Chem Fundam, 15 (1): 59 - 64

Peterson S D, Schaal W et Karlén A (2006). Improved CoMFA Modeling by optimization of settings. J Chem Inf Model, 46 (1): 355–364.

Pitzer K S (1955). The volumetric and thermodynamic properties of fluids. I. Theoretical basis and virial coefficients1. J Am Chem Soc JACS, 77 (13): 3427–3433.

Pitzer K S, Lippmann D Z, Curl Jr R F, Huggins C M et Petersen D E *(*1955). The volumetric and thermodynamic properties of fluids. II. Compressibility factor, vapor pressure and entropy of vaporization. J Am Chem Soc JACS, 77 (13): 3433–3440.

Poling B. E., Prausnitz J. M., O'Connell J. P. (2001). The properties of gases and liquids. Fifth Edition, McGRAW-HILL, New York.

Przezdziecki J W et Sridhar T (1985). Prediction of liquid viscosities. AIChE J, 31 (2): 333 - 335

Qiang W, Qingzhu J et Peisheng M (2012). Prediction of the acentric factor of organic compounds with the positional distributive contribution method. J Chem Eng Data, 57 (1): 169–189.

Redlich O et Kwong J N S (1949). On the thermodynamics of solutions. Chem Rev, 44**:** 233 - 244.

Reid R C, Prausnitz J M, Sherwood T K et Poling B E (1987). The properties of gases and liquids. Fourth Edition, McGRAW-HILL, New York.

Rekker R (1977). The hydrophobic fragmental constant. Elsevier, Amsterdam.

Riedel L (1949). Estimation of unknown critical pressures of organic compounds. Z Elektrochem, 53: 222 – 228.

Riedel L (1954). Kritischer Koeffizient, Dichte des gesättigten Dampfes und Verdampfungswärme. Untersuchungen über eine Erweiterung des Theorems der übereinstimmenden Zustände. Teil III. Chem Ing Tech, 26 (12): 679 - 683

Rivail J L (1994). Eléments de chimie quantique à l'usage des chimistes, Inter Editions, Paris.

Roy K et Narayan Das N (2014). A review on principles, theory and practices of 2D-QSAR. Current Drug Metabol, 15 (4): 346 – 379

Roy K, Kar S et Narayan Das N (2015). A primer on QSAR/QSPR modeling: fundamental concepts. Springer International Publishing AG, Switzerland.

Rumelbart D E, McClelland J L et PDP Research Group (1988). Parallel distributed processing (Vol. 1). IEEE.

SCAN- Software for Chemometric Analysis- 1995, version 1,1- for Windows, Minitab USA.

Schrödinger E (1926). Quantization as an eigenvalue problem, Ann Phys Leipzig, 348 (4): 361 - .

Schuur J H, Selzer P et Gasteiger J (1996). The coding of the three-dimensional structure of molecules by molecular transforms and its application to structure-spectra correlations and studies of biological activity. J Chem Inf Comput Sci, 36 (2): 334 – 344.

Shen M, Béguin C, Golbraikh A, Stables J P, Kohn H et Tropsha A (2004). Application of predictive QSAR models to database mining: identification and experimental validation of novel anticonvulsant compounds. J Med Chem, 47 (9): 2356 – 2364.

Skander N et Chitour C E (2007). Group-contribution estimation of the critical properties of hydrocarbons. Oil Gas Sci Technol, 62 (3): 391 - 398.

Slater J C (1928). The self consistent field and the structure of atoms, Phys Rev, 32 (3): 339 -

Slater J C (1929). The theory of complex spectra, Phys Rev 34 (10): 1293 -

Smith J M, Van Ness H et Abbott M (2005). Introduction to chemical engineering thermodynamics. McGraw-Hill Education. New York

Snee R D (1977). Validation of regression models: Methods and examples. Technometrics, 19 (4): 415-428.

Soave G (1972). Equilibrium constants from a modified Redlich–Kwong equation of state. Chem Eng Sci, 27 (6): 1197 - 1203.

Sola D, Ferri A, Banchero M, Manna L et Sicardi S (2008). QSPR prediction of N-boiling point and critical properties of organic compounds and comparison with a group-contribution method. Fluid Phase Equilib, 263 (1): 33–42.

Stewart J J P (1989). Optimization of parameters for semiempirical methods II. Applications. J Comput Chem, 10 (2): 209–220.

Stewart J J P (2007). Optimization of parameters for semiempirical methods V: Modification of NDDO approximations and application to 70 elements. J Mol Model, 13 (12): 1173–1213.

Szabo N et Osllund S (1982). Modern quantum chemistry, Macmillan, New York.

Taylor P J (1991). Quantitative drug design. the rational design, mechanistic study and therapeutic applications of chemical compounds. In: Hansch C, Sammes P G, Taylor J B (eds) Comprehensive medicinal chemistry, vol 4. Pergamon Press, Oxford; pp 241–294

Thermodynamics Research Center (TRC), Texas A&M University, http://trcweb.tamu.edu, 1999.

Todeschini R et Consonni V (2000) Handbook of molecular descriptors. Wiley-VCH, Weinheim.

Todeschini R et Gramatica P (1997). SD-modelling and prediction by WHIM descriptors. Part 5. Theory development and chemical meaning of WHIM descriptors. Quant Struct-Act Rel, 16 (2): 113 - 119.

Todeschini R, Ballabio D, Consonni V, Mauri A, Pavan M, (2009) MobyDigs Software for Multilinear Regression Analysis and Variable Subset Selection by Genetic Algorithm, Release 1,1 for windows, Milano.

Todeschini R, Consonni V et Gramatica P (2009). Chemometrics in QSAR. In: Brown S, Tauler R, Walczak R (eds) Comprehensive chemometrics, vol 4. Elsevier, Oxford, pp 129–172

Todeschini R, Lasagni M et Marengo E (1994). New molecular descriptors for 2D and 3D structures. Theory. J Chemom, 8 (4): 263–272.

Tomassone R, Lesquoy E et Miller C (1983). La régression: nouveaux regards sur une ancienne méthode statistique. Masson. Paris.

Tropsha A, Gramatica P et Gombar V K (2003). The importance of being earnest: validation is the absolute essential for successful application and interpretation of QSPR models. QSAR Comb Sci, 22 (1): 69 - 77.

Tsonopoulos C (1974). An empirical correlation of second virial coefficients. AIChE J, 20 (2): 263- 272

Turner B E, Costello C L et. Jurs P C (1998). Prediction of critical temperatures and pressures of industrially important organic compounds from molecular structure. J Chem Inf Comput. Sci, 38 (4): 639–645.

Tute M S (1990). History and objectives of quantitative drug design. In: Hansch C, Sammes PG, Taylor JB (eds) Comprehensive medicinal chemistry, vol 4. Pergamon Press, Oxford, pp 1–31

Voityuk A A et Rösch N (2000). AM1/d Parameters for Molybdenum. J Phys Chem A, 104 (17): 4089–4094.

Wagner W (1973). New vapour pressure measurements for argon and nitrogen and a new method for establishing rational vapour pressure equations. Cyrogenics, 13 (8): 470 - 482.

Wagner W (1977). A new correlation method for thermodynamic data applied to the vapor pressure curve of argon, nitrogen, and water. Report PC/TlS, IUPAC Thermodynamic Table Project Centre. London

Wang Q, Jia Q et Ma P (2012). Prediction of the acentric factor of organic compounds with the positional distributive contribution method. J. Chem. Eng. Data, 57 (1): 169–189.

Wehrens R, Putter H et Putter L M (2000). The bootstrap: a tutorial. Chemom Int, Lab Syst, 54 (1): 35- 52.

Weisberg S (2005). Applied Linear Regression. Third edition, John Wiley and sons Inc, New Jersey.

Wold S, Eriksson L (1995). Statistical validation of QSAR results, In: H, Van de Waterbeemd ed, Chemometrics methods in molecular design, VCH, New York, Vol, 2, pp, 309- 318.

Wu W, Walczak B, Massart D L, Heuerding S, Erni F, Last I R et Prebble K A (1996). Artificial neural networks in classification of NIR spectral data: design of the training set. Chemom Intell Lab Syst, 33 (1): 35-46.

Xu J, Wang L, Liu L, Bai Z et Wang L (2011). QSPR study of absorption maxima of azobenzenes dyes. Bull Korean Chem Soc, 32 (11): 3865-3872.

Yonaba H, Anctil F et Fortin V (2010). Comparing sigmoid transfer functions for neural network multistep ahead streamflow forecasting. J Hydrol Eng, 15 (4): 275 – 283.

ANNEXES

Tableau A-1: Valeurs des facteurs acentriques (ω_{EXP}) et des descripteurs du modèle.

