الجمهورية الجزائرية الديمقراطية الشعبية
République Algérienne Démocratique et Populaire
وزارة التعليم العالي و البحث العلمي
Ministère de l'enseignement supérieur et de la recherche scientifique

UNIVERSITÉ BADJI MOKHTAR
- ANNABA-

جامعة باجي مختار- عنابة

Année / 2018

Faculté des Sciences
Département de Chimie

THÈSE
Présentée pour obtenir le diplôme de Doctorat en Sciences
Option : Chimie Analytique et Environnement

THEME

METHODES STATISTIQUE ET NEURONALE POUR LA PREDICTION DE QUELQUES PROPRIETES PHYSIQUES (Teb, Tc, Pc)

Par : KERTIOU Nour-eddine

Devant le jury

Membres de Jury:

Président :	**Mr. Ahmed DJELLAL**	**Professeur, Université d'Annaba**
Directeur de thèse :	**Mr. Djelloul MESSADI**	**Professeur, Université d'Annaba**
Examinateur :	**Mr. Abdelbaki CHEMAM**	**Professeur, Ecole ESTI d'Annaba**
Examinateur :	**Mr. Noureddine ZENATI**	**MCA, Université de Souk- Ahras**
Examinateur :	**Mr. Brahim HARKATI**	**MCA, Université de Tebessa**
Examinateur :	**Mr. Mohamed Abdessalem DEMES**	**MRA, C.R.B.Ts Constantine**

Dédicace

Je dédie ce modeste travail à :

- *La mémoire de mon père*
- *Ma mère*
- *Ma famille : Fatma, Ahmed, Chaima, Sara, et Yahia*
- *Mes frères et sœurs*
- *Mes amis*
- *Enfin, toute l'équipe du labo LASEA.*

K-Nour-eddine

Remerciements

Avant tout, je remercie Dieu le tout puissant de m'avoir donné la force et la foi et de m'avoir permis d'arriver à ce stade.

La présente étude a été réalisée au laboratoire de Sécurité Environnementale et Alimentaire de l'Université d'ANNABA sous la direction de monsieur le Pr. MESSADI. Aussi, je me permets de lui exprimer ma profonde reconnaissance, pour le bienveillant intérêt qu'il a accordé quant à la réalisation de cette étude.

J'exprime ma profonde et respectueuse gratitude à Monsieur Ahmed Djellal, Professeur à l'Université d'Annaba, qui m'a fait l'honneur d'accepter de présider le jury de cette thèse.

J'adresse à Monsieur Abdelbaki CHEMAM, Professeur à l'Ecole Supérieure des Technologies Industrielles d'Annaba, l'expression de mes sincères remerciements et de mon entière gratitude, pour faire partie du jury.

Mes vifs remerciements vont également à monsieur Noureddine ZENATI Maître de conférences à l'université de Souk Ahras, monsieur Brahim HARKATI Maître de conférences à l'université de Tebessa, et monsieur Mohamed Abdelessalam DEMES Maître de recherche au centre CRBTs Constantine, pour l'honneur qu'ils nous font en acceptant d'examiner notre travail.

C'est avec beaucoup de gratitude enfin que je remercie tous les membres de l'équipe LASEA pour leur soutien leur amitié et leur aide. J'ai eu beaucoup de plaisir à partager de bons moments à leurs côtés.

Enfin, Je tiens à présenter ma reconnaissance et mes remerciements à ma famille, qui est ma source d'inspiration et mon plus grand soutien, en particulier ma femme Fatma.

Résumé

Deux modèles QSPR ont été développés pour la prédiction de la température d'ébullition ainsi que de la température et pression critiques. Les données, concernant 165 paraffines, ont été séparées en deux sous-ensembles disjoints comprenant respectivement 125 éléments pour le calcul et le test (éventuel) du modèle, et 40 éléments pour sa validation externe. Deux modèles ont ainsi été créés sur le même ensemble de données: un modèle de régression multilinéaire et un modèle de réseaux de neurones artificiels.

Des descripteurs moléculaires théoriques ont été calculés en utilisant des logiciels de modélisation moléculaire du commerce. La sélection des descripteurs réalisée par algorithme génétique et la taille du modèle à été déterminée en utilisant la variation de coefficient de détermination R^2 en fonction de la variation de descripteurs (point de brisure).

Les valeurs des paramètres statistiques (R^2, Q^2, SDEC, SDEP, SDEPext) obtenues attestent de la pertinence des modèles développés, avec une supériorité établie le modèle non linéaire (RNA).

Mots-clés:

Paraffines – Température d'ébullition et grandeurs critiques – Descripteurs moléculaires théoriques – Modèles hybrides.

Abstract

Two models QSAR were developed for the prediction of the boiling point as well as critical temperature and pressure. The data, concerning 165 paraffin's, were separate in two disjoined subsets including/understanding respectively 125 elements for calculates and the possible test of the models, and 40 elements for its external validation. Two models were thus created on the same whole of data: a multilinear model of regression and a model of artificial neural networks.

Theoretical molecular descriptors were calculated by using software of molecular modeling of the trade. The best model and the number of descriptors in the final QSPR model was determined on the basis of the correlation coefficient R^2 (breaking point), and the selection of the descriptors realized by genetic algorithm.

Values of the statistical parameter (R^2, Q^2, SDEC, SDEP, SDEPext) obtained attest relevance of the models developed, with a superiority established for the models of artificial neurons.

Key words:

Paraffin's - boiling point – sizes criticize – molecular descriptors theoretical – hybrid Models.

ملـخص:

تم تطوير نموذجين بطريقة الـ QSPR لتنبؤ درجة الغليان، و كذلك الدرجة و الضغط الحرجين .

المعطيات الخاصة بـ 165 مركب برافيني (paraffines) تم تقسيمها الى مجموعتين الاولى تحتوي على 125 عنصر لحساب و تجريب النموذج اما الثانية تحتوي على 40 عنصر لتصديق الخارجي للنموذج.

النموذجين المتحصل عليهما لنفس المعطيات هما : نموذج التراجع المتعدد الخطي و نموذج الشبكة العصبونية الاصطناعية.

الموصفات الجزيئية النظرية تم حسابها باستعمال برمجيات النمذجة الجزيئية المتوفرة في السوق . حجم النموذج تم تحديده عن طريق تغير معامل الاثبات بدلالة تغير حجم النموذج (النقطة الفاصلة)، اما اختيار المواصفات عن طريق الخوارزمية المورثية .

قيم المعالم الاحصائية ($SDEP_{ext}$, SDEP , SDEC , Q^2, R^2) المتحصل عليها تؤكد تعلق النماذج المطورة مع تفوق معتبر لنموذج الشبكة العصبونية الاصطناعي.

الكلمات الدالة:

برافين (paraffines) – درجة الغليان و قيم حرجة – مواصفات جزيئية نظرية – نماذج مهجنة.

Sommaire

PARTIE APPLICATION

SYMBOLES ET ABREVIATIONS

ACP: Analyse en composantes principales.

AM1 : Austin Model 1.

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

EQMP Ecart quadratique moyen de prédiction.

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

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

eistd : Résidus standartisés.

F : Statistique de Fisher.

FIV: Facteur d'inflation de la variance.

GA: Algorithme génétique (Genetic Algorithm).

hii : Eléments diagonaux de la matrice chapeau.

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

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

n: Dimension de la population (échantillon).

PLS(ou MCP): Moindres carrés partiels.

PRESS : Somme des carrés des erreurs de prédiction.

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

QSAR : Quantitative Structure/ Activity Relationships.
(Relations Quantitatives Structure/ Activité).

QSPR : Quantitative Structure/ Propriety Relationships.
(Relations Quantitatives Structure/ Propriété).

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

R^2 : Coefficient de détermination.

RLM (MLR): Régression linéaire multiple.

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

RNA: Réseaux de neurones artificiels.

S : Erreur standard.

SCE : Somme des carrés des écarts.

SCT : Somme des carrés totale.

t : t de Student.

t_i : Résidu studentisé externe.

yi: Valeur observée.

$\hat{y}_i$: Valeur estimée.

$\hat{y}_{(i)}$: Valeur prédite.

Liste des tableaux

Liste des figures

INTRODUCTION GENERALE

İNTRODUCTION GÉNÉRALE

Les n-praffines à haut poids moléculaire sont présentes dans les pétroles, et les n-paraffines pures sont utiles comme composés modèles pour le développement et le test de corrélations servant à la prédiction des propriétés des fractions de pétrole. Les n-praffines, les n-oléfines et les n-alcools sont tous présents dans les huiles synthétiques obtenues par la réaction de Fischer et Tropsch. Le développement ultérieur de ce processus technologique nécessite la connaissance des propriétés de ces fluides, ou leur estimation avec une précision raisonnable. En outre, les propriétés des n-prarffines et de leurs dérivés sont importantes à connaître dans les industries de fabrication des cires à base de paraffines, et de différents polymères linéaires.

La température d'ébullition d'un composé reflète les interactions entre molécules du liquide, ainsi que la différence entre les fonctions de partition moléculaires internes dans le liquide et le gaz obtenue à l'ébullition.

La température d'ébullition peut être mesurée facilement, et sa prédiction reste d'une portée limitée. Cependant, c'est la propriété physico-chimique la plus utilisée dans les exercices de modélisation par réseaux de neurones artificiels (RNA).

La température critique est la température au-dessus de laquelle un gaz ne peut être liquéfié, ce qui signifie également qu'au-dessus de la température critique une substance ne peut se présenter distinctement sous les phases gazeuse et liquide. Pareillement, la pression critique est définie comme la pression la plus élevée à laquelle les phases gazeuse et liquide, d'une substance donnée, peuvent encore co-exister.

Températures et pressions critiques sont nécessaires dans les calculs thermodynamiques, et importantes pour les ingénieurs de l'industrie chimique. Elles sont également importantes dans de nombreux domaines de la recherche pharmaceutique, comme par exemple l'ingénierie des cristaux qui utilise les fluides supercritiques.

Les méthodes de contributions de groupes (GCM) peuvent être appliquées pour la prédiction des propriétés de composés homologues. En général, ces méthodes sont inefficaces

quand on extrapole à des nombres d'atomes de carbone élevés. Les équations théoriques, qui présentent un comportement asymptotique correct, fournissent des résultats médiocres concernant les composés à petits nombres d'atomes de carbone.

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

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

Des logiciels informatiques spécialisés permettent le calcul de plus de 10 000 descripteurs moléculaires appartenant à différentes classes. Plutôt que de rechercher à expliquer la variable dépendante (propriété physique considérée) par tous les régresseurs (descripteurs moléculaires), on peut chercher seulement un ensemble réduit de régresseurs qui donne une reconstitution aussi satisfaisante de la variable à expliquer. Parmi les stratégies mises en œuvre pour la sélection d'un ensemble limité de variables explicatives, on peut citer : les méthodes de pas-à-pas, ainsi que les algorithmes évolutifs et génétiques.

Nous avons appliqué des méthodes hybrides: algorithme génétique/régression multilinéaire (GA/MLR), et algorithme génétique/réseaux de neurones artificiels (GA/RNA) pour modéliser, séparément, la température d'ébullition, la température et la pression critiques de 165 paraffines industriellement importantes dans la perspective du génie chimique.

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

Dans la Partie Généralités, nous avons développé tout ce qui a trait au pré-traitement des molécules (introduction des molécules, optimisation de leur géométrie) en vue du calcul des

descripteurs moléculaires à l'aide de différents logiciels de modélisation moléculaire. Nous y avons également développé les connaissances théoriques de base utilisées tout au long de ce travail: algorithmes génétiques, régression multilinéaire, réseaux de neurones artificiels, traitement statistique pour l'évaluation de la qualité de l'ajustement (robustesse des modèles; détection des observations aberrantes; test de randomisation; validation externe).

Dans la Partie Application, nous présentons et discutons les modèles calculés :

GA/MLR et GA/RNA, pour la température critique;

GA/MLR et GA/RNA, pour la température d'ébullition;

Et enfin, GA/MLR et GA/RNA, pour la pression critique.

PARTİE THÉORİQUE

I-LA MODÉLİSATİON MOLÉCULAİRE

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

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

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

II-OPTİMİSATİON DE LA GÉOMETRİE DES MOLÉCULES

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

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

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

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

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

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

$$\psi_m(n)=\varphi_m(n)\alpha(n)$$
$$\bar{\psi}_m(n)=\varphi_m(n)\beta(n) \tag{2}$$

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

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

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

$$H(1,2,\dots,n)=\sum_{i=1}^{n}h_{(i)}^{c}+\sum_{i<j}\frac{e^2}{r_{ij}} \tag{3}$$

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

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

$$E=<\psi|H|\psi> \tag{4}$$

Lorsqu'on développe cette intégrale on arrive [1] au résultat :

$$E=\sum_{i=1}^{m}2h_{ii}^{c}+\sum_{i=1}^{m}\sum_{j=1}^{m}(2J_{ij}-K_{ij}) \tag{5}$$

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

$$\boldsymbol{h_{ii}^{c}=<\psi_i(\mu)\left|h_{(\mu)}^{c}\right|\psi_i(\mu)>} \tag{6}$$

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

$$\boldsymbol{J_{ij}=\iint\psi_i^{*}(\mu)\psi_i(\mu)\frac{e^2}{r_{\mu\nu}}\psi_j^{*}(\nu)\psi_j(\nu)d\tau_\mu d\tau_\nu} \tag{7}$$

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

$$\boldsymbol{K_{ij} = \iint \psi_i^*(\mu)\psi_i^*(\nu)\frac{e^2}{r_{\mu\nu}}\psi_j(\mu)\psi_j(\nu)d\tau_\mu d\tau_\nu} \quad (8)$$

est **l'intégrale biélectronique moléculaire d'échange** ; elle représente également une somme de répulsions entre charges élémentaires, l'électron occupant deux orbitales moléculaires ψ_i et ψ_j. $r_{\mu\nu}$ représente la distance entre les deux électron μ et ν.

Remarques :

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

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

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

$$E=E^c+E^{RE} \quad (9)$$

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

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

$$E_T = E + \sum_{N<L}\frac{Z_K E_L e^2}{R_{KL}} \quad (10)$$

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

La relation (5) est équivalente à :

$$E = \sum_{i=1}^{m}\{h_{ii}^c + [h_{ii}^c + \sum_{j=1}^{m}(2J_{ij} + K_{ij}]\} \quad (5)$$

Le terme :

$$e_i = h_{ii}^c + \sum_{j=1}^{m}(2J_{ij} + K_{ij}) \quad (11)$$

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

E se réduit donc à :

$$E = \sum_{i=1}^{m}(h_{ii}^c + e_i) \quad (12)$$

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

$$E = \sum_{i=1}^{m} 2\, e_i \quad (13)$$

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

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

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

II-1-2. Détermination des Orbitales ou équations de Hartree-Fock

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

Quelles propriétés doivent posséder les ψ_i pour être acceptables au sens de la mécanique ondulatoire, et qu'elles puissent s'adapter au système particulier envisagé ?
Il faut que le déterminant de Slater soit une solution approchée de l'équation de Schrödinger totale :

$$H(1,2,\dots.,n)\psi(1,2,\dots.,n) = E\psi(1,2,\dots.,n) \quad (14)$$

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

Comme la variation du déterminant de Slater s'exprime par la variation du jeu des $\{\psi_i\}$, il faudrait avoir, pour une variation première du jeu d'orbitales choisies, une variation première de l'énergie totale nulle, et pour cela il faut que les ψ_i soient solutions des équations de Hartree-Fock [2-4]:

$$\{\delta\psi_i\} \rightarrow \delta E^1 = 0 \quad (15)$$

Ces deux conditions contiennent les équations de Hartree-Fock :

$$F_{(\mu)}\psi_i(\mu) = e_i\psi_i(\mu) \quad (16)$$

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

$$\boldsymbol{F_{(\mu)} = \left[h^c_{(\mu)} + \sum_{i=1}^{m} 2J_l(\mu) - K_l(\mu)\right]} \quad (17)$$

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

II-1-3. Equations de Roothaan et Hall

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

Chaque orbitale moléculaire psi se présentera sous la forme :

$$\psi_i(\mu) = \sum_{p=1}^{N} C_{pi}\varphi_p(\mu) \tag{18}$$

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

Les équations de Hartree-Fock prennent, en tenant compte de (18), une expression vectorielle assez simple :

$$\sum_{p=1}^{N} C_{pi}\left[F_{pq} - e_i S_{pq}\right] = 0 \qquad , \qquad q \in [1, N] \tag{19}$$

Les coefficients :

$$\boldsymbol{S_{pq}} = \int \boldsymbol{\varphi_p^*\, \varphi_q d\tau} \tag{20}$$

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

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

L'équation (19) est un système linéaire homogène (*N* équations à *N* inconnues) qu'on peut écrire sous la forme matricielle :

$$[\boldsymbol{F} - \boldsymbol{e_i S}]C_i = \boldsymbol{0} \tag{21}$$

où $\boldsymbol{F}$ est la matrice $[F_{pq}]$; $\boldsymbol{S}$ est la matrice $[S_{pq}]$; $\boldsymbol{C}$ est la matrice $[C_{pi}]$.

$$F_{pq} = h_{pq}^{c} + \sum_{l=1}^{N}\sum_{m=1}^{N} p_{lm}[< pq \left| lm > - \frac{1}{2} < pm \right| lq >] \tag{22}$$

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

$$h_{pq}^{c} = \left\langle \varphi_p(\mu) \middle| h_{(\mu)}^{c} \middle| \varphi_q(\mu) \right\rangle \tag{23}$$

-* $\langle pq|lm\rangle = \iint \varphi_p(\mu)\varphi_q(\mu)\frac{e^2}{r_{\mu\nu}}\varphi_l(\nu)\varphi_m(\nu)d\tau_\mu d\tau_\nu$ (24)

$\langle pm|lq\rangle = \int \varphi_p(\mu)\varphi_m(\mu)\frac{e^2}{r_{\mu\nu}}\varphi_l(\nu)\varphi_q(\nu)d\tau_\mu d\tau_\nu$ (25)

-* $p_{lm} = \sum_{i=1}^{N} 2C_{li}C_{mi}$ =éléments de la matrice densité (26)

-* $\boldsymbol{P} = [p_{lm}]$ = matrice densité (27)

II-1-4. Quelques remarques sur les processus de résolution des équations de Hartree-Fock-Roothaan.

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

$$\boldsymbol{F\ C_i = e_i\ S\ C_i} \quad (21)$$

Löwdin [5] a proposé un procédé qui permet de se ramener dans tous les cas au calcul des valeurs propres et vecteurs propres d'une matrice moyennant une transformation de la base des orbitales atomiques (**orthogonalisation de Löwdin**).

Multiplions à gauche les deux membres de (21) par la matrice $\mathbf{S}^{-1/2}$, qui n'est jamais singulière puisque **S** ne l'est pas ; il vient successivement:

$$\boldsymbol{S^{-1/2}\ F\ C_i = e_i\ S^{-1/2} S\ C_i}$$

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

Soit en posant :

$$\boldsymbol{S^{-1/2}\ F\ S^{-1/2} = \bar{F}} \quad \text{et} \quad \boldsymbol{S^{-1/2}\ C_i = \bar{C}_i} \quad (28)$$

$$\boldsymbol{\bar{F}\ \bar{C}_i = e_i\ \bar{C}_i}\ , \text{ c'est-à-dire } [\boldsymbol{\bar{F} - e_i\ I}\ ]\boldsymbol{\bar{C}_i = 0} \quad (29)$$

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

$$\boldsymbol{\bar{F}\ \bar{C} = e_i\ \bar{C}_i} \quad (29)$$

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

II-1-5. Détermination des intégrales de la méthode de Hartree-Fock-Roothaan (HFR)

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

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

$$h_{pq}^{c} = < \varphi_p(\mu) \left| h_{\mu}^{c} \right| \varphi_q(\mu) > \quad (30)$$

Il existe deux types d'intégrales de ce genre : **monocentres** lorsque $\varphi_p$et $\varphi_q$appartiennent au même atome R ; **bicentres**, lorsque $\varphi_p$et $\varphi_q$appartiennent à des atomes différents.

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

$$h_{pq}^{c} = -\frac{\hbar^2}{2m}\int \varphi_p(\mu)\Delta_{(\mu)}\varphi_q(\mu)d\tau_\mu - \sum_k Z_K \int \varphi_p(\mu)\frac{e^2}{r_{k\mu}}\varphi_q(\mu)d\tau_\mu \quad (31)$$

$\overleftarrow{Intégrales\ cinétiques}$ $\overleftarrow{intégrales\ d'attractions\ nucléaires}$

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

*- **Intégrales bi-électroniques**

$$G_{pq} = \sum_l \sum_m p_{lm}\left[\langle pq|lm\rangle - \frac{1}{2}\langle pm|lq\rangle\right] \quad (32)$$

$\langle pq|lm\rangle$, $\langle pm|lq\rangle$ et p_{lm} sont respectivement définis par les relations (24), (25) et (26).

On a plusieurs types d'intégrales :

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

Le calcul des intégrales biélectroniques prend le plus grand temps, et il n'est pas possible, en prenant des orbitales de Slater (33) d'en donner des expressions analytiques.

$$\varphi_{n,l,m}(k,\vec{r}) = Nr^{n-1}e^{-kr}y_{l,m}(\theta,\varphi) \quad (33)$$

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

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

II-2. Les méthodes semi-empiriques

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

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

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

II-2-1. Définition du semi-empirisme

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

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

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

II-2-2. Quelques théories semi-empiriques

La première théorie semi-empirique, ou théorie de Pople-Pariser-Parr (PPP), introduite en 1953 par Pariser et Parr [6, 7], et utilisée la même année par Pople [8], permet d'étudier les systèmes conjugués sans tenir compte du squelette σ.