N°	Composé	ω_{EXP}	Mol. Wt.	MATS2m	ESpm05u	R1u+	R1p+
1	Chlorotrifluoromethane	0,175	104,459	-0,199	5,485	0,02	0,018
2	Trichloromonofluoromethane	0,195	137,368	-0,02	5,485	0,041	0,04
3	Tetrafluoromethane	0,177	88,005	0,25	5,485	0	0
4	Fluoroform	0,267	70,014	0,333	3,434	0,051	0,019
5	Difluoromethane	0,278	52,024	0,5	0	0,046	0,018
6	Chloromethane	0,151	50,488	0	0	0,055	0,043
7	Methyl fluoride	0,204	34,033	0	0	0,033	0,013
8	Methane	0,011	16,043	0	0	0	0
9	methanol	0,565	32,042	0	0	0,427	0,074
10	Methanethiol	0,15	48,109	0	0	0,369	0,231
11	Pentafluoroethyl chloride	0,251	154,467	-0,211	6,494	0,125	0,124
12	Acetylene	0,189	26,038	0	0	0,149	0,057
13	1,1,1-tri?uoroethane	0,259	84,041	-0,167	5,485	0,152	0,058
14	ethanoic acid	0,445	60,053	-0,333	3,434	0,323	0,079
15	Ethane	0,099	30,07	0	0	0,157	0,06
16	1,2-Propadiene	0,122	40,065	1	0	0,239	0,091
17	Propene	0,142	42,081	1	0	0,203	0,077
18	Cyclopropane	0,13	42,081	0	3,434	0,19	0,072
19	Propylamine	0,283	59,111	-0,333	0	0,194	0,061
20	1,3-Butadiene	0,195	54,092	1	0	0,164	0,063
21	Cyclobutane	0,185	56,108	1	0	0,15	0,057
22	cis-2-butene	0,203	56,108	1	0	0,182	0,069
23	Isopropyl Alcohol	0,665	60,096	-0,556	3,434	0,316	0,065
24	Ethyl Acetate	0,361	88,106	-0,1	3,714	0,18	0,082
25	Propyl methanoate	0,32	88,106	0,5	0	0,173	0,076
26	2-chlorobutane	0,267	92,568	-0,375	3,714	0,15	0,083
27	Butane	0,2	58,123	1	0	0,148	0,056
28	Isobutane	0,186	58,123	1	3,434	0,159	0,06
29	Diethyl sulfide	0,295	90,189	-0,583	0	0,166	0,11
30	1-pentene	0,237	70,134	1	0	0,184	0,07
31	2-methyltetrahydrofuran	0,292	86,134	-0,314	4,111	0,165	0,063
32	1-Propanol	0,629	60,096	-0,333	0	0,309	0,063
33	pentane	0,252	72,15	1	0	0,144	0,055
34	2-methylbutane	0,229	72,15	1	3,714	0,15	0,057
35	neopentane	0,197	72,15	1	5,485	0,137	0,052
36	3-methyl-1-butanol	0,559	88,15	-0,04	3,714	0,258	0,054
37	ethyl propyl ether	0,328	88,15	-0,4	0	0,152	0,058
38	benzene	0,21	78,114	1	0	0,13	0,057
39	cyclohexane	0,211	84,161	1	0	0,116	0,046
40	1-hexene	0,281	84,161	1	0	0,169	0,07
41	propyl propanoate	0,373	116,16	0,143	3,932	0,139	0,053

N°	Composé	ω_{EXP}	Mol. Wt.	MATS2m	ESpm05u	R1u+	R1p+
42	3-methylbutyl methanoate	0,4	116,16	0,524	3,714	0,172	0,065
43	hexane	0,3	86,177	1	0	0,127	0,052
44	2-methylpentane	0,278	86,177	1	3,714	0,143	0,055
45	2,2-dimethylbutane	0,233	86,177	1	5,602	0,136	0,052
46	1-hexanol	0,573	102,177	-0,067	0	0,169	0,056
47	4-methyl-2-pentanol	0,552	102,177	-0,167	4,394	0,225	0,054
48	toluene	0,264	92,141	1	3,932	0,245	0,093
49	4-methylphenol	0,51	108,14	-0,086	4,615	0,228	0,087
50	butyl-2-propenoate	0,312	128,175	0,205	3,932	0,139	0,071
51	methylcyclohexane	0,235	98,188	1	3,932	0,127	0,048
52	ethylcyclopentane	0,27	98,188	1	4,263	0,133	0,051
53	propyl butanoate	0,399	130,187	0,205	3,932	0,123	0,047
54	heptane	0,35	100,204	1	0	0,141	0,054
55	2-methylhexane	0,331	100,204	1	3,714	0,139	0,053
56	Ethanol	0,649	46,069	-1	0	0,335	0,066
57	3,3-dimethylpentane	0,269	100,204	1	5,707	0,136	0,052
58	1-heptanol	0,588	116,203	-0,048	0	0,16	0,054
59	1,3-dimethylbenzene	0,327	106,167	1	4,615	0,195	0,074
60	cyclooctane	0,254	112,215	1	0	0,099	0,044
61	1-octene	0,393	112,215	1	0	0,166	0,065
62	octane	0,399	114,231	1	0	0,128	0,052
63	2-methylheptane	0,378	114,231	1	3,714	0,135	0,052
64	4-methylheptane	0,371	114,231	1	3,932	0,128	0,049
65	2,4-dimethylhexane	0,344	114,231	1	4,511	0,13	0,05
66	3-ethyl-2-methylpentane	0,331	114,231	1	4,796	0,132	0,05
67	3-ethyl-3-methylpentane	0,305	114,231	1	5,802	0,129	0,049
68	2,2,4-trimethylpentane	0,304	114,231	1	5,74	0,122	0,046
69	2,3,3-trimethylpentane	0,291	114,231	1	5,969	0,114	0,043
70	2,2,3,3-tetramethylbutane	0,248	114,231	1	6,494	0,102	0,039
71	1-octanol	0,594	130,23	-0,036	0	0,172	0,051
72	propylbenzene	0,345	120,194	1	4,111	0,171	0,071
73	1-ethyl-4-methylbenzene	0,364	120,194	1	4,71	0,178	0,068
74	1,2,3-trimethylbenzene	0,367	120,194	1	5,252	0,148	0,057
75	1,3,5-trimethylbenzene	0,399	120,194	1	5,017	0,178	0,068
76	3-methylbutyl butanoate	0,583	158,241	0,278	4,511	0,124	0,047
77	nonane	0,445	128,258	1	0	0,137	0,052
78	2,2-dimethylheptane	0,383	128,258	1	5,602	0,118	0,047
79	2,2,3,3-tetramethylpentane	0,304	128,258	1	6,538	0,115	0,044
80	2,2,3,4-tetramethylpentane	0,301	128,258	1	6,066	0,111	0,042
81	2,2,4,4-tetramethylpentane	0,314	128,258	1	6,293	0,106	0,04
82	naphthalene	0,304	128,174	1	4,949	0,111	0,059
83	butylbenzene	0,393	134,221	1	4,111	0,133	0,07
84	-methylpropylbenzene	0,383	134,221	1	4,615	0,152	0,069

N°	Composé	ω_{EXP}	Mol. Wt.	MATS2m	ESpm05u	R1u+	R1p+
85	1,4-diethylbenzene	0,403	134,221	1	4,796	0,122	0,055
86	1-(1-methylethyl)-4-methylbenzene	0,376	134,221	1	5,142	0,159	0,061
87	1,2,4,5-tetramethylbenzene	0,423	134,221	1	5,485	0,125	0,048
88	trans-bicyclo[4,4,0]decane	0,303	138,253	1	4,949	0,095	0,044
89	decane	0,49	142,285	1	0	0,12	0,05
90	3,3,5-trimethylheptane	0,383	142,285	1	5,861	0,111	0,043
91	2,2,3,3-tetramethylhexane	0,366	142,285	1	6,538	0,11	0,043
92	2,2,5,5-tetramethylhexane	0,377	142,285	1	6,293	0,096	0,036
93	1-decanol	0,661	158,284	-0,022	0	0,175	0,039
94	1-methylnaphthalene	0,348	142,2	1	5,352	0,21	0,08
95	2-methylnaphthalene	0,374	142,2	1	5,252	0,243	0,093
96	undecane	0,537	156,312	1	0	0,117	0,047
97	1,1'-biphenyl	0,404	154,211	1	4,949	0,155	0,067
98	odecane	0,576	170,338	1	0	0,117	0,048
99	diphenylmethane	0,481	168,238	1	4,796	0,139	0,072
100	tridecane	0,618	184,365	1	0	0,118	0,049
101	phenanthrene	0,479	178,233	1	5,707	0,103	0,057
102	anthracene	0,501	178,233	1	5,638	0,094	0,054
103	tetradecane	0,644	198,392	1	0	0,107	0,049
104	pentadecane	0,685	212,419	1	0	0,108	0,043
105	hexadecane	0,718	226,446	1	0	0,115	0,048
106	2,2,4,4,6,8,8-heptamethylnonane	0,548	226,446	1	6,792	0,085	0,032
107	heptadecane	0,753	240,473	1	0	0,086	0,046
108	octadecane	0,8	254,5	1	0	0,092	0,043
109	nonadecane	0,845	268,527	1	0	0,093	0,043
110	eicosane	0,865	282,554	1	0	0,094	0,045
111	Methylamine	0,283	31,057	0	0	0,275	0,065
112	1,1-difluoroethane (R-152a)	0,276	66,051	-0,333	3,434	0,157	0,06
113	Acetone	0,307	58,08	-0,556	3,434	0,176	0,067
114	Propane	0,152	44,097	1	0	0,164	0,062
115	1-Butene	0,194	56,108	1	0	0,191	0,073
116	trans-2-butene	0,218	56,108	1	0	0,178	0,068
117	2-methylpropene	0,199	56,108	1	3,434	0,181	0,069
118	1-butanol	0,59	74,123	-0,167	0	0,269	0,054
119	2-methyl-1-propanol	0,59	74,123	-0,063	3,714	0,249	0,062
120	2-methyl-2-propanol	0,613	74,123	-0,375	5,485	0,295	0,056
121	diethyl ether	0,281	74,123	-0,583	0	0,167	0,063
122	1-Butanamine	0,338	73,138	-0,167	0	0,187	0,057
123	1-pentyne	0,394	68,119	1	0	0,285	0,109
124	2-methyl-2-butene	0,339	70,134	1	3,714	0,168	0,064
125	3-methyl-1-butene	0,211	70,134	1	3,714	0,183	0,07
126	2-pentanone	0,346	86,134	-0,28	3,714	0,155	0,061
127	3-pentanone	0,342	86,134	-0,28	3,932	0,159	0,061