La première théorie des orbitales moléculaires semi-empirique tri-dimensionnelle est **l'approximation au recouvrement différentiel nul (CNDO pour : Complete Neglect of Differential Overlap)**, introduite par Pople, Santry et Segal [9], pour être appliquée à tous les électrons de valence de molécules quelconques organiques ou minérales.

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

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

et - Théorie de Hartree- Fock (HF).

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

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

Cependant, en 1965, les ressources computationnelles nécessaires pour l'approche SCF complète n'étaient pas encore disponibles. La pratique des théories des orbitales moléculaires nécessitaient donc encore des approximations. Le principal problème réside dans le calcul et le stockage des intégrales tétracentres notées $\langle \mu\nu|\lambda\sigma\rangle$, nécessaires pour le calcul des interactions électron- électron dans le cadre de l'approximation SCF. Les indices $\mu, \nu, \lambda \; et \; \sigma$ dénotent quatre centres d'orbitales atomiques de sorte que le nombre de telles orbitales à calculer croît proportionnellement à N^4, où N est le nombre d'orbitales atomiques. En fait, le nombre de telles intégrales n'est pas exactement égal à la puissance quatrième du nombre de fonctions de base parce que beaucoup d'entr'elles sont reliées par symétrie. Ce qui était une tâche très difficile en 1965 ; ainsi Pople, Santry et Segal ont introduit [9] l'approximation que seules les intégrales pour lesquelles $\mu = \nu$ et $\lambda = \sigma$ c'est-à-dire : $\langle \mu\mu|\nu\nu\rangle$ seront prises en compte et que, de plus, toutes les orbitales atomiques seront traitées de la même façon (comme si elles étaient des orbitales s), de sorte que l'équation (34) s'applique, où μ est centrée sur l'atome A et λ sur l'atome B et ainsi γ_{AB} ne dépend que des identités de A et B, et peut être traité comme paramètre.

$$\langle \mu\mu|\lambda\lambda\rangle = \gamma_{AB} \tag{34}$$

Une première approximation, due à Pariser et Parr [6, 7] consiste à traiter le terme mono-centre γ_{AA}comme différence entre le potentiel d'ionisation PI_A et l'affinité électronique AE_A de A [Eq.(35)] :

$$\gamma_{AA} = PI_A - AE_A \tag{35}$$

Les termes di-centres sont alors données par l'éq.(36) :

$$\gamma_{AB} = \frac{\gamma_{AA}+\gamma_{BB}}{2+r_{AB}(\gamma_{AA}+\gamma_{BB})} \quad (36)$$

Ce qui conduit à : $\gamma_{AB} = (\gamma_{AA} + \gamma_{BB})/2$ pour une distance interatomique, r_{AB}, nulle et $\gamma_{AB} \approx {}^{1}/_{r_{AB}}$ pour des distances interatomiques plus grandes. Ces expressions (Eqs. (34) –(36)) montrent la simplicité de la technique CNDO, qui a été utilisée pour calculer les propriétés électroniques comme les moments dipolaires ou les énergies d'excitation, généralement à partir des géométries expérimentales. Il ya eu beaucoup de modifications des eqs.(35 et (36), mais elles restent d'une simplicité comparable. Pareillement, des expressions simplifiées ont aussi été utilisées pour les intégrales mono-électroniques.

Cependant, la méthode CNDO montra des insuffisances systématiques directement imputées aux simplifications ébauchées précédemment, aussi fut-elle remplacée par la méthode **INDO (Intermediale Neglect of Differential Overlap)**, introduite en 1967 par Pople, Beveridge et Dobosh [10]. L'approximation qui conduit à l'éq. (34) s'étant avérée très sévère, elle fut remplacée par des valeurs individuelles pour les différents types d'interactions entre deux orbitales atomiques. Ces valeurs individuelles, souvent désignées par G_{ss}, G_{sp}, G_{pp} et G^2_{pp} dans la littérature, peuvent être ajustées pour donner un accord avec l'expérience meilleur que celui obtenu avec la méthode CNDO. Cependant, en INDO les termes di-centres sont maintenus du même type que ceux apparaissant dans les éqs. (35) et (36). Cette approximation conduit à des affaiblissements systématiques, comme par exemple dans le traitement des interactions entre doublets isolés.

Pour surmonter ces carences, Pople et collaborateurs revinrent à une approche plus complète que celle qu'ils proposèrent initialement en 1965 [9] : **l'approximation au recouvrement différentiel diatomique nul (NDDO : Neglect of Diatomic Differential Overlap)**.

Dans la NDDO, toutes les intégrales tétracentres insuffisances $\langle \mu\nu|\lambda\sigma \rangle$ dans lesquelles μ et ν sont sur le même centre, comme le sont λ et σ (mais pas nécessairement sur le même comme le sont μ et ν) sont prises en compte. De plus, les intégrales pour lesquelles les deux centres atomiques sont différents sont traitées de manière analogue que les intégrales mono-centres en INDO, entrainant, une amélioration de la description des interactions (doublet isolé)-(doublet isolé) par rapport aux méthodes précédentes. La NDDO forme la base de presque toutes les autres

méthodes semi-empiriques qui, à quelques exceptions ont été développées par MJS Dewar et son école.

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

Les méthodes MINDO sont maintenant largement obsolètes.

La méthode avantageuse pour la plupart des techniques modernes d'orbitales moléculaires semi-empiriques est la MNDO, qui a été publiée par Dewar et Thiel en 1977 [11]. La MNDO est une méthode NDDO dans laquelle Dewar et Thiel ont introduit un formalisme basé sur les multipoles pour le calcul des intégrales bi-électroniques. Elle a été paramétrée pour reproduire les chaleurs de formation expérimentales, les géométries, les moments dipolaires et les potentiels d'ionisation. Elle s'avéra très supérieure aux méthodes MINDO pour la plupart des grandeurs calculées. Cependant la MNDO présente une faiblesse qui limite sévèrement son utilité ; elle ne reproduit pas la liaison hydrogène. Cette faiblesse a été surmontée de façon pragmatique par Burstein et Isaev [12] qui modifièrent simplement le potentiel de répulsion cœur-cœur par addition de fonctions gaussiennes en vue d'obtenir des liaisons hydrogène. Ce « fixe » a été adopté par le groupe Dewar pour leur méthode suivante AM1 [13] qui est par ailleurs identique à la MNDO. AM1, en retour, s'avéra présenter une faiblesse dans le traitement des composés nitrosés et hypervalents. Ces faiblesses ont été abordées par Stewart dans une nouvelle paramérisation nommée PM3 [14], qui est par ailleurs identiques à AM1. Cependant, MNDO, MNDO/H, AM1 et PM3 sont pour l'essentiel identiques du point de vue quanto-mécanique. Leurs différences se limitent à la « correction » classique des potentiels entre atomes et pour laquelle les paramètres sont traités comme variables dans la procédure de paramétrisation.

II-2-3. Limites et avantages des méthodes semi-empiriques [15]

La négligence de toutes les intégrales bi-électroniques tri et tétracentres réduit la matrice de Fock d'un ordre formel M^4 à M^2. Toutefois, le temps requis pour la diagonalisation de la matrice ***F*** croît comme le cube de la dimension de la matrice. La diagonalisation d'une matrice devient importante lorsque la dimension dépasse ~ 10 000 × 10 000. De nombreuses itérations sont nécessaires pour la résolution des équations SCF, et habituellement la géométrie est également optimisée, nécessitant de nombreux calculs pour différentes géométries. Ce qui situe la

limite actuelle des méthodes semi-empiriques à environ 1000 atomes. Il est à noter que la méthode classique de résolution des équations HF par diagonalisation de la matrice de Fock s'impose rapidement comme l'étape limitante réelle dans les méthodes semi-empiriques. Des développements récents se sont ainsi focalisés sur la formulation de méthodes alternatives pour l'obtention d'orbitales SCF sans passer par la diagonalisation [16, 17]. De telles méthodes utilisent des ajustements (combinaisons) linéaires avec le nombre d'atomes, ce qui permet d'effectuer des calculs pour des systèmes comprenant plusieurs milliers d'atomes.

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

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

Remarque : l'interaction de configuration CI résoud le problème de la corrélation électronique en considérant plus d'un schéma d'occupation des orbitales moléculaires (OM) et en combinant les micro-états obtenus par permutation des positions électroniques sur toutes les OM disponibles. Dans sa forme la plus simple, un calcul CI consiste en un calcul SCF préliminaire qui fournit les OM qui seront utilisées telles quelles tout au long du reste du traitement. Des micro-états sont alors construits en déplaçant les électrons des orbitales occupées à celles vacantes selon des schémas pré-établis. La matrice CI est alors calculée, ses éléments diagonaux représentent les énergies des micro-états et les éléments non diagonaux leurs interactions. Cette

matrice est diagonalisée en vue d'obtenir les énergies des différents états (fondamental et excités) de la molécule comme combinaisons linéaires des micro-états. De nouveau les énergies sont fournies par les valeurs propres et les coefficients de la combinaison linéaire par les vecteurs propres. Cette procédure conduit à la stabilisation de l'état fondamental, et fournit également les énergies et les fonctions d'onde des états excités. Le problème est que si l'on doit considérer chacun des arrangements possibles de tous les électrons dans toutes les OM (CI complète), les calculs deviennent par trop importants même pour des molécules de taille moyenne avec un ensemble de base pas trop important (parce qu'il y a de trop nombreuses orbitales virtuelles).

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

La forme la plus économique est celle pour laquelle seuls les micro-états dans lesquels un électron est promu de l'état fondamental à une orbitale virtuelle (excitations simples) sont utilisées. Ce qu'on désigne, dans une forme abrégée, par CIS. En ajoutant toutes les excitations doubles (pour lesquels deux électrons sont promus) on est conduit à CISD, et ainsi de suite (Figure 1).