N°	Composé	ω_{EXP}	Mol. Wt.	MATS2m	ESpm05u	R1u+	R1p+
128	ethyl propanoate	0,39	102,133	0,05	3,932	0,156	0,059
129	1-pentanol	0,579	88,15	-0,1	0	0,271	0,053
130	methylcyclopentane	0,227	84,161	1	4,111	0,14	0,053
131	4-methyl-2-pentanone	0,351	100,161	-0,167	4,394	0,139	0,058
132	ethyl butanoate	0,463	116,16	0,143	3,932	0,139	0,06
133	butyl ethanoate	0,407	116,16	0,143	3,714	0,153	0,058
134	2-methylpropyl ethanoate	0,456	116,16	0,167	4,394	0,158	0,06
135	3-methylpentane	0,273	86,177	1	3,932	0,14	0,057
136	2,3-dimethylbutane	0,248	86,177	1	4,615	0,124	0,047
137	2-methyl-1-pentanol	0,498	102,177	-0,028	3,932	0,221	0,053
138	3-methylhexane	0,323	100,204	1	3,932	0,137	0,052
139	2,2-dimethylpentane	0,287	100,204	1	5,602	0,13	0,05
140	2,3-dimethylpentane	0,297	100,204	1	4,71	0,138	0,053
141	2,4-dimethylpentane	0,304	100,204	1	4,394	0,127	0,048
142	2,2,3-trimethylbutane	0,25	100,204	1	5,889	0,128	0,049
143	1,4-dimethylbenzene	0,322	106,167	1	4,615	0,184	0,07
144	3-methylheptane	0,371	114,231	1	3,932	0,131	0,05
145	3-ethylhexane	0,362	114,231	1	4,111	0,133	0,05
146	2,2-dimethylhexane	0,339	114,231	1	5,602	0,125	0,052
147	2,3-dimethylhexane	0,347	114,231	1	4,71	0,119	0,047
148	2,5-dimethylhexane	0,357	114,231	1	4,394	0,13	0,049
149	3,3-dimethylhexane	0,32	114,231	1	5,707	0,129	0,049
150	3,4-dimethylhexane	0,338	114,231	1	4,796	0,116	0,044
151	2,2,3-trimethylpentane	0,298	114,231	1	5,916	0,119	0,045
152	2,3,4-trimethylpentane	0,316	114,231	1	5,142	0,126	0,048
153	1-methylethylbenzene	0,326	120,194	1	4,796	0,165	0,07
154	1,2,4-trimethylbenzene	0,377	120,194	1	5,142	0,164	0,062
155	cis-bicyclo[4,4,0]decane	0,276	138,253	1	4,949	0,1	0,045
156	2-Butanone	0,322	72,107	-0,375	3,714	0,164	0,067
157	propyl acetate	0,389	102,133	0,05	3,714	0,159	0,061
158	3-ethylpentane	0,311	100,204	1	4,111	0,141	0,054

Tableau A-2: Valeurs des EQM en fonction du nombre de neurones.

nombre de neurones	EQMC	EQMP	$EQMP_{ext}$
1	0,123682	0,12721	0,0732421
2	0,0534817	0,0599535	0,0629162
3	0,0519868	0,0550091	0,0617974
4	0,0509257	0,0538799	0,0600765
5	0,0496429	0,0530283	0,0577533
6	0,0487858	0,0532481	0,0568758
7	0,0490442	0,0533105	0,0577845
8	0,0495357	0,0537687	0,0576972
9	0,0506053	0,0534352	0,0604932
10	0,0497437	0,0534138	0,0595659

Tableau A-3: Valeurs des EQM en fonction du nombre d'itérations.

nombre d'itérations	EQMC	EQMP	$EQMP_{ext}$
1647	0,0459787	0,0511572	0,0594339
1657	0,0477829	0,0536774	0,0577878
1667	0,0459853	0,0511888	0,0595836
1677	0,047612	0,0512183	0,0577736
1687	0,043218	0,046261	0,047012
1697	0,0418	0,0447	0,0459
1708	0,0449813	0,0486448	0,052942
1718	0,0459785	0,0528176	0,0593577
1728	0,0459761	0,0529962	0,059367
1738	0,0459803	0,0503089	0,0595678
1748	0,0459948	0,0509301	0,059679
1758	0,0459784	0,0511851	0,0595007
1768	0,0478142	0,0524525	0,0578195
1778	0,0478835	0,0514617	0,0578534
1788	0,045972	0,0517525	0,0594701
1798	0,0459803	0,0520025	0,0595198
1808	0,0459828	0,0521115	0,0595845
1818	0,0459716	0,0512895	0,0594147
1828	0,0459792	0,0515707	0,0595596
1838	0,0476862	0,0516189	0,0577867
1848	0,0459766	0,0540851	0,0595139
1858	0,0459703	0,0499208	0,0594904
1868	0,0460064	0,0530497	0,0597586
1878	0,0459701	0,0527947	0,0594701
1888	0,0478583	0,0520603	0,0579191
1898	0,0459871	0,0510007	0,0596275
1909	0,0459854	0,0533482	0,0596268

Tableau A-4: Valeurs des volumes critiques (V_C) et des descripteurs du modèle.

N°	Composé	V_C	Mol. Wt.	Mor18v	Mor17p
1	Chlorotrifluoromethane	180,3	104,459	-0,043	-0,038
2	Trichloromonofluoromethane	248	137,368	-0,165	0,054
3	Carbon Tetrachloride	276	153,822	-0,254	0,259
4	Tetrafluoromethane	140,7	88,005	-0,064	0,027
5	Trichloromethane	240	119,377	-0,151	0,126
6	Fluoroform	133	70,014	-0,058	0,013
7	Difluoromethane	121	52,024	-0,035	-0,022
8	Chloromethane	143	50,488	-0,042	-0,055
9	Methyl fluoride	113,3	34,033	-0,028	-0,06
10	nitromethane	173	61,04	0,033	-0,07
11	Methane	98,6	16,043	-0,04	-0,091
12	methanol	118	32,042	-0,033	-0,071
13	Methanethiol	145	48,109	-0,053	-0,024
14	Pentafluoroethyl chloride	256	154,467	-0,018	-0,002
15	etrafluoroethene	172	100,016	-0,079	0,074
16	Acetylene	112,2	26,038	0,035	-0,054
17	1,1,1-tri?uoroethane	193,6	84,041	-0,056	-0,06
18	Ethylene	131,1	28,054	-0,083	-0,032
19	1,1-dichloroethane	236	98,959	-0,087	-0,092
20	1,2-dichloroethane	220	98,959	-0,051	-0,161
21	Methyl formate	172	60,053	-0,017	-0,061
22	Ethyl Chloride	199	64,514	-0,056	-0,153
23	Ethane	145,5	30,07	-0,063	-0,165
24	Ethanethiol	207	62,136	-0,059	-0,124
25	Dimethyl sulfide	201	62,136	-0,087	-0,123
26	Propyne	163,5	40,065	-0,02	-0,101
27	Cyclopropane	162,8	42,081	0,066	-0,383
28	1,2-dichloropropane	287,66	112,187	-0,061	-0,238
29	methyl ethanoate	228	74,079	-0,03	-0,131
30	Ethyl formate	229	74,079	-0,011	-0,149
31	n-Propyl chloride	254	78,541	-0,075	-0,174
32	Propane	200	44,097	-0,082	-0,225
33	1-Propanol	219	60,096	-0,068	-0,207
34	N,N-dimethylmethanamine	254	59,111	-0,088	-0,229
35	Furan	218	68,075	-0,136	-0,244
36	Thiophene	219	84,142	-0,21	-0,064
37	1-Butyne	208	54,092	-0,013	-0,175
38	1,3-Butadiene	221	54,092	-0,169	-0,045
39	Cyclobutane	218	56,108	0,125	-0,447
40	2-methylpropene	238,8	56,108	-0,179	-0,149
41	1,4-dioxane	238	88,106	-0,076	-0,156

N°	Composé	V_c	Mol. Wt.	Mor18v	Mor17p
42	thyl Acetate	286	88,106	-0,089	-0,22
43	propyl methanoate	285	88,106	-0,038	-0,229
44	2-chlorobutane	312	92,568	-0,071	-0,308
45	Butane	255	58,123	-0,08	-0,313
46	2-methyl-2-propanol	275	74,123	-0,046	-0,284
47	diethyl ether	280	74,123	-0,077	-0,284
48	1,2-dimethoxyethane	270,64	90,126	-0,079	-0,262
49	1,2-Butanediol	303,05	90,126	-0,073	-0,258
50	Diethyl sulfide	318	90,189	-0,078	-0,362
51	pyridine	254	79,101	-0,217	-0,126
52	cyclopentene	245	68,119	0,022	-0,291
53	cyclopentane	260	70,134	-0,023	-0,288
54	1-pentene	298,4	70,134	-0,179	-0,204
55	2-methyl-2-butene	292	70,134	-0,197	-0,254
56	3-methyl-2-butanone	310	86,134	-0,062	-0,331
57	2-methyltetrahydrofuran	267	86,134	0,015	-0,329
58	pentanoic acid	336,2	102,133	-0,074	-0,271
59	ethyl propanoate	345	102,133	-0,095	-0,247
60	pentane	311	72,15	-0,09	-0,392
61	neopentane	303,2	72,15	-0,037	-0,393
62	1-pentanol	326	88,15	-0,066	-0,373
63	benzene	256	78,114	-0,342	-0,129
64	phenol	229	94,113	-0,349	-0,091
65	yclohexene	296,88	82,145	-0,262	-0,222
66	4-methyl-3-penten-2-one	353,43	98,147	-0,144	-0,243
67	cyclohexane	308	84,161	-0,184	-0,346
68	1-hexene	355,1	84,161	-0,145	-0,31
69	thyl butanoate	421	116,16	-0,124	-0,324
70	3-methylbutyl methanoate	411,4	116,16	-0,123	-0,296
71	hexane	368	86,177	-0,095	-0,529
72	2-methylpentane	366,7	86,177	-0,069	-0,577
73	2,2-dimethylbutane	359,1	86,177	-0,041	-0,515
74	1-hexanol	381	102,177	-0,105	-0,467
75	4-methylphenol	277	108,14	-0,389	-0,141
76	butyl-2-propenoate	427,54	128,175	-0,123	-0,288
77	cycloheptane	359	98,188	0,026	-0,619
78	methylcyclohexane	368	98,188	-0,165	-0,451
79	is-1,3-dimethylcyclopentane	363,3	98,188	-0,091	-0,366
80	heptanoic acid	429,7	130,187	-0,108	-0,424
81	2-methylhexane	421	100,204	-0,084	-0,645
82	3-ethylpentane	415,8	100,204	-0,109	-0,63
83	2,2,3-trimethylbutane	397,6	100,204	-0,027	-0,626
84	1-heptanol	435	116,203	-0,093	-0,535