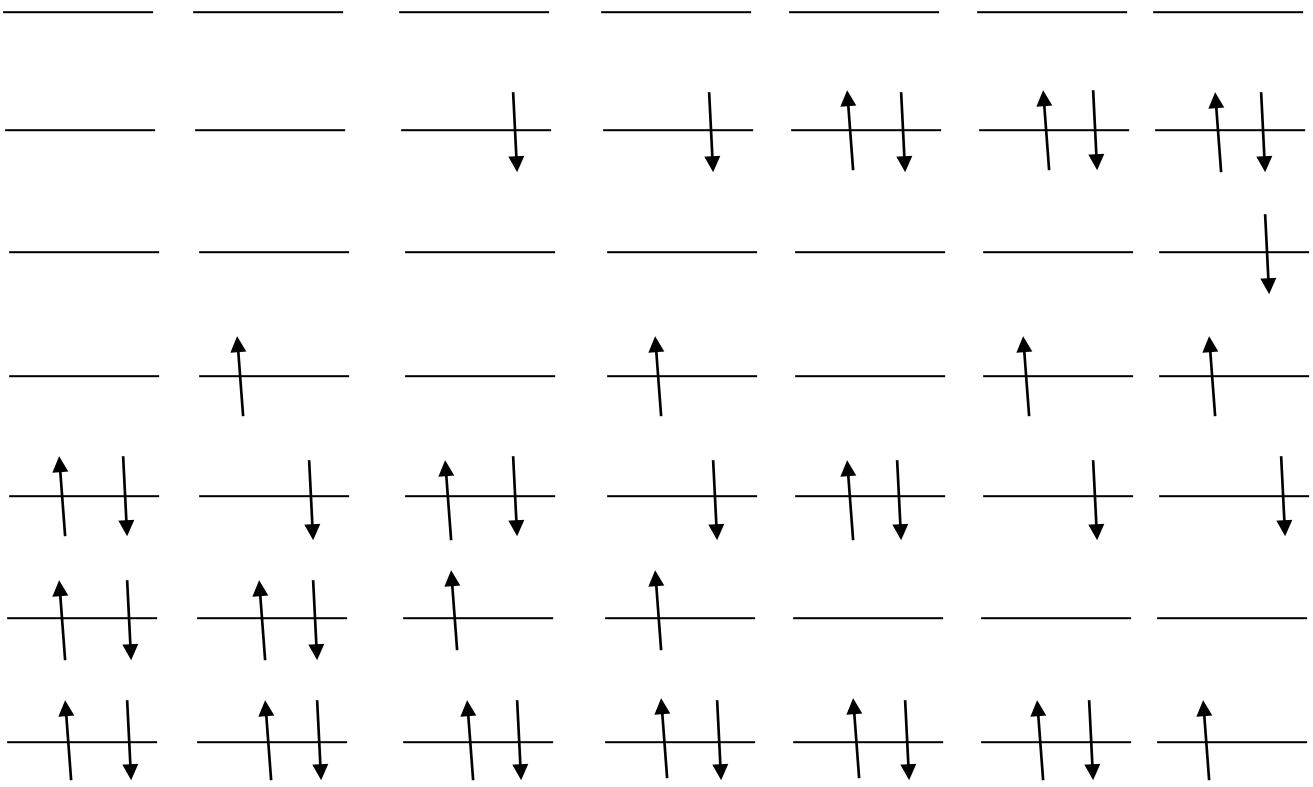


Figure 1. Déterminants de Slater excités générés à partir d'une référence HF

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

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

$$\psi_{CI} = a_0\phi_{SCF} + \sum_{Simples\ (S)} a_s\phi_s + \sum_{Doubles\ (D)} a_D\phi_D + \cdots = \sum_{i=0} a_i\phi_i \qquad (37)$$

La méthode des multiplicateurs indéterminés de Lagrange [18] est ensuite appliquée pour minimiser l'énergie :

$$E = (\langle\psi|\hat{H}|\psi\rangle/\langle\psi|\psi\rangle) \qquad (38)$$

Les méthodes semi-empiriques partagent les avantages/désavantages des méthodes de champ de force (cf : III), elles sont davantage performantes avec les systèmes pour lesquels on dispose de données expérimentales en quantités, mais il leur est impossible de faire des prédictions pour des types de composés totalement inconnus. La dépendance des données

expérimentales n'est pas aussi sévère que pour la méthode du champ de force, à cause de la forme complexe de la fonctionnelle du modèle. Les méthodes NDDO nécessitent uniquement des paramètres atomiques, et nullement des paramètres di-, tri- et tétra-atomiques comme dans les méthodes de champ de force. Une fois un atome donné paramétré, tous les types de composés possibles contenant cet élément peuvent être traités. Le plus petit nombre de paramètres et la forme plus complexe de la fonctionnelle ont l'inconvénient, par rapport aux méthodes de champ de force, qu'il est très difficile de « réparer » un problème spécifique par re-paramétrisation.

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

Les méthodes semi-empiriques fournissent une méthode de calcul de la fonction d'onde électronique, qui peut être utilisée pour la prévision d'une variété de propriétés. Il n'y a rien qui entrave le calcul, par exemple, de la polarisabilité d'une molécule, bien qu'il soit connu des calculs *ab-initio* que l'obtention de bons résultats nécessite un grand ensemble de base polarisé incluant des fonctions diffuses. Les méthodes semi-empiriques comme AM1 ou PM3 n'ont qu'une base minimale (absence de polarisation et de fonctions diffuses), la corrélation électronique n'est qu'implicitement incluse par les paramètres et aucune donnée de polarisabilité n'a été utilisée pour dériver ces paramètres. Il est douteux que de tels calculs puissent conduire à des résultats comparables à ceux fournis par l'expérience, et ils nécessitent, pour le moins, un calibrage soigné [15]. Encore une fois, il convient de souligner que la capacité d'effectuer un calcul ne garantit pas la fiabilité des résultats obtenus.

II-3. Analyse des distributions de charges

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

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

Il existe une famille d'analyses de population, parmi lesquelles nous citerons celles de Coulson et Longuet-Higgins [19], exprimée en termes de charges (ou « densités de charge ») et d'ordres de liaison, celle de Mulliken [20], que nous rappellerons brièvement, et qui fait intervenir les populations atomiques et de recouvrement.

II-3-1. Analyse de population de Mulliken

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

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

$$P_\nu = \sum_k^{OM.occupées} N_k C_{k\nu}^* C_{k\nu} \qquad (39)$$

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

$$R_{\mu\nu} = 2\sum_k^{OM.occupées} C_{k\mu}^* C_{k\nu} S_{\mu\nu} \qquad (40)$$

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

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

$$\sum_\mu R_{\mu\nu} = \sum_\mu \sum_\nu P_{\mu\nu} S_{\mu\nu} = N \quad \text{[Décomposition sur les OA]} \qquad (41)$$

$$\int \psi^* \psi \, d\tau = N \quad \text{[Décomposition sur les OM]} \qquad (42)$$

Posons :

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

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

$$Q_M = \sum_{\mu(M)} q_\mu\,(M) \qquad (43)$$

q_μ=densité électronique de l'orbitale μ ;

Q_M= densité électronique de l'atome M.

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

$$\delta_M = Z_M - Q_M \qquad (44)$$

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

II-3-2. Calcul du moment dipolaire

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

Dans ce schéma, l'expression en u.a du moment dipolaire d'une molécule, dans la convention des chimistes, est [21]

$$\vec{\mu} = \sum_P \sum_Q \sum_{r\in P} \sum_{s\in Q} P_{rs}^{PQ} \int \varphi_r^* \vec{r}\, \varphi_s d_s d_r - \vec{\mu}_{nucl} \qquad (45)$$

Avec :

$$P_{rs}^{PQ} = \sum_i n_i C_{ir} C_{is} \qquad (46)$$

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

On peut alors faire les substitutions suivantes :

$\vec{r}=\vec{r}_p+\vec{\xi}$, dans les termes tels que $\boldsymbol{P = Q}$ (47)

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

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

$$\vec{\xi}_{rr'}^{P} = \int \varphi_r^* \vec{\xi}\, \varphi_{r'} d\tau \;\; ; \;\; \vec{\chi}_{rs}^{PQ} = \frac{\int \varphi_r^* \vec{\chi} \varphi_s d\tau}{S_{rs}^{PQ}} \qquad (48)$$

Le moment dipolaire (45) devient [20] :

$$\vec{\mu} \;\; = \sum_p \delta_p \vec{r}_p \; + \vec{\mu}_{hybrid} + \vec{\mu}_{recouvr} \qquad (49)$$

Avec :

$$\vec{\mu}_{hybrid} = \sum_p \sum_{r,r'\in P} P_{rr'}^{PP} \vec{\xi}_{rr'}^{P} \qquad (50)$$

Et :

$$\vec{\mu}_{recouvr} = \sum_p \sum_{r,r' \in P} \sum_Q \sum_{s \in Q} P_{rs}^{PQ} S_{rs}^{PQ} \vec{\chi}_{rs}^{PQ} \quad (51)$$

II-3-3. Application

Nous avons réuni dans la figure 2 quelques applications [22] des indices électroniques de la méthode des orbitales moléculaires.

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

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

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

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

$$P_f = \frac{q_{max} - q_{min}}{F(r_{max})} \quad (52)$$

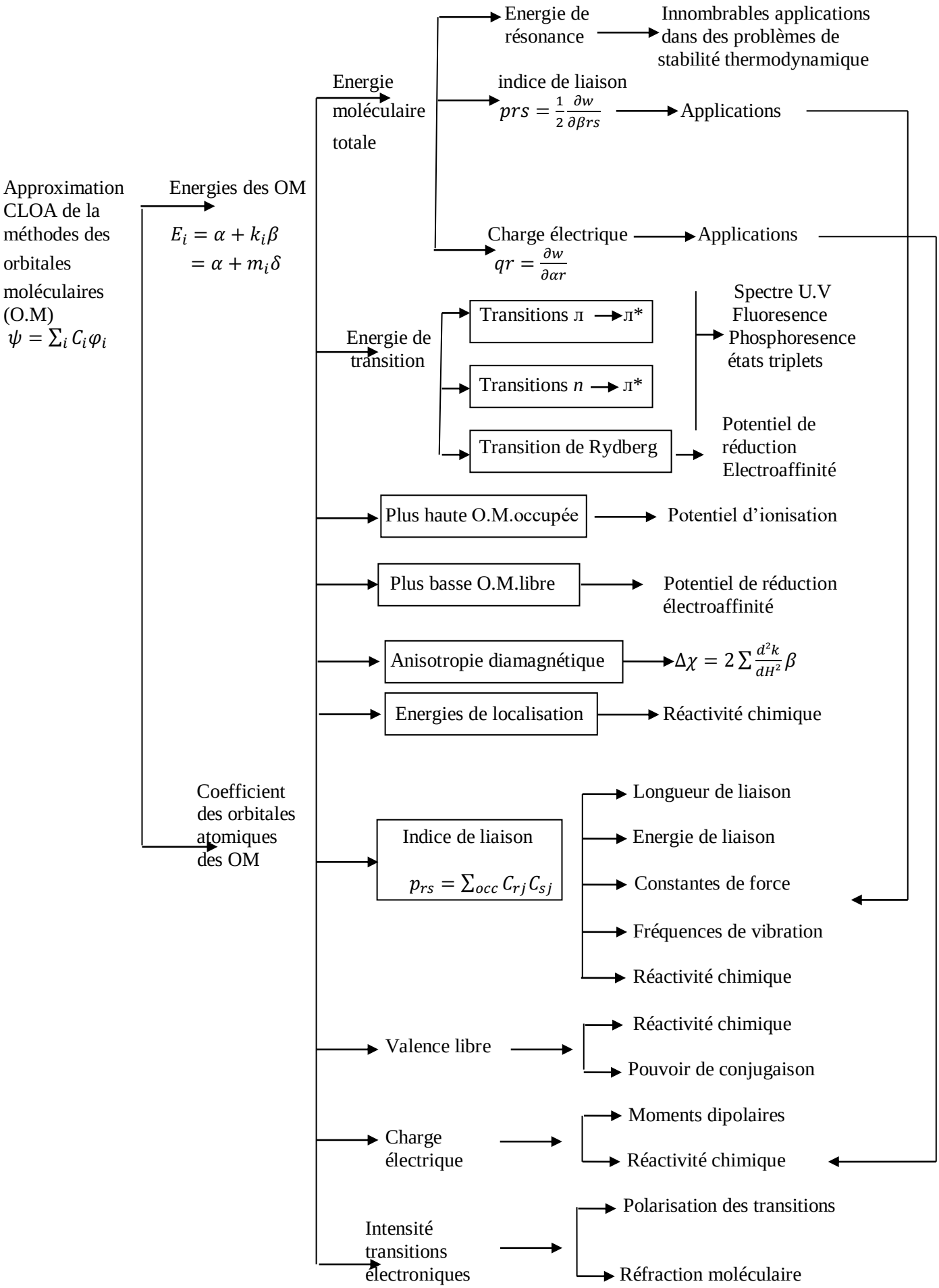


Figure - 2 : Les indices électroniques [22] de la méthode des orbitales moléculaires et leurs applications.