N°	Composé	V_c	Mol. Wt.	Mor18v	Mor17p
85	1,4-dimethylbenzene	378	106,167	-0,41	-0,315
86	cyclooctane	410	112,215	-0,016	-0,686
87	1-octene	468	112,215	-0,205	-0,51
88	octane	492	114,231	-0,189	-0,648
89	2-methylheptane	488,2	114,231	-0,116	-0,687
90	4-methylheptane	476	114,231	-0,126	-0,724
91	2,4-dimethylhexane	472	114,231	-0,179	-0,725
92	3-ethyl-2-methylpentane	445,3	114,231	-0,123	-0,659
93	3-ethyl-3-methylpentane	455,1	114,231	-0,089	-0,758
94	2,2,4-trimethylpentane	469,7	114,231	-0,081	-0,584
95	2,3,3-trimethylpentane	455,1	114,231	-0,113	-0,698
96	2,2,3,3-tetramethylbutane	482	114,231	-0,101	-0,629
97	1-octanol	490	130,23	-0,112	-0,671
98	propylbenzene	440	120,194	-0,431	-0,364
99	1-ethyl-4-methylbenzene	440	120,194	-0,473	-0,369
100	1,2,3-trimethylbenzene	435	120,194	-0,55	-0,219
101	1,2,4-trimethylbenzene	435	120,194	-0,534	-0,275
102	1,3,5-trimethylbenzene	430	120,194	-0,447	-0,449
103	1-nonene	526	126,242	-0,175	-0,626
104	nonane	555	128,258	-0,134	-0,799
105	2,2,5-trimethylhexane	519	128,258	-0,136	-0,831
106	2,2,3,3-tetramethylpentane	478	128,258	-0,183	-0,732
107	2,2,3,4-tetramethylpentane	490	128,258	-0,076	-0,739
108	2,2,4,4-tetramethylpentane	504	128,258	0,048	-0,789
109	2,3,3,4-tetramethylpentane	493	128,258	-0,055	-0,786
110	1-nonanol	544	144,257	0,048	-0,846
111	naphthalene	407	128,174	-0,612	-0,122
112	1,2,3,4-tetrahydronaphthalene	408	132,205	-0,508	-0,367
113	butylbenzene	497	134,221	-0,429	-0,398
114	1,4-diethylbenzene	480,5	134,221	-0,482	-0,398
115	1-(1-methylethyl)-4-methylbenzene	497	134,221	-0,489	-0,458
116	trans-bicyclo[4,4,0]decane	480	138,253	-0,233	-0,574
117	1-decene	584	140,269	-0,25	-0,652
118	decane	624	142,285	-0,17	-0,85
119	1-decanol	600	158,284	-0,116	-0,862
120	undecane	689	156,312	-0,145	-0,966
121	1,1'-biphenyl	497	154,211	-0,838	-0,271
122	1,3,5-triethylbenzene	624,14	162,276	-0,5	-0,629
123	1,3-dimethyltricyclo[3,3,1,13,7]decane	571,45	164,292	-0,351	-0,744
124	dodecane	754	170,338	-0,244	-1,048
125	1-dodecanol	718	186,338	-0,165	-0,966
126	diphenylmethane	563	168,238	-0,708	-0,246
127	tridecane	823	184,365	-0,206	-1,18

N°	Composé	V_c	Mol. Wt.	Mor18v	Mor17p
128	phenanthrene	554	178,233	-0,907	-0,226
129	anthracene	554	178,233	-0,881	-0,116
130	tetradecane	894	198,392	-0,173	-1,23
131	pentadecane	966	212,419	-0,223	-1,182
132	hexadecane	1034	226,446	-0,194	-1,496
133	heptadecane	1103	240,473	-0,252	-1,543
134	1,4-diphenylbenzene	729	230,309	-1,264	-0,187
135	octadecane	1189	254,5	-0,133	-1,788
136	1,1-di?uoroethene	153,5	64,035	-0,073	0,013
137	1,1-di?uoroethane (R-152a)	181	66,051	-0,027	-0,106
138	ethanoic acid	171	60,053	-0,039	-0,06
139	Ethanol	167	46,069	-0,036	-0,156
140	Dimethyl ether	170	46,069	-0,057	-0,144
141	Dimethylamine	180	45,084	-0,082	-0,128
142	Propene	184,6	42,081	-0,131	-0,091
143	Acetone	209	58,08	-0,028	-0,174
144	Isopropyl Alcohol	220	60,096	-0,051	-0,208
145	methyl ethyl ether	221	60,096	-0,073	-0,212
146	2-Propanamine	221	59,111	-0,071	-0,2
147	1-Butene	240,8	56,108	-0,158	-0,135
148	trans-2-butene	237,7	56,108	-0,154	-0,173
149	cis-2-butene	233,8	56,108	-0,164	-0,193
150	2-Butanone	267	72,107	-0,045	-0,235
151	Tetrahydrofuran	224	72,107	-0,017	-0,228
152	Butanoic acid	292	88,106	-0,08	-0,178
153	Isobutane	262,7	58,123	-0,076	-0,294
154	1-butanol	275	74,123	-0,091	-0,269
155	2-methyl-1-propanol	273	74,123	-0,061	-0,271
156	Cyclopentanone	268	84,118	-0,004	-0,195
157	cis-2-pentene	302,1	70,134	-0,174	-0,25
158	3-methyl-1-butene	304,9	70,134	-0,211	-0,158
159	2-pentanone	301	86,134	-0,046	-0,345
160	3-pentanone	336	86,134	-0,077	-0,288
161	propyl acetate	345	102,133	-0,065	-0,258
162	2-methylbutane	308,3	72,15	-0,071	-0,434
163	2-pentanol	329	88,15	-0,095	-0,384
164	ethyl propyl ether	339	88,15	-0,099	-0,336
165	methylcyclopentane	319	84,161	-0,052	-0,317
166	cyclohexanol	333,88	100,161	0,024	-0,514
167	4-methyl-2-pentanone	340,6	100,161	-0,044	-0,384
168	hexanoic acid	377,2	116,16	-0,093	-0,339
169	butyl ethanoate	412,8	116,16	-0,073	-0,325
170	2-methylpropyl ethanoate	413	116,16	-0,053	-0,334

N°	Composé	V_c	Mol. Wt.	Mor18v	Mor17p
171	methylpentane	366,7	86,177	-0,124	-0,415
172	2,3-dimethylbutane	357,6	86,177	-0,066	-0,565
173	toluene	316	92,141	-0,378	-0,206
174	ethylcyclopentane	375	98,188	-0,02	-0,446
175	trans-1,3-dimethylcyclopentane	363,3	98,188	-0,095	-0,349
176	3-methylhexane	404	100,204	-0,129	-0,6
177	2,2-dimethylpentane	415,8	100,204	-0,107	-0,621
178	2,3-dimethylpentane	393	100,204	-0,118	-0,605
179	2,4-dimethylpentane	417,5	100,204	-0,12	-0,495
180	3,3-dimethylpentane	414,1	100,204	-0,047	-0,633
181	1,3-dimethylbenzene	375	106,167	-0,421	-0,279
182	3-methylheptane	471,1	114,231	-0,123	-0,674
183	3-ethylhexane	460,5	114,231	-0,105	-0,742
184	2,2-dimethylhexane	478	114,231	-0,164	-0,66
185	2,3-dimethylhexane	468,2	114,231	-0,117	-0,688
186	2,5-dimethylhexane	482	114,231	-0,115	-0,719
187	3,3-dimethylhexane	442,8	114,231	-0,123	-0,72
188	3,4-dimethylhexane	458,8	114,231	-0,124	-0,713
189	2,2,3-trimethylpentane	436	114,231	-0,127	-0,642
190	2,3,4-trimethylpentane	456,2	114,231	-0,204	-0,611
191	1-methylethylbenzene	434,7	120,194	-0,419	-0,36
192	cis-bicyclo[4,4,0]decane	480	138,253	-0,073	-0,888

Tableau A-5: Valeurs des volumes critiques (V_C) prédites par les MCG.

Composé	V_C	$V_{C(JR)}{}^{a}$	$V_{C(Ly)}{}^{b}$
trans-bicyclo[4,4,0]decane	480	477,6039	488,1062
2,2,3,4-tetramethylpentane	490	516,6405	513,1396
1,2,4-trimethylbenzene	435	431,5661	424,0650
1-ethyl-4-methylbenzene	440	431,5661	425,0651
2,3,4-trimethylpentane	456,2	465,6283	468,1290
2,2,3-trimethylpentane	436	466,6286	462,1273
3-ethyl-2-methylpentane	445,3	471,6299	472,1301
3,4-dimethylhexane	458,8	471,6299	472,1301
3,3-dimethylhexane	442,8	472,6302	466,1284
2,5-dimethylhexane	482	471,6299	472,1301
2,4-dimethylhexane	472	471,6299	472,1301
2,3-dimethylhexane	468,2	471,6299	472,1301
2,2-dimethylhexane	478	472,6302	466,1284
3-methylheptane	471,1	477,6316	476,1312
1,4-dimethylbenzene	378	375,5531	370,0523
2,2,3-trimethylbutane	397,6	410,6150	407,1140
2,3-dimethylpentane	393	415,6164	417,1168
2,2-dimethylpentane	415,8	416,6166	411,1151
3-ethylpentane	415,8	421,6180	421,1179
2-methylhexane	421	421,6180	421,1179
cis-1,3-dimethylcyclopentane	363,3	367,5896	375,5916
ethylcyclopentane	375	368,5898	374,0912
2,2-dimethylbutane	359,1	360,6031	356,1018
methylpentane	366,7	365,6045	366,1046
ethyl butanoate	421	395,5740	395,0739
4-methyl-2-pentanone	340,6	371,5783	371,0782
1-hexene	355,1	352,5859	350,0853
2-pentanol	329	328,5813	329,0814
2-methylbutane	308,3	309,5910	311,0915
propyl acetate	345	339,5608	340,0609
3-pentanone	336	321,5661	320,0658
2-pentanone	301	321,5661	320,0658
2-methyl-2-butene	292	296,5723	286,0697
cis-2-pentene	302,1	295,5720	295,0719
1-pentene	298,4	296,5723	295,0719
cyclopentene	245	243,5464	247,5472
2-methyl-1-propanol	273	272,5676	274,0680
Butane	255	259,5795	260,0796
2-chlorobutane	312	302,5416	305,0420
2-methylpropene	238,8	241,5589	231,0563
cis-2-butene	233,8	239,5584	240,0585