III- LA MÉCANİQUE MOLÉCULAİRE

Si une molécule est trop grosse pour subir un traitement semi-empirique, il est toujours possible de modéliser son comportement en évitant complètement la mécanique quantique. Les méthodes désignées par mécanique moléculaire, établissent une expression algébrique simple de l'énergie d'un composé, sans avoir à calculer une fonction d'onde ou une densité électronique totale [23]. L'expression de l'énergie consiste en des équations classiques simples, comme l'équation de l'oscillateur harmonique, dans le but de décrire l'énergie associée à l'étirement de liaison, de flexion, de rotation, et aux forces intermoléculaires, telles que les interactions de Van der Waals et de liaison hydrogène. Toutes les constantes apparaissant dans ces équations doivent être obtenues à partir de données expérimentales ou d'un calcul *ab initio*.

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

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

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

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

pas des électrons. Pour représenter l'aspect électrostatique, il n'est même pas besoin des charges atomiques partielles si l'on utilise, par exemple, des dipôles de liaisons. Au contraire de la mécanique quantique, la mécanique moléculaire nécessite plus d'informations que le numéro atomique seul. En fait, chaque atome doit être décrit de manière plus détaillée.

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

III-2. Forme fonctionnelle des champs de force courants

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

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

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

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

III-3. Quelques exemples

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

-MM2, MM3, et MM4 : (http://europa. chem. uga. edu/allinger/mm2mm3.html).
Introduit par Allinger *et al*.[24-27], largement utilisé pour le traitement de petites molécules.

-AMBER : (Assisted Method Building and Energy Refinement) (http://amber. scripps. edu)
Introduit par Cornell *et al.* [28] très largement utilisé dans le traitement des protéines et des acides nucléiques.

-CHARMM : (Chemistry at Harvard molecular Modeling) (http://yuri. harvard. edu) Développé par Mackerall, Karplus *et al.*, [29-31] qui est largement utilisé pour la simulation de petites molécules et jusqu'aux complexes solvatés de grandes macromolécules biologiques.
CHARMm est une version commerciale disponible de CHARMM qui est également applicable aux petits composés organiques [32].
-MMFF : (MerckMolecular Force Field)
Développé par Halgren [33-34], il est similaire à MM3 dans la forme, mais en diffère par son application focalisée sur les processus de condensation de phases (en dynamique moléculaire). Il reproduit l'exactitude MM3 pour les petites molécules, et est applicable aux protéines et autres systèmes d'importance biologique.

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

1) La forme de la fonction de chaque terme énergétique.
2) Le nombre de termes croisés (qui reflètent le couplage entre coordonnées internes) inclus.

3) Le type d'information utilisé pour ajuster les paramètres.

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

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

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

$$\upsilon(r^N) = \sum_{liaisons} \frac{k_i}{2}(I_i - I_{i,0})^2 + \sum_{angles} \frac{k_i}{2}(\theta_i - \theta_{i,0})^2 \sum_{torsions} \frac{V_n}{2}(1 - \cos(n\omega - \gamma)) +$$
$$\sum_{i=1}^{N} \sum_{j=i+1}^{N} \left(4\varepsilon_{ij} \left[\left(\frac{\sigma_{ij}}{r_{ij}} \right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}} \right)^{6} \right] + \frac{q_i q_j}{4\pi\varepsilon_0 r_{ij}} \right) \qquad (53)$$

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

Les diverses contributions sont représentées schématiquement sur la figure suivante :

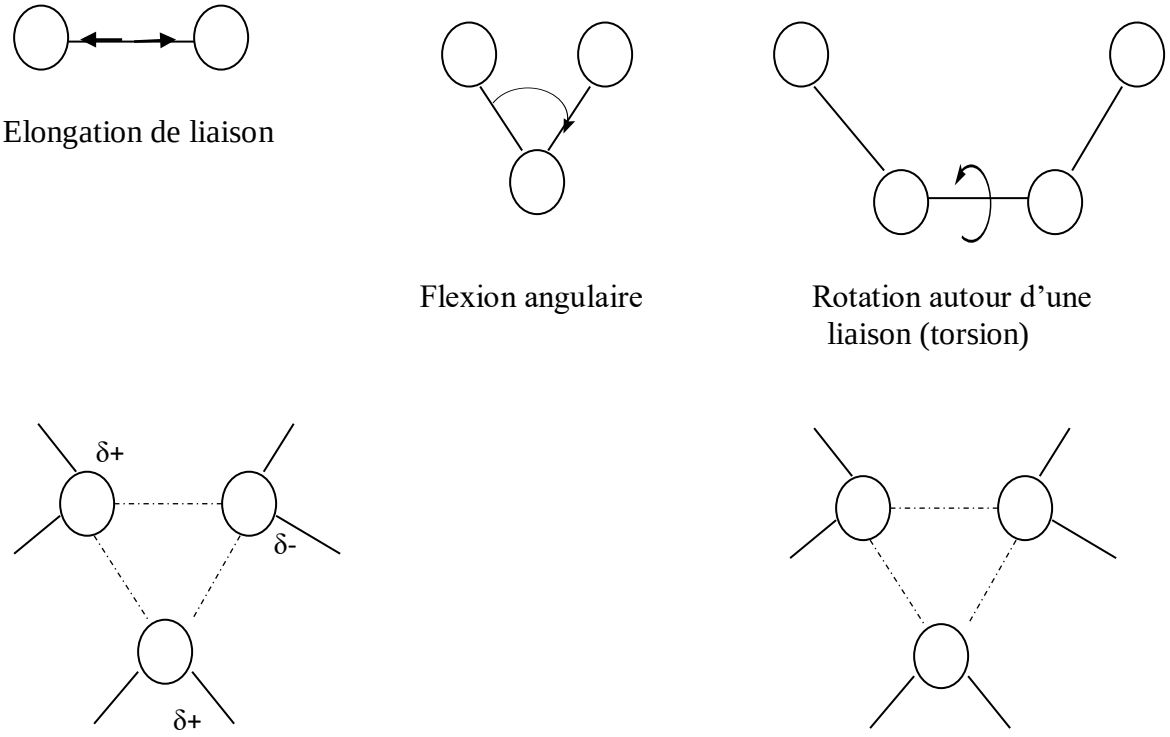


Figure 3: Représentation schématique des quatre contributions d'un champ de force de MM : élongation de liaison, flexion angulaire, termes de torsion et interactions non liées.

Le premier terme de l'équation (53) modélise l'interaction entre paires d'atomes liés, représentée dans ce cas par un potentiel harmonique qui fournit la variation de l'énergie lorsque la longueur de liaison l_i dévie de sa valeur de référence (à l'équilibre) $l_{i,0}$. Le second terme est une sommation sur tous les angles de valence de la molécule, encore modélisé par un potentiel harmonique (un angle de valence est l'angle formé par trois atomes A- B- C où A et C sont toux deux liés à B). Le troisième terme dans l'équation (53) renseigne sur la variation d'énergie lorsqu'une liaison tourne. La quatrième contribution est le terme de non liaison, qui est calculé entre toutes les paires d'atomes (i et j) localisés sur différentes molécules ou appartenant à la même molécule mais séparés par au moins 3 liaisons (c'est- à – dire avec une relation l, n où $n \geq 4$). Dans un champ de force simple le terme de non liaison comprend un terme de potentiel coulombien pour les interactions électrostatiques et le potentiel de Lennard- Jones pour les interactions de Van der Waals.

III-5. Champ de force MM2 et MM+ [35]

III-5- 1. Champ de force MM2

* Elongation des liaisons : MM2 a été paramétré pour ajuster les distances moyennes calculées à partir du mouvement vibrationnel à température ambiante à celles obtenues par diffraction électronique.

Pour mieux reproduire la courbe de Morse, MM2 fait intervenir un terme quadratique et un autre cubique :

$$v(l) = \frac{k}{2}(l - l_0)^2 \left[1 - k'(l - l_0)\right]^2 \qquad (54)$$

* Variation des angles : les déviations des angles de leurs valeurs de références sont souvent exprimées en utilisant la loi de Hooke ou potentiel harmonique :

$$v(\theta) = \frac{k}{2}(\theta - \theta_0)^2 \qquad (55)$$

Comme pour les termes d'élongation des liaisons, la précision du champ de force peut être augmentée par l'incorporation de termes d'ordres supérieurs. MM2 contient un terme d'ordre 4 en plus du terme quadratique.

$$v(\theta) = \frac{k}{2}(\theta - \theta_0)^2 \left[1 - k'(\theta - \theta_0)\right]^2 \qquad (56)$$

* Torsion des angles dièdres : correspond à la rotation d'une liaison selon l'angle dièdre ω formé par 4 atomes. Le champ de force MM2 utilise 3 termes d'une série de Fourier :

$$v(\omega) = \frac{v_1}{2}(1 + \cos\omega) + \frac{v_2}{2}(1 - \cos 2\omega) + \frac{v_3}{2}(1 + \cos 3\omega) \qquad (57)$$

Une interprétation physique a été attribuée à chacun des 3 termes à partir de l'analyse de calculs *ab initio* effectués sur des hydrocarbures fluorés simples.

* Angle dièdre impropre ou déviation extra- planaire. Examinons comment la cyclobutanone serait modélisée en utilisant uniquement un champ de force comportant des termes

standards pour l'élongation des liaisons et les variations angulaires du type de ceux apparaissant dans l'équation (57). La structure d'équilibre obtenue avec un tel champ de force sera caractérisée par la localisation de l'atome d'oxygène hors du plan formé par l'atome de carbone contigu (à l'oxygène) et les 2 carbones qui lui sont liés (Figure.4).

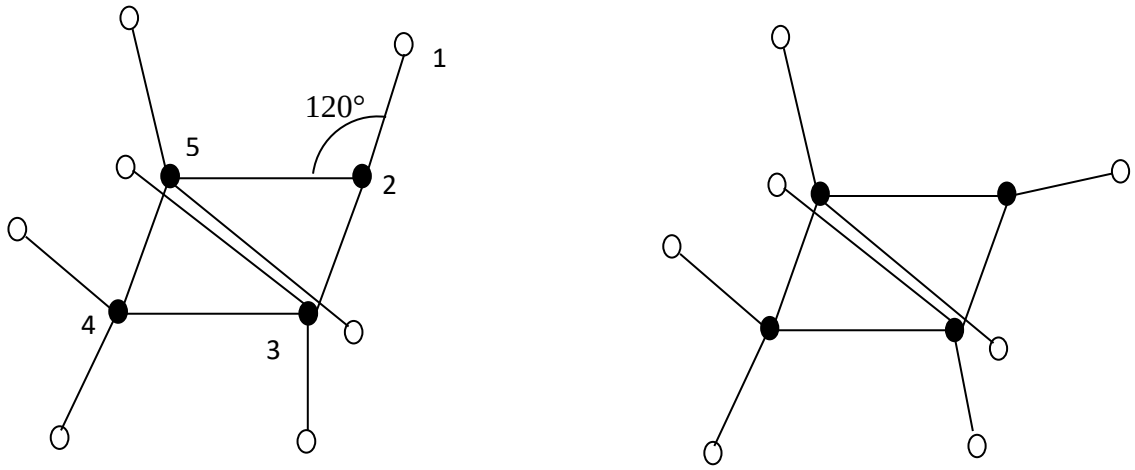


Figure 4: Sous un terme extra- planaire, l'atome d'oxygène de la cyclobutanone est prévu de se situer hors du plan du cycle / gauche plutôt que dans le plan.

Dans cette configuration, les angles vers l'oxygène adoptent des valeurs proches de la valeur de référence 120°. Expérimentalement, il est établi que l'atome d'oxygène demeure dans le plan du cyclobutane bien que les angles C-C=O soient grands (133°). Ceci parce que l'énergie de liaison π, qui est maximalisée dans l'arrangement coplanaire, sera plus réduite si l'atome d'oxygène est dévié hors du plan. Pour réaliser la géométrie désirée il est nécessaire d'ajouter un (ou des) terme (s) additionnel (s) dans le champ de force qui maintienne (nt) le carbone sp^2 et les 3 atomes qui lui sont liés dans le même plan. La plus simple façon de le réaliser est d'utiliser un terme de variation angulaire extra- planaire.

Il y a plusieurs façons d'incorporer des termes de variation extra- planaire dans un champ de force. Une approche possible est de traiter les 4 atomes comme un angle de torsion impropre pour lequel les 4 atomes (Figure.5) ne sont pas liés dans la séquence 1- 2- 3- 4. Une façon de définir, dans ce cas, une torsion impropre consiste à impliquer les atomes 1- 5- 3- 2 de la figure.

Un potentiel de torsion de la forme suivante :

$$v(\omega) = k(1 - \cos 2\omega) \qquad (58)$$

peut être utilisé pour maintenir l'angle de rotation impropre à 0° ou 180°.

Il existe diverses autres façons pour inclure la contribution de la variation d'angle extra- planaire. Par exemple, une définition qui est la plus proche de la notion de « variation extra- planaire » implique le calcul de l'angle entre une liaison à partir de l'atome central et le plan défini par l'atome central et les 2 autres atomes (Figure.5). La valeur 0° correspond aux 4 atomes coplanaires. Une troisième approche consiste à calculer la hauteur de l'atome central au- dessus du plan défini par les 3 autres atomes (Figure.5). Avec ces deux définitions la déviation de la coordonnée extra- planaire (angle ou distance) peut être modélisée à l'aide d'un potentiel harmonique de la forme :

$$v(\theta) = \frac{k}{2}\,\theta^2 \quad ; \quad v(h) = \frac{k}{2}\,h^2 \qquad (59)$$

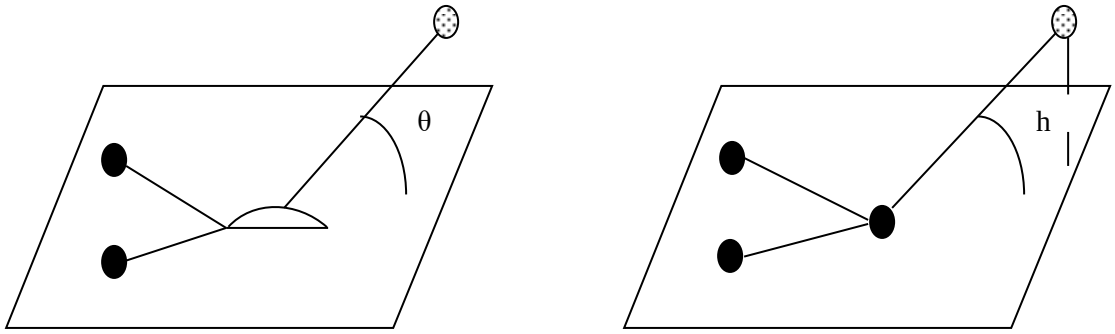


Figure 5: Deux façons pour modéliser les contributions de la variation d'angle extra- planaire.

* Termes de croisement : Les termes de croisement permettent de tenir compte des interactions entre l'élongation des liaisons, la variation des angles et la torsion des angles dièdres. Par exemple, si les liaisons C-O et O-H de l'angle C-O-H sont étirées, alors la distance entre les atomes terminaux (C et H) est augmentée, ce qui rend plus facile la diminution de l'angle COH. Pareillement, la diminution de l'angle COH tend a être accompagnée d'une augmentation des longueurs des liaisons O-H et C-O. Pour tenir compte de cette interaction, on peut ajouter un terme de croisement « élongation- variation angulaire ». (stretch- bend) de la forme :

$$v_{\Delta\theta} = \frac{1}{2}\,k_{12}(\Delta l_1 + \Delta l_2)\,\Delta\theta \qquad (60)$$

avec $\Delta l_1 = l_1 - l_{10}$; $\Delta l_2 = l_2 - l_{20}$ et $\Delta\theta = \theta - \theta_0$

l_{10}, l_{20} et θ_0 représentent les valeurs de références pour l_1, l_2 et θ respectivement.

Les termes de croisement les plus utilisés sont (Figure 6) :

* élongation- élongation et élongation- variation angulaire, pour deux liaisons à un même atome ;

* élongation- torsion angle dièdre, variation angulaire- torsion angle dièdre et variation angulaire-variation angulaire pour 2 angles avec un atome central commun.

Le champ de force MM2 fait intervenir uniquement un terme de croisement élongation-variation angulaire.

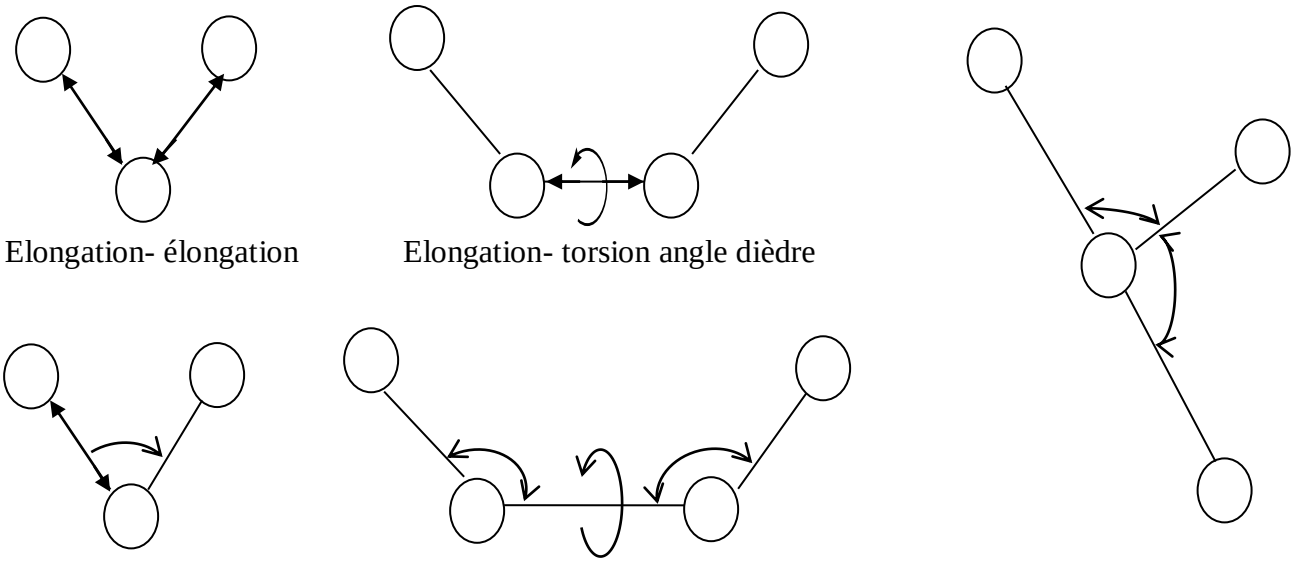


Figure 6: Illustration schématique des termes de croisement supposés les plus importants pour les champs de forces.

* Interactions électrostatiques : Le terme électrostatique v_{es} est généralement défini par la somme des interactions électrostatiques entre toutes les paires d'atomes exceptées les paires 1, 2 et 1, 3 : $v_{es} = \sum l \geq 4\, v_{es,ij}$, où les atomes i, j vérifient la relation ($l \geq 4$).

v_{es} est généralement calculé en affectant des charges atomiques partielles à chaque atome et en appliquant un potentiel coulombien.

MM2 n'utilise pas cette procédure mais associe un moment dipolaire avec chaque liaison puis calcule v_{es} comme somme des énergies potentielles d'interactions entre moments de liaisons. Chaque moment dipolaire étant localisé au centre, et dirigé le long de cette liaison. Les valeurs des moments de liaisons de certains types de liaisons ont été choisies pour ajuster les moments dipolaires expérimentaux de petites molécules.

L'équation (61) décrit un modèle de charge, basé essentiellement sur les dipôles centrés, utilisé dans MM2 [24].

$$v_{es} = \frac{\mu_i \mu_j}{kr^3}(\cos \chi - 3 \cos \alpha_i \cos \alpha_j) \qquad (61)$$

χ et α_i, α_j désignent respectivement les angles entre les dipôles et les angles entre chaque dipôle et le vecteur de connexion.