Composé	V_C	$V_{C(JR)}$	$V_{C(Ly)}$
trans-2-butene	237,7	239,5584	240,0585
1Butene	240,8	240,5586	240,0585
2-Propanamine	221	226,5756	229,0765
Isopropyl Alcohol	220	216,5540	219,0546
1-Propanol	219	222,5555	223,0556
Propane	200	203,5664	205,0669
methyl ethanoate	228	227,5352	230,0356
Dimethylamine	180	182,5661	187,0677
Ethanethiol	207	201,5518	205,0527
Dimethyl ether	170	165,5415	170,0427
Ethanol	167	166,5418	168,0422
Ethane	145,5	147,5538	150,0547
Ethyl Chloride	199	196,5180	199,0182
Methyl formate	172	182,5244	115,7455
ethanoic acid	171	172,5231	173,0232
1,1-difluoroethene	153,5	165,5153	157,0145
methanol	118	110,5281	113,0287
octadecane	1189	1043,7693	1030,2658
1,4-diphenylbenzene	729	719,5507	702,0495
eptadecane	1103	987,7557	975,2524
hexadecane	1034	931,7420	920,2391
pentadecane	966	875,7284	865,2257
tetradecane	894	819,7148	810,2123
anthracene	554	555,5337	554,0336
phenanthrene	554	555,5337	554,0336
tridecane	823	763,7012	755,1989
diphenylmethane	563	547,5508	537,0499
1-dodecanol	718	726,6784	718,1763
odecane	754	707,6876	700,1856
1,3-dimethyltricyclo[3,3,1,13,7]decane	571,45	565,6129	571,1140
1,3,5-triethylbenzene	624,14	599,6060	589,1041
1,1'-biphenyl	497	491,5389	482,0381
undecane	689	651,6740	645,1722
1-decanol	600	614,6511	608,1495
decane	624	595,6604	590,1589
1-decene	584	576,6406	570,1390
cis-bicyclo[4,4,0]decane	480	477,6039	488,1062
1-(1-methylethyl)-4-methylbenzene	497	481,5783	476,0775
1,4-diethylbenzene	480,5	487,5793	480,0781
butylbenzene	497	487,5793	481,0783
1,2,3,4-tetrahydronaphthalene	408	437,5586	438,0587
naphthalene	407	409,5303	408,0302
1-nonanol	544	558,6374	553,1361

Composé	V_C	$V_{C(JR)}$	$V_{C(Ly)}$
2,3,3,4-tetramethylpentane	493	516,6405	513,1396
2,2,4,4-tetramethylpentane	504	517,6408	507,1380
2,2,3,3-tetramethylpentane	478	517,6408	507,1380
2,2,5-trimethylhexane	519	522,6422	517,1407
nonane	555	539,6468	535,1456
1-nonene	526	520,6269	515,1256
1,3,5-trimethylbenzene	430	431,5661	424,0650
1,2,3-trimethylbenzene	435	431,5661	424,0650
1-methylethylbenzene	434,7	425,5652	422,0647
propylbenzene	440	431,5661	426,0653
1-octanol	490	502,6238	498,1227
2,2,3,3-tetramethylbutane	482	461,6272	452,1246
2,3,3-trimethylpentane	455,1	466,6286	462,1273
2,2,4-trimethylpentane	469,7	466,6286	462,1273
3-ethyl-3-methylpentane	455,1	472,6302	466,1284
3-ethylhexane	460,5	477,6316	476,1312
4-methylheptane	476	477,6316	476,1312
2-methylheptane	488,2	477,6316	476,1312
octane	492	483,6332	480,1323
1-octene	468	464,6133	460,1122
cyclooctane	410	401,5979	396,0966
1,3-dimethylbenzene	375	375,5531	370,0523
1-heptanol	435	446,6101	443,1093
3,3-dimethylpentane	414,1	416,6166	411,1151
2,4-dimethylpentane	417,5	415,6164	417,1168
3-methylhexane	404	421,6180	421,1179
heptanoic acid	429,7	452,5874	448,0865
trans-1,3-dimethylcyclopentane	363,3	367,5896	375,5916
methylcyclohexane	368	360,5879	363,5886
cycloheptane	359	353,5862	351,5857
butyl-2-propenoate	427,54	432,5709	430,0705
4-methylphenol	277	448,4576	450,0766
toluene	316	319,5402	316,0397
1-hexanol	381	390,5965	388,0958
2,3-dimethylbutane	357,6	359,6028	362,1035
2-methylpentane	366,7	365,6045	366,1046
hexane	368	371,6062	370,1058
3-methylbutyl methanoate	411,4	400,5749	331,7920
2-methylpropyl ethanoate	413	389,5728	391,0731
butyl ethanoate	412,8	395,5740	395,0739
hexanoic acid	377,2	396,5741	393,0735
cyclohexanol	333,88	323,5682	326,5688
methylcyclopentane	319	312,5762	319,0778

Composé	V_C	$V_{C(JR)}$	$V_{C(Ly)}$
cyclohexane	308	305,5745	307,0749
4-methyl-3-penten-2-one	353,43	358,5621	346,0599
cyclohexene	296,88	291,5582	292,0583
phenol	229	282,5263	279,0260
benzene	256	263,5275	262,0274
ethyl propyl ether	339	333,5826	335,0829
1-pentanol	326	334,5828	333,0824
neopentane	303,2	259,8342	258,3641
pentane	311	315,5928	315,0926
ethyl propanoate	345	339,5608	340,0609
pentanoic acid	336,2	340,5610	338,0606
2-methyltetrahydrofuran	267	277,5570	282,5581
3-methyl-2-butanone	310	315,5648	316,0649
3-methyl-1-butene	304,9	290,5708	291,0710
cyclopentane	260	257,5628	262,5640
cyclopentanone	268	264,5427	268,0433
pyridine	254	256,5328	/
Diethyl sulfide	318	313,5793	315,0796
1,2-Butanediol	303,05	291,5615	292,0616
1,2-dimethoxyethane	270,64	295,5623	300,0632
diethyl ether	280	277,5689	280,0695
2-methyl-2-propanol	275	267,5664	264,0655
1-butanol	275	278,5691	278,0690
Isobutane	262,7	253,5776	253,5776
propyl methanoate	285	294,5497	225,7681
Ethyl Acetate	286	283,5479	285,0481
1,4-dioxane	238	235,5398	234,0395
Butanoic acid	292	284,5480	283,0478
Tetrahydrofuran	224	222,5441	226,0448
2-Butanone	267	265,5526	265,0525
Cyclobutane	218	209,5511	218,0532
1,3-Butadiene	221	221,5391	220,0389
1-Butyne	208	221,5391	222,0392
Thiophene	219	219,5272	233,0289
Furan	218	194,5173	196,0174
N,N-dimethylmethanamine	254	221,5740	247,0825
methyl ethyl ether	221	221,5552	225,0561
n-Propyl chloride	254	252,5300	254,0301
Ethyl formate	229	238,5369	170,7564
Acetone	209	209,5392	210,0393
1,2-dichloropropane	287,66	295,5111	299,0113
Cyclopropane	162,8	161,5394	173,5423
Propene	184,6	184,5450	185,0451

Composé	V_C	$V_{C(JR)}$	$V_{C(Ly)}$
Propyne	163,5	165,5254	167,0256
Dimethyl sulfide	201	201,5518	205,0527
1,1-difluoroethane (R-152a)	181	177,5270	182,0276
1,2-dichloroethane	220	245,5021	248,0021
1,1-dichloroethane	236	239,5020	244,0021
Ethylene	131,1	129,5316	130,0317
1,1,1-trifluoroethane	193,6	190,5217	190,0216
Acetylene	112,2	109,5114	112,0117
tetrafluoroethene	172	201,5100	184,0092
entafluoroethyl chloride	256	264,5061	261,0061
Methanethiol	145	145,5379	150,0391
Methane	98,6	/	/
itromethane	173	173,5284	173,0283
Methyl fluoride	113,3	109,5228	113,0235
Chloromethane	143	140,5069	144,0071
Difluoromethane	121	127,5162	131,0167
Fluoroform	133	139,5123	145,0127
Trichloromethane	240	232,4851	237,9848
Tetrafluoromethane	140,7	152,5099	153,0099
Carbon Tetrachloride	276	276,4772	276,9771
Trichloromonofluoromethane	248	245,4855	245,9855
Chlorotrifluoromethane	180,3	183,5019	184,0019

[a] Valeurs prédites par la MCG de Joback et Reid (**J&R87**) [b] Valeurs prédites par celle de Lydersen (**Ly55**)

Relation structure/ facteur acentrique d'alcools et de phénols : approche algorithme génétique – régression linéaire multiple.

Structure / acentric factor relationship of alcohols and phenols: genetic algorithm – multiple linear regression approach

Hamza Haddag[1], Amel Bouakkadia[1, 2], Leila. Lourici[1, 3*], Nasr Eddine Chakri[1], Djelloul Messadi[1]

[1]*Laboratoire de Sécurité Environnementale et Alimentaire, Université BADJI Mokhtar, BP 12, 23000 Annaba, Algérie.*

[2]*Université Abbès Laghrour Khenchela , Algérie.*

[3]*Université Chadli Bendjedid -36000 - El Tarf, Algérie.*

Soumis le 04/09/2016 *Révisé le 05/02/2017* *Accepté le 21/02/2017*

ملخص

المعامل الغيرمركزي لـ 18 مركبا هيدروكسيليا (كحولات،فينولات)، ربطت خطيا بمواصفين جز يئيين من الصنف الهندسي تم اختيارهما بواسطة الخوارزمي الجيني من بين 1600حسبت باستعمال برنامج النمذجة الجزيئية.DRAGON. الأحصاءات المختلفة (معاملا التحديد المتعدد و التنبؤ، جذورالأخطاء المربعة المتوسطة ...) تبين جودة، متانة و قدرة التنبؤ الداخلية الجيدة للنموذج. لم نحصي أي ملاحظة نافذة أو شاذة.