* Interactions de Van der Waals : la plupart des champs de forces utilisent le potentiel 12- 6 de Lennard Jones.

MM2 utilise le potentiel de Buckingham, avec un terme attractif proportionnel à r^{-6} et un terme répulsif proportionnel à $e^{-\alpha r}$ où α est un paramètre :

$$v_{vdw} = A\, e^{-\alpha r} - \frac{B}{r^6} \qquad (62)$$

* Liaisons conjuguées : MM2 utilise une procédure générale qui consiste à effectuer des calculs semi- empiriques sur les électrons π pour en tirer les ordres de liaisons, qui sont ensuite utilisés pour affecter des longueurs de liaisons et des constantes de force de référence pour les liaisons conjuguées.

III-5-2. Champ de force MM+

C'est une extension du Champ de force MM2, avec l'ajout de quelques paramètres additionnels. MM+ est un champ de force robuste, il a l'aptitude de prendre en considération les paramètres négligés dans d'autres champs de force et peut donc s'appliquer pour des molécules plus complexes tels que les composés inorganiques.

Le tableau suivant [18] compare les trois techniques computationnelles majeures évoquées.

Tableau - 1: Etude comparative des techniques *ab initio*, semi- empirique et mécanique moléculaire.

ab initio	**Semi- empirique**	**Mécanique moléculaire**
- Prise en compte de tous les électrons. - Limité à quelques dizaines d'atomes. Nécessite un super calculateur - Peut être appliquée à des composés inorganiques, organométalliques, et aux fragments moléculaires (composants catalytiques d'enzymes). - Vide, solvatation implicite. - Applicable à l'état fondamental, et aux états de transition et excité.	- Ignore certains électrons (simplification). - Limité à quelques centaines d'atomes. - Peut être appliquée à des composés inorganiques, organiques, organométalliques et de petits oligomères (peptides, nucléotides, saccharides). - Vide, solvatation implicite. - Applicable à l'état fondamental, et aux états de transition et excité.	- Ignore tous les électrons. Seuls les noyaux sont considérés. - Molécules contenant des milliers d'atomes - Peut être appliquée aux composés inorganiques, organiques, oligonucléotides, peptides, saccharides, métallo-organiques et inorganiques. - Vide, solvatation implicite ou explicite. - Applicable uniquement à l'état fondamental.

IV-LA DYNAMİQUE MOLÉCULAİRE

La dynamique moléculaire a débuté avec l'arrivée, en 1957, des premiers ordinateurs [36]. Mais les premières simulations réelles ont été faites par Rahman [37], grâce à ses travaux sur la simulation de l'argon liquide, en 1964, avec un temps de simulation de 10^{-11} s, puis de l'eau liquide [38] en 1971.

IV—1. Principe de la dynamique moléculaire

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

$$\vec{F}_i = m_i \vec{a}_i = m_i \frac{d^2 \vec{r}_i(t)}{dt^2} \quad (63)$$

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

m_i : masse de l'atome i.

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

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

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

$$E_c = \sum_{i=1}^{N} \frac{|\vec{P}_i|}{2n_i} \quad (64)$$

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

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

$$E_c = \frac{3K_b T}{2}(3N - N_c) \quad (65)$$

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

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

$$\vec{F}_i = \frac{d\vec{E}(r_i \ldots r_n)}{dr_i(t)} \quad (66)$$

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

r_i : coordonnées cartésiennes de l'atome i.

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

$$\vec{a}_i = \frac{d\vec{V}_i}{dt} \quad (67)$$

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

$$\vec{V}_i = \frac{d\vec{r}_i}{dt} \quad (68)$$

pas des électrons. Pour représenter l'aspect électrostatique, il n'est même pas besoin des charges atomiques partielles si l'on utilise, par exemple, des dipôles de liaisons. Au contraire de la mécanique quantique, la mécanique moléculaire nécessite plus d'informations que le numéro atomique seul. En fait, chaque atome doit être décrit de manière plus détaillée.

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

III-2. Forme fonctionnelle des champs de force courants

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

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

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

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

III-3. Quelques exemples

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

-MM2, MM3, et MM4 : (http://europa. chem. uga. edu/allinger/mm2mm3.html).
Introduit par Allinger *et al*.[24-27], largement utilisé pour le traitement de petites molécules.

-AMBER : (Assisted Method Building and Energy Refinement) (http://amber. scripps. edu)
Introduit par Cornell *et al.* [28] très largement utilisé dans le traitement des protéines et des acides nucléiques.

-CHARMM : (Chemistry at Harvard molecular Modeling) (http://yuri. harvard. edu) Développé par Mackerall, Karplus *et al.*, [29-31] qui est largement utilisé pour la simulation de petites molécules et jusqu'aux complexes solvatés de grandes macromolécules biologiques.
CHARMm est une version commerciale disponible de CHARMM qui est également applicable aux petits composés organiques [32].
-MMFF : (MerckMolecular Force Field)
Développé par Halgren [33-34], il est similaire à MM3 dans la forme, mais en diffère par son application focalisée sur les processus de condensation de phases (en dynamique moléculaire). Il reproduit l'exactitude MM3 pour les petites molécules, et est applicable aux protéines et autres systèmes d'importance biologique.

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

1) La forme de la fonction de chaque terme énergétique.
2) Le nombre de termes croisés (qui reflètent le couplage entre coordonnées internes) inclus.

3) Le type d'information utilisé pour ajuster les paramètres.

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

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

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

$$\upsilon(r^N) = \sum_{liaisons} \frac{k_i}{2}(I_i - I_{i,0})^2 + \sum_{angles} \frac{k_i}{2}(\theta_i - \theta_{i,0})^2 \sum_{torsions} \frac{V_n}{2}(1 - \cos(n\omega - \gamma)) +$$
$$\sum_{i=1}^{N} \sum_{j=i+1}^{N} \left(4\varepsilon_{ij} \left[\left(\frac{\sigma_{ij}}{r_{ij}} \right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}} \right)^{6} \right] + \frac{q_i q_j}{4\pi\varepsilon_0 r_{ij}} \right) \qquad (53)$$

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

Les diverses contributions sont représentées schématiquement sur la figure suivante :

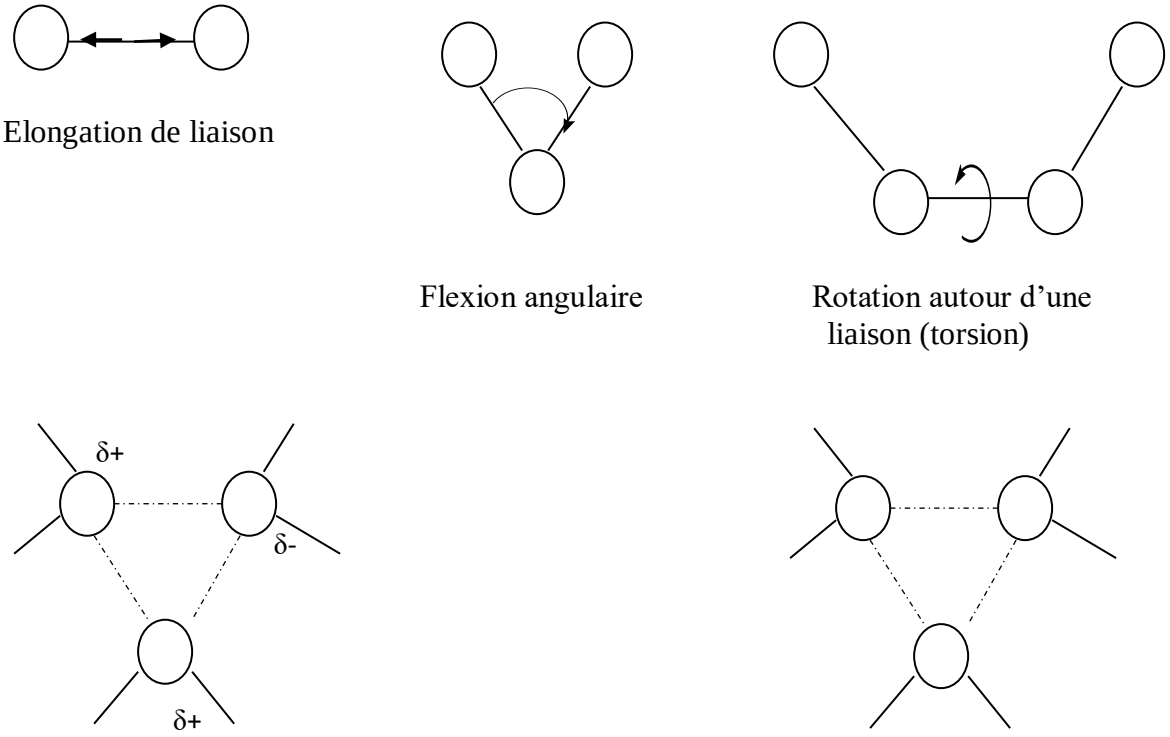


Figure 3: Représentation schématique des quatre contributions d'un champ de force de MM : élongation de liaison, flexion angulaire, termes de torsion et interactions non liées.

Le premier terme de l'équation (53) modélise l'interaction entre paires d'atomes liés, représentée dans ce cas par un potentiel harmonique qui fournit la variation de l'énergie lorsque la longueur de liaison l_i dévie de sa valeur de référence (à l'équilibre) $l_{i,0}$. Le second terme est une sommation sur tous les angles de valence de la molécule, encore modélisé par un potentiel harmonique (un angle de valence est l'angle formé par trois atomes A- B- C où A et C sont toux deux liés à B). Le troisième terme dans l'équation (53) renseigne sur la variation d'énergie lorsqu'une liaison tourne. La quatrième contribution est le terme de non liaison, qui est calculé entre toutes les paires d'atomes (i et j) localisés sur différentes molécules ou appartenant à la même molécule mais séparés par au moins 3 liaisons (c'est- à – dire avec une relation l, n où $n \geq 4$). Dans un champ de force simple le terme de non liaison comprend un terme de potentiel coulombien pour les interactions électrostatiques et le potentiel de Lennard- Jones pour les interactions de Van der Waals.

III-5. Champ de force MM2 et MM+ [35]

III-5- 1. Champ de force MM2

* Elongation des liaisons : MM2 a été paramétré pour ajuster les distances moyennes calculées à partir du mouvement vibrationnel à température ambiante à celles obtenues par diffraction électronique.