الكلمات الجوهرية *: كحولات و فينولات – التمثيل الرقمي للتركيب الكيميائي – المعامل الغير مركزي– التراجع المتعدد الخطي – النموذج الهجين PSR.*

Abstract

The acentric factors of 18 hydroxy compounds (alcohols, phenols) were linearly correlated with 2 molecular descriptors of geometrical type selected by genetic algorithm, among more than 1600 derived from the molecular modeling software DRAGON. The different statistics calculated (multiple determination and prediction coefficients; roots of the mean quadratic errors; Y-scrambling) show the quality, the robustness and the good internal predictive capacity of the constructed model. No outliers or influential observation was found.

***Key words**: Alcohols and phenols – Numerical representation of chemical structure – Acentric factor – Multiple linear regression – Hybrid SPR model.*

Résumé

Les facteurs acentriques de 18 composés hydroxylés (alcools, phénols), ont été corrélés linéairement avec 2 descripteurs moléculaires de type géométrique sélectionnés par algorithme génétique, parmi plus de 1600 calculés en utilisant le logiciel de modélisation moléculaire DRAGON. Les différentes statistiques établies (cœfficient de détermination multiple et de prédiction ; racines des erreurs quadratiques moyennes ; test de randomisation) montrent la qualité, la robustesse et les bonnes capacités prédictives internes du modèle construit. Aucune observation aberrante ou influente n'a été relevée.

***Mots clés** : Alcools et phénols – Représentation numérique de la structure chimique – Facteur acentrique – Régression linéaire multiple – Modèle RSP hybride.*

[*]*Auteur correspondent : leilalourici@yahoo.fr*

1. INTRODUCTION

Le facteur acentrique ω est un paramètre parmi les plus courants des corps purs. Comme proposé à l'origine par Pitzer [1,2] ω représente la non sphéricité d'une molécule. De ce fait, le facteur acentrique est très utilisé pour la détermination de nombreuses propriétés thermodynamiques (facteur de compressibilité, pression de vapeur, enthalpie de vaporisation, coefficients de l'équation du viriel) et dans les études des équilibres de phases des substances [3-4].
Pour les gaz monoatomiques, ω est essentiellement nul, et pour le méthane sa valeur est encore très petite. Cependant, ω croît avec la masse moléculaire des hydrocarbures, de même qu'avec la polarité.
Si, à présent, ω est très largement utilisé pour caractériser la complexité d'une molécule du point de vue de la géométrie et de la polarité [5], les grandes valeurs du facteur acentrique de certains composés polaires (ω >0,4) ne sont pas significatives dans l'acception originelle de cette propriété.
L'objectif de ce travail vise à utiliser la méthodologie RSP (pour Relation Structure/ Propriété), dans l'approche algorithme génétique/ régression linéaire multiple (AG/RLM), pour relier les facteurs acentriques, compris entre 0,433 et 0,665, d'un ensemble hétérogène d'alcools et de phénols, à des descripteurs moléculaires reflétant certaines particularités des molécules prises en compte. L'interprétation de ces descripteurs permettrait d'avoir un aperçu sur les facteurs vraisemblablement liés aux facteurs acentriques des alcools et phénols considérés.
La qualité de l'ajustement et la robustesse du modèle ont été vérifiées.

2. METHODOLOGIE

2.1 Ensemble de données :

Les facteurs acentriques d'un ensemble hétérogène d'alcools et de phénols ont été prélevés dans la littérature [3]. Ces données (Tab. 1) se rapportent à 13 alcanols à chaînes ouvertes (linéaires ou ramifiées) ou fermées, et 5 dérivés phénoliques ; on y relève plusieurs isomères (chaîne, position).

2.2 Descripteurs moléculaires :

La représentation numérique de la structure chimique (descripteurs moléculaires) est une étape importante de l'investigation RSP. La qualité du modèle élaboré est étroitement liée au mode de détermination de ces descripteurs.
Nous avons utilisé le logiciel de modélisation moléculaire HyperChem 6.03 [6] pour représenter les molécules, puis obtenir les géométries finales à l'aide de la méthode semi-empirique AM1. Tous les calculs ont été exécutés dans le cadre du formalisme de Hartree-Fock avec contrainte de spin (ou RHF, pour Restricted Hartree-Fock) sans interaction de configuration. Les structures moléculaires ont été optimisées à l'aide de l'algorithme Polak-Ribiere avec pour critère une racine du carré moyen du gradient égale à 0,001 kcal.mol^{-1}. Les géométries ainsi optimisées ont été transférées dans le logiciel informatique DRAGON [7] pour le calcul de plus de 1600 descripteurs appartenant à 20 classes différentes.
En utilisant les options correspondantes du logiciel DRAGON, nous avons d'abord éliminé les descripteurs à valeurs constantes (écarts types inférieurs à 0,0001) qui n'apportent aucune information, ensuite ceux qui sont hautement corrélés ($R \geq 0{,}9$) et qui véhiculent une information redondante. Pour chaque paire de descripteurs corrélés, est éliminé automatiquement celui qui présente les plus hautes corrélations croisées avec les autres descripteurs.

2.3 Choix d'un sous-ensemble de descripteurs (VSS, pour Variable Subset Selection) par Algorithme génétique (GA/VSS):

On dispose souvent de plus de descripteurs qu'il n'est nécessaire. Et plutôt que de chercher à expliquer la variable dépendante (facteur acentrique) par tous les régresseurs (descripteurs moléculaires) disponibles, 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, les algorithmes évolutifs et génétiques [8,9]; la comparaison se fait souvent à l'avantage de ces derniers.
La modélisation de processus génétiques a initié le développement des algorithmes génétiques, qui peuvent être exploités dans une grande variété de problèmes d'optimisation [10].
Dans un algorithme génétique adapté à l'optimisation, une solution potentielle est considérée comme un individu dans une population. La valeur de la fonction de coût associée à une solution mesure "l'adaptation" de l'individu associé à son environnement. Un algorithme génétique simule l'évolution,

sur plusieurs générations, d'une population initiale dont les individus sont mal adaptés au moyen d'opérateurs génétiques de reproduction et de mutation. Après un certain nombre de générations, la population est constituée d'individus bien adaptés, autrement dit des solutions supposées "bonnes" au problème d'optimisation.
Dans ce travail la sélection des descripteurs a été réalisée par algorithme génétique dans le logiciel MobyDigs [11], en maximisant le coefficient de prédiction Q^2_{LOO}.

2.4 Modèle de régression multiple :

Par souci de simplicité on utilise la régression linéaire multiple (RLM) qui impose des transformations linéaires dans les relations entre descripteurs et propriétés étudiées.
Un modèle de régression multiple entre une variable expliquée Y et l variables explicatives $X_1, X_2, ..., X_l$, s'écrit pour tout $i = 1,2,...,n$:

$$y_i = \beta_0 + \sum_{j=1}^{l} \beta_j x_{ij} + \varepsilon_i \qquad (1)$$

où les $y_i, x_{i1}, ..., x_{il}$ sont des données respectivement relatives aux variables $Y, X_1, ..., X_l$.
Les estimateurs des coefficients β_j sont calculés en utilisant la méthode des moindres carrés ordinaires. Les variables aléatoires ε_i représentent les termes d'erreur non observables du modèle. On peut estimer ces erreurs par les résidus ordinaires e_i, différences entre les valeurs observées y_i et les valeurs estimées $\hat{y}_i$.
L'estimation par les moindres carrés des coefficients de régression suppose que les données suivent la loi normale, ce qui sera vérifié systématiquement.
Deux paramètres statistiques sont couramment utilisés pour l'évaluation de la qualité du modèle :

- Le coefficient de détermination multiple :

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

où $\bar{y}$ est la valeur moyenne des valeurs observées.

- La racine de l'écart quadratique moyen de prédiction :

$$EQMP = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_{(i)})^2} = \sqrt{\frac{PRESS}{n}} \qquad (3)$$

Il est intéressant de considérer, également, la racine de l'écart quadratique moyen calculé sur l'ensemble de calibrage (EQMC), c'est-à-dire l'ensemble qui a servi à la construction du modèle.

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

La validation croisée par "Leave –One –Out" (LOO) [12] consiste à calculer le modèle sur (n-1) composés, et à utiliser le modèle ainsi obtenu pour calculer le facteur acentrique du composé écarté, noté $\hat{y}_{(i)}$. On répète le procédé pour chacun des n composés. La somme des carrés des erreurs de prédiction, désignée par le symbole PRESS dans l'équation (3), est une mesure de la dispersion des estimations. On l'utilise pour définir le coefficient de prédiction :

$$Q^2_{LOO} = \frac{SCT - PRESS}{SCT} \quad (5)$$

Contrairement à R^2, qui augmente avec le nombre de paramètres du modèle, le facteur Q^2_{LOO} affiche une courbe avec maximum (ou avec palier) obtenu pour un certain nombre de descripteurs, puis décroît de façon monotone. Ce fait confère une grande importance au coefficient Q^2_{LOO} .Une valeur de Q^2_{LOO} > 0,5 est considérée comme satisfaisante, et une valeur supérieure à 0,9 est excellente [13].

3. RESULTATS ET DISCUSSION

3.1 Sélection des descripteurs moléculaires :

L'optimisation par algorithme génétique (GA-VSS) conduit à de nombreux modèles de différentes dimensions. Parmi les modèles sélectionnés nous avons retenu le plus simple à deux variables explicatives (de coefficient de corrélation r= 0,092 pour une valeur de p= 0,716) qui sont des descripteurs moléculaires géométriques : l'autocorrélation à levier pondéré de distance topologique 3/ pondérée par les volumes atomiques de van der Waals v ($\mathrm{HATS3v}$), et la seconde composante de l'indice de taille WHIM dirigé/ pondérée par les polarisabilités p ($\mathrm{L2p}$).

Les descripteurs moléculaires à invariant holistique pondéré (WHIM) [14,15], permettent de saisir dans le détail les informations relatives à la taille, la forme, la symétrie et la distribution des atomes d'une molécule par rapport à des cadres de références fixes. Le calcul des descripteurs WHIM repose sur l'analyse en composantes principales de la matrice de covariance des coordonnées atomiques pondérées, dont les éléments sont définis par :

$$s_{jk} = \frac{\sum_{i=1}^{n} w_i \left(q_{ij} - \overline{q}_j\right)\left(q_{ij} - \overline{q}_k\right)}{\sum_{i=1}^{n} w_i} \quad (6)$$

où n représente le nombre d'atomes de la molécule, w_i , le poid du i^{ème} atome, q_{ij} la j^{ème} coordonnée cartésienne de l'atome i (j=1,2,3) alors que $\overline{q}_j$ est la moyenne de cette j^{ème} coordonnée.

Six modèles de pondération, rapportés à l'échelle de l'atome de carbone, sont proposés, et selon le mode adopté on obtient différentes matrices de covariance et différents axes principaux (c'est-à-dire des composantes $\mathbf{t_m}$, m=1, 2,3) pour la molécule.

On distingue les descripteurs WHIM dirigés, calculés individuellement selon les directions des composantes principales, et les descripteurs WHIM non dirigés, ou globaux, calculés pour la molécule entière à partir des combinaisons des premiers.

Les indices de taille WHIM dirigés, Lkw , sont définis par les valeurs propres $\lambda_k \left(k = 1,2,3\right)$ de la matrice de covariance des coordonnées atomiques pondérées de la molécule. Chaque vecteur propre mesure la dispersion (variance pondérée) des atomes projetés sur l'axe principal considéré, renseignant ainsi sur la dimension de la molécule selon cette direction principale.

Le descripteur moléculaire L2p qui est lié à la dimension des molécules sur le deuxième axe principal, met également en évidence le rôle de la polarisabilité.

Les descripteurs "Assemblage de géométrie, topologie et poids atomiques" GETAWAY (pour GEometry, Topology, and Atom Weights AssempblY) [16,17] sont basés sur les formules d'autocorrélation spatiales, en pondérant les atomes dans les molécules par des propriétés physico-chimiques, et par les informations 3D contenues dans les éléments des matrices influence moléculaires **H** et influence/distances **R** qui en est déduite (par minimisation des interactions entre paires d'atomes trop éloignés). La matrice **H,** elle-même, est définie à partir de la matrice moléculaire **M** des coordonnées cartésiennes x, y, z des atomes (y compris les hydrogènes) prises par rapport au barycentre de la molécule, considérée dans la conformation choisie. Les éléments diagonaux h_{ii} de **H** (ou leviers) renseignent sur "l'influence" de chaque atome de la molécule quant à déterminer la forme

globale de celle-ci; en fait, les atomes périphériques possèdent toujours de plus grands h_{ii} que les atomes voisins du barycentre de la molécule. De plus, l'ampleur du levier maximal d'une molécule dépend de sa grosseur et de sa forme. Notons enfin, ce qui peut être déduit de la géométrie moléculaire, que les valeurs des leviers sont sensibles à des changements conformationnels significatifs, et aux longueurs de liaison qui tiennent compte des types d'atomes et de la multiplicité des liaisons.

Les descripteurs d'autocorrélation à levier pondéré de distance topologique k (=3, dans notre cas), sont calculés à partir de l'équation :

$$HATSkw = \sum_{n=1}^{n_{AT}\text{-}1} \sum_{j>i} \left(w_i h_{ii}\right)\left(w_j h_{jj}\right)\delta\left(k; d_{ij}\right) \qquad (7)$$

$$k = 0,1,2,...,8$$

n_{AT} est le nombre d'atomes de la molécule ; d_{ij} est la distance topologique entre les atomes i et j, c'est-à-dire le nombre de liaisons du chemin le plus court reliant ces deux atomes; w_i est une pondération atomique physico-chimique (volume de van der Waals dans le cas présent);

$\delta\left(k; d_{ij}\right)$ est une fonction delta de Dirac ($\delta = 1$ si $d_{ij} = k$, sinon zéro).

Ils apportent une information sur la position effective, dans l'espace moléculaire, des substituants et des fragments de la molécule. De plus, ils renseignent, jusqu'à un certain point, sur la dimension et la forme moléculaire, ainsi que sur les propriétés atomiques spécifiques.

3.2 Modèle AG/MLR :

Avant de procéder au développement effectif des équations de régression, la qualité statistique des variables dépendante et explicatives a été vérifiée.

Tableau 1 : Valeurs des facteurs acentriques et des descripteurs moléculaires sélectionnés

N° (i)	Nom	ω_i	L2p	HATS3v	h_{ii}	e_{istd}
1	o-Cresol	0,433	1,726	0,118	0,265	0,2882
2	Phenol	0,438	1,481	0,115	0,186	1,3342
3	m-Cresol	0,454	1,516	0,123	0,178	0,5987
4	Pentafluorophenol	0,502	1,539	0,148	0,150	-0,9329
5	p-Cresol	0,505	1,297	0,137	0,101	-0,2194
6	Cyclohexanol	0,528	1,479	0,218	0,230	1,8875
7	Methanol	0,556	0,239	0,092	0,267	1,0597
8	Heptan-1-ol	0,560	0,565	0,107	0,146	-0,5199
9	Hexan-1-ol	0,560	0,541	0,119	0,123	0,2334
10	Butan-2-ol	0,577	0,703	0,177	0,086	1,2230
11	Pentan-1-ol	0,579	0,54	0,134	0,102	-0,0220
12	Octan-1-ol	0,587	0,563	0,100	0,166	-2,4464
13	2-Methylpropan-1-ol	0,592	1,220	0,192	0,112	-1,6401
14	Butan-1-ol	0,593	0,496	0,160	0,105	0,7535
15	2-Methylpropan-2-ol	0,612	1,218	0,237	0,262	-0,6252
16	Propan-1-ol	0,623	0,487	0,189	0,147	0,7080
17	Ethan-1-ol	0,644	0,377	0,203	0,214	1,0411
18	Propan-2-ol	0,665	0,792	0,212	0,161	-2,2746

Les diagrammes de probabilités établis à partir des données du tableau 1 montrent que les variables considérées se distribuent selon la loi normale, puisque les R obtenus sont systématiquement supérieurs aux R critiques (R_C) donnés par les tables pour les niveaux α=1% et α=5% , pour n=18 individus (Tab. 2).

Tableau 2 : Vérification de la loi de Laplace-Gauss pour n=18 individus.

	ω	HATS3v	L2p
R (%)	97,84	97,42	94,99
R_C (%)	94,55 (pour $\alpha = 5\%$) et 92,18 (pour $\alpha = 1\%$)		

Le modèle basé sur les descripteurs sélectionnés a pour équation :

$$\hat{\omega} = 0{,}510(\pm 0{,}022) + 0{,}938(\pm 0{,}124)\,\mathrm{HATS3v} - 0{,}107(\pm 0{,}011)\,\mathrm{L2p} \quad (8)$$

Il vérifie les hypothèses d'un modèle statistique linéaire à effets fixes. En effet la figure 1 reproduit la distribution des résidus normalisés RESN (Rapport : résidus ordinaires/ racine du carré moyen des écarts) en fonction des valeurs ajustées AJUST, qui semble aléatoire (sans tendance particulière), ce qui montre la constance des variances σ^2, c'est-à-dire leur indépendance des régresseurs et de la variable dépendante ajustée.

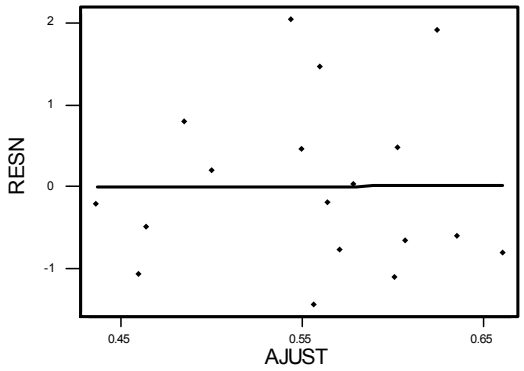


Figure 1: Graphe des résidus normalisés en fonction des facteurs acentriques ajustés.

La quasi-linéarité ($R = 0{,}9675; R_C = 0{,}9455$) du diagramme des scores normaux (Fig. 2) est un indice de normalité. La statistique de Durbin-Watson [18], $d=1{,}85$, est plus grande que la valeur supérieure donnée par les tables pour 2 régresseurs, et pour tout risque raisonnable α, ce qui établit l'indépendance des résidus.

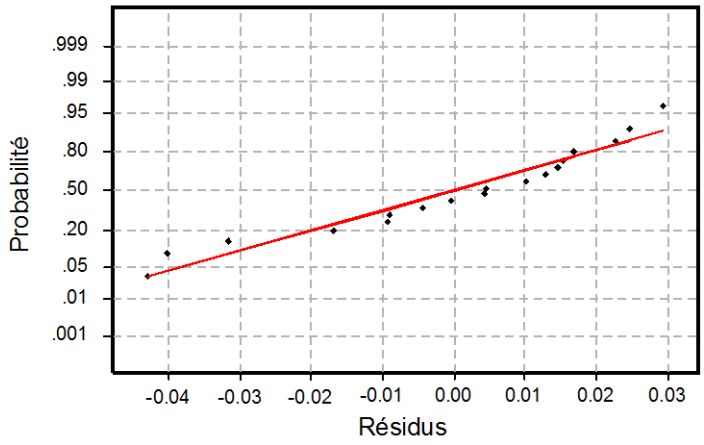


Figure 2 : Diagramme des scores normaux

Les diagnostiques statistiques du modèle sont rapportés ci-après :

$$R^2(\%) = 89{,}83\ ;\ Q^2(\%) = 85{,}35\ ;\ R^2_{adj} = 88{,}47\ ;\ EQMC = 0{,}020\ ;\ EQMP = 0{,}025\ ;$$
$$F = 66{,}24\ ;\ SE = 0{,}023$$

Les valeurs de R^2 et de R^2_{adj} montrent la qualité de l'ajustement, alors que la petite différence entre R^2 et Q^2 renseigne sur la robustesse du modèle qui, en outre, est hautement significatif (grande valeur du paramètre de Fisher F). De plus, la similitude de $EQMC$ et $EQMP$ signifie que la capacité de prédiction interne du modèle n'est pas trop dissemblable de son pouvoir d'ajustement

Les modèles RSP, à cause (souvent) de leur complexité et de la sophistication des outils de chimiométrie employés, peuvent constituer une source de corrélation fortuite. Dans le but d'établir que le modèle obtenu n'est pas dû au hasard, nous avons appliqué le test de randomisation de y . Ce test consiste à générer un vecteur "facteur acentrique" par permutation aléatoire des composantes du vecteur réel. On calcule alors sur le vecteur obtenu (considéré comme vecteur expérimental réel) un modèle RSP, selon la méthode habituelle. Ce procédé est répété plusieurs fois (100 dans notre cas).

La figure 3 qui représente le graphe des coefficients statistiques Q^2 et R^2 permet de comparer les résultats obtenus pour les modèles randomisés (carrés pleins) au modèle de départ (astérisque). Il est clair que les statistiques obtenues pour les vecteurs modifiés des facteurs acentriques sont plus petites (la majorité des valeurs de Q^2 sont même négatives) que celles du modèle RSP réel, ce qui permet d'assurer qu'une relation structure/facteur acentrique réelle a été établie.

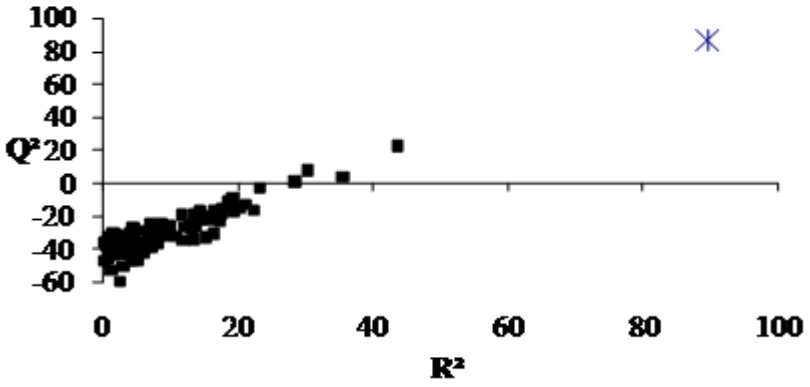


Figure 3 : Test de randomisation associé au modèle RSP

Pour détecter les observations aberrantes nous avons utilisé les résidus de prédiction standardisés [12] :

$$e_{istd} = \frac{e_{(i)}}{\sqrt{s^2_{(i)}\left(1 - h_{ii}\right)}} \tag{9}$$

pour lesquels l'estimation $S^2_{(i)}$ de σ^2 est calculée selon :

$$S^2_{(i)} = \frac{\left[n - \left(l+1\right)\right] CME - e_i^2 / \left(1 - h_{ii}\right)}{n - l - 2} \tag{10}$$

pour $\left(n-1\right)$ observations, la i[eme] étant exclue ; e_i est le résidu ordinaire ; CME est le carré moyen des écarts, et $\left(l+1\right)$ le nombre de paramètres du modèle.

Les valeurs absolues des résidus de prédiction standardisés (tableau 1, dernière colonne) étant toutes inférieures en valeur absolue à 3 unités d'écart type $\left(\left|e_{istd}\right| < 3\sigma\right)$ aucune donnée aberrante n'est ainsi détectée pour le modèle.

Les leviers h_{ii}, éléments diagonaux de la matrice **H** de passage du vecteur y au vecteur $\hat{y}$, permettent de juger de l'influence d'une observation i dans la détermination de l'équation de régression (lorsque h_{ii} est supérieur à la valeur critique $3\left(l+1\right)/n$). Toutes les valeurs reproduites dans la colonne h_{ii} (Tab. 1) étant inférieures à $3 \times 3/18 = 0,5$, aucune observation n'est influente.

Du fait de la différence entre les bases de données modélisées (du point de vue source et nombre de données) d'une part et des complexités des méthodes utilisées d'autre part, la comparaison des erreurs de calcul a été privilégiée. Les erreurs de calcul du facteur acentrique par les méthodes de contribution de groupes (MCG) [3] se distribuent entre 0,04 et 0,07 en unité log. Elles sont supérieures à celles des travaux cités dans le tableau 3 où AAD est la moyenne des valeurs absolues des déviations, et AAD% la valeur relative.

Tableau 3 : Comparaison avec les travaux antérieurs

	n	Méthode	AAD (AAD%)	SE	EQMC
Carande et al. [19]	614	RQSP (RVS)[b]	0,0310 (6,9)	0,023	0,048
Wang et al. [20]	477(48)[a]	MCG	0,0613 (10,39)	-	-
Mokshina et al. [21]	331	RQSP (RF)[c]	0,0140 (-)	0,027	-
Notre travail	18	RSP (AG/RLM)	0,0172 (3,05)	0,023	0,020

[a] 48 alcools pour lesquels les résultats sont rapportés.
[b] Régression par vecteurs supports.
[c] Random Forest pour la sélection des descripteurs.

Si le modèle de Mokshina et al. est le meilleur, il ne reste pas moins difficile à mettre en œuvre.la méthode de calcul des descripteurs n'étant pas automatisée contrairement au modèle bilinéaire que nous présentons dont les variables explicatives sont calculables rapidement par les logiciels disponibles.

4. CONCLUSION

Des logiciels informatiques (DRAGON, HyperChem …) permettent le calcul de nombreux descripteurs moléculaires utilisés pour modéliser une grande variété de propriétés. On trouve dans d'autres (MobyDigs …), du moins partiellement, une série d'outils créés pour la validation des modèles de régression, dont l'utilisation permet de mettre en évidence des situations particulières.

Ainsi, les facteurs acentriques d'un mélange hétérogène de composés hydroxylés (alcools, phénols), dont plusieurs isomères, ont pu être corrélés avec 2 indices structuraux de type géométrique. Le modèle hautement significatif obtenu, dont nous avons pu vérifier les hypothèses de départ, permet de reproduire les facteurs acentriques observés avec une précision moyenne inférieure à 3 % ; il possède une robustesse et une capacité prédictive satisfaisantes. Nous n'avons pas relevé d'observation présentant des valeurs extrêmes des caractéristiques établies, et qui puisse être considérée comme aberrante ou influente.

5. REFERENCES

[1] Pitzer K.S., 1955. The volumetric and thermodynamic properties of fluids. I. Theoretical basis and virial coefficients, Journal of the American Chemical Society, Vol. 77 (13), 3427–3433.

[2] Pitzer K.S., Lippmann D.Z., Curl R.F, Huggins C.M. & Petersen D.E., 1955. The volumetric and thermodynamic properties of fluids. II. Compressibility factor, vapor pressure and entropy of vaporization, Journal of the American Chemical Society, Vol. 77 (13), 3433–3440.

[3] Poling B.E., Prausnitz J.M. &. O'Connell J.P., 2001. The properties of gases & liquids, Fifth Ed., Mc Graw-Hill. 803p.

[4] Gharagheizi F., Eslamimanesh A., Sattari M., Mohammadi A.H. & Richon D., 2015. Computation of the second virial coefficient of chemical compounds using a corresponding states based method. In Advances in Chemistry Research. J. C. Taylor (Eds), Nova Science Publishers, Inc., Vol. 24, 91-112.

[5] Todeschini R., Consonni V., 2008. Handbook of molecular descriptors. R. Mannhold, H. Kubinyi & H. Timmermann, eds, Wiley, VCH. 688p

[6] Hyperchem ^{TM} Release 6.03 for windows, Molecular Modeling System (2000).

[7] Todeschini R., Consonni V. & Pavan M., 2005. DRAGON, Software for the Calculation of Molecular Descriptors. Release 5.3 for windows, Milano.

[8] Leach A.R., 2001. Molecular modelling: principles and applications. Second Ed., Prentice Hall. 744p

[9] Leach A.R & Gillet V.L., 2007. An introduction to chemoinformatics: Revised Ed.; Springer. 274p

[10] Chambers L., 1995. Practical handbook of genetic algorithms: Applications Volume I; CRC Press. 568p.

[11] Todeschini R., Ballabio D., Consonni V., Mauri A., Pavan M., MobyDigs Software for Multilinear Regression Analysis and Variable Subset Selection by Genetic Algorithm. Release 1.0 for Windows, Milano (2004).

[12] Draper N.R. & Smith H., 1998. Applied regression analysis, Third Ed., Wiley series in Probability and Statistics. 736p.

[13] Eriksson L., Jaworska J., Worth A.P., Cronin M.T.D., Mc Dowell R.M. & Gramatica P., 2003. Methods for reliability and uncertainty assessment and for applicability evaluations of classification and regression-based QSARs, Environmental Health Perspectives, Vol. 111 (10), 1361–1375.

[14] Todeschini R., Lasagni M. & Marengo E., 1994. New molecular descriptors for 2D and 3D structures. Theory, Journal of Chemometrics, Vol. 8 (4), 263–272.

[15] Todeschini R. & Gramatica P., 1997. 3D-Modelling and prediction by WHIM descriptors. Part 5. Theory development and chemical meaning of WHIM descriptors, Quantitative Structure-Activity Relationships, Vol. 16, 113-119.

[16] Consonni V., Todeschini R. & Pavan M., 2002. Structure/response correlations and similarity/diversity analysis by GETAWAY descriptors. 1. Theory of the novel 3D molecular descriptors, Journal of Chemical Information and Computer Sciences, Vol. 42 (3), 682–692.

[17] Consonni V., Todeschini R., Pavan M. & Gramatica P., 2002. Structure/response correlations and similarity/diversity analysis by GETAWAY descriptors. 2. Application of the novel 3D molecular descriptors to QSAR/QSPR studies, Journal of Chemical Information and Computer Sciences, Vol. 42 (3), 693–705.

[18] Durbin J. & Watson G.S., 1971, Testing for serial correlation in least squares regression III, Biometrika, Vol. 58 (1), 1-19.

[19] Carande W.H., Kazakov A., Muzny C. & Frenkel M., 2015. Quantitative Structure-Property Relationship Predictions of Critical Properties and Acentric Factors for Pure Compounds, Journal of Chemical & Engineering Data, Vol. 60, 1377–1387.

[20] Wang Q.,Jia Q. & Ma P., 2012. Prediction of the Acentric Factor of Organic Compounds with the Positional Distributive Contribution Method, Journal of Chemical & Engineering Data, Vol. 57, 169–189.

[21] Mokshina E.G., Kuz'min V.E. & Nedostup V.I., 2014. QSPR Modeling of Critical Parameters of Organic Compounds Belonging to Different Classes in Terms of the Simplex Representation of Molecular Structure, Russian Journal of Organic Chemistry, Vol. 50 (3), 314–321.