Pour mieux reproduire la courbe de Morse, MM2 fait intervenir un terme quadratique et un autre cubique :

$$v(l) = \frac{k}{2}(l - l_0)^2 \left[1 - k'(l - l_0)\right]^2 \qquad (54)$$

* Variation des angles : les déviations des angles de leurs valeurs de références sont souvent exprimées en utilisant la loi de Hooke ou potentiel harmonique :

$$v(\theta) = \frac{k}{2}(\theta - \theta_0)^2 \qquad (55)$$

Comme pour les termes d'élongation des liaisons, la précision du champ de force peut être augmentée par l'incorporation de termes d'ordres supérieurs. MM2 contient un terme d'ordre 4 en plus du terme quadratique.

$$v(\theta) = \frac{k}{2}(\theta - \theta_0)^2 \left[1 - k'(\theta - \theta_0)\right]^2 \qquad (56)$$

* Torsion des angles dièdres : correspond à la rotation d'une liaison selon l'angle dièdre ω formé par 4 atomes. Le champ de force MM2 utilise 3 termes d'une série de Fourier :

$$v(\omega) = \frac{v_1}{2}(1 + \cos\omega) + \frac{v_2}{2}(1 - \cos 2\omega) + \frac{v_3}{2}(1 + \cos 3\omega) \qquad (57)$$

Une interprétation physique a été attribuée à chacun des 3 termes à partir de l'analyse de calculs *ab initio* effectués sur des hydrocarbures fluorés simples.

* Angle dièdre impropre ou déviation extra- planaire. Examinons comment la cyclobutanone serait modélisée en utilisant uniquement un champ de force comportant des termes

standards pour l'élongation des liaisons et les variations angulaires du type de ceux apparaissant dans l'équation (57). La structure d'équilibre obtenue avec un tel champ de force sera caractérisée par la localisation de l'atome d'oxygène hors du plan formé par l'atome de carbone contigu (à l'oxygène) et les 2 carbones qui lui sont liés (Figure.4).

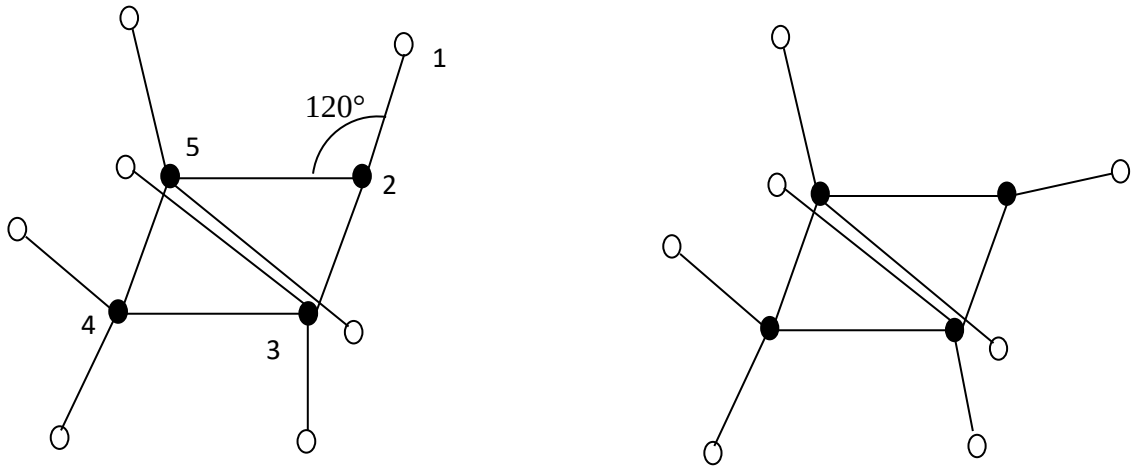


Figure 4: Sous un terme extra- planaire, l'atome d'oxygène de la cyclobutanone est prévu de se situer hors du plan du cycle / gauche plutôt que dans le plan.

Dans cette configuration, les angles vers l'oxygène adoptent des valeurs proches de la valeur de référence 120°. Expérimentalement, il est établi que l'atome d'oxygène demeure dans le plan du cyclobutane bien que les angles C-C=O soient grands (133°). Ceci parce que l'énergie de liaison π, qui est maximalisée dans l'arrangement coplanaire, sera plus réduite si l'atome d'oxygène est dévié hors du plan. Pour réaliser la géométrie désirée il est nécessaire d'ajouter un (ou des) terme (s) additionnel (s) dans le champ de force qui maintienne (nt) le carbone sp^2 et les 3 atomes qui lui sont liés dans le même plan. La plus simple façon de le réaliser est d'utiliser un terme de variation angulaire extra- planaire.

Il y a plusieurs façons d'incorporer des termes de variation extra- planaire dans un champ de force. Une approche possible est de traiter les 4 atomes comme un angle de torsion impropre pour lequel les 4 atomes (Figure.5) ne sont pas liés dans la séquence 1- 2- 3- 4. Une façon de définir, dans ce cas, une torsion impropre consiste à impliquer les atomes 1- 5- 3- 2 de la figure.

Un potentiel de torsion de la forme suivante :

$$v(\omega) = k(1 - \cos 2\omega) \qquad (58)$$

peut être utilisé pour maintenir l'angle de rotation impropre à 0° ou 180°.

Il existe diverses autres façons pour inclure la contribution de la variation d'angle extra- planaire. Par exemple, une définition qui est la plus proche de la notion de « variation extra- planaire » implique le calcul de l'angle entre une liaison à partir de l'atome central et le plan défini par l'atome central et les 2 autres atomes (Figure.5). La valeur 0° correspond aux 4 atomes coplanaires. Une troisième approche consiste à calculer la hauteur de l'atome central au- dessus du plan défini par les 3 autres atomes (Figure.5). Avec ces deux définitions la déviation de la coordonnée extra- planaire (angle ou distance) peut être modélisée à l'aide d'un potentiel harmonique de la forme :

$$v(\theta) = \frac{k}{2}\,\theta^2 \quad ; \quad v(h) = \frac{k}{2}\,h^2 \qquad (59)$$

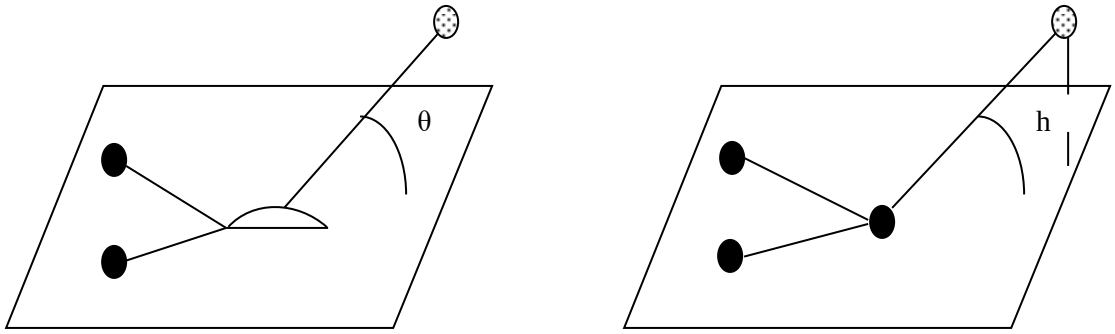


Figure 5: Deux façons pour modéliser les contributions de la variation d'angle extra- planaire.

* Termes de croisement : Les termes de croisement permettent de tenir compte des interactions entre l'élongation des liaisons, la variation des angles et la torsion des angles dièdres. Par exemple, si les liaisons C-O et O-H de l'angle C-O-H sont étirées, alors la distance entre les atomes terminaux (C et H) est augmentée, ce qui rend plus facile la diminution de l'angle COH. Pareillement, la diminution de l'angle COH tend a être accompagnée d'une augmentation des longueurs des liaisons O-H et C-O. Pour tenir compte de cette interaction, on peut ajouter un terme de croisement « élongation- variation angulaire ». (stretch- bend) de la forme :

$$v_{\Delta\theta} = \frac{1}{2}\,k_{12}(\Delta l_1 + \Delta l_2)\,\Delta\theta \qquad (60)$$

avec $\Delta l_1 = l_1 - l_{10}$; $\Delta l_2 = l_2 - l_{20}$ et $\Delta\theta = \theta - \theta_0$

l_{10}, l_{20} et θ_0 représentent les valeurs de références pour l_1, l_2 et θ respectivement.

Les termes de croisement les plus utilisés sont (Figure 6) :

* élongation- élongation et élongation- variation angulaire, pour deux liaisons à un même atome ;

* élongation- torsion angle dièdre, variation angulaire- torsion angle dièdre et variation angulaire-variation angulaire pour 2 angles avec un atome central commun.

Le champ de force MM2 fait intervenir uniquement un terme de croisement élongation-variation angulaire.

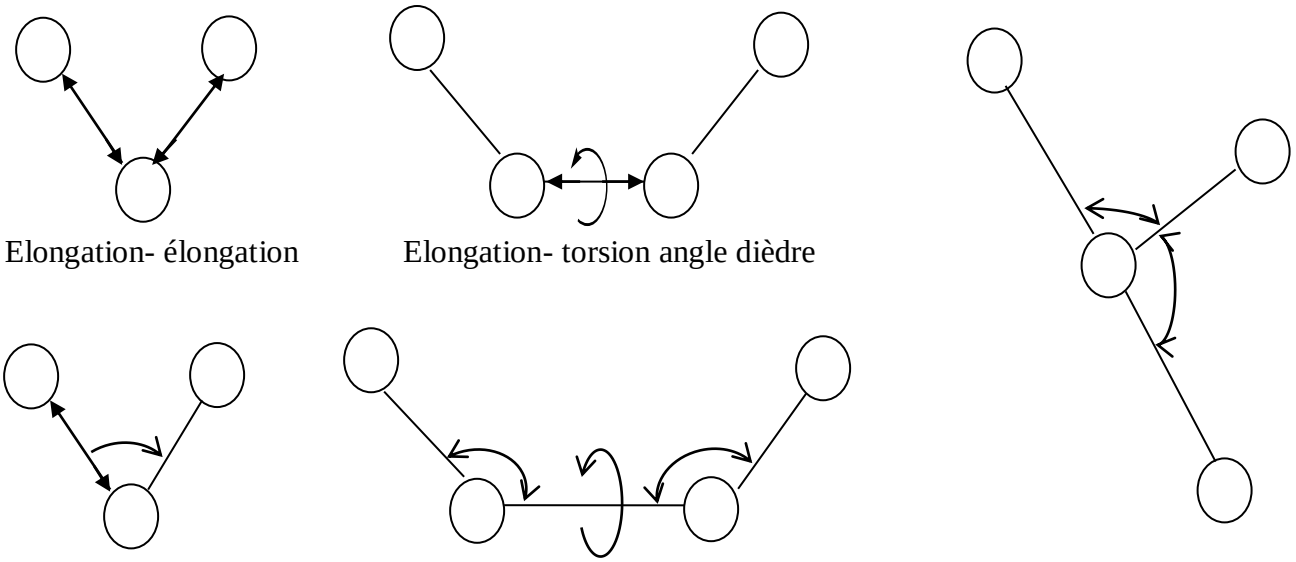


Figure 6: Illustration schématique des termes de croisement supposés les plus importants pour les champs de forces.

* Interactions électrostatiques : Le terme électrostatique v_{es} est généralement défini par la somme des interactions électrostatiques entre toutes les paires d'atomes exceptées les paires 1, 2 et 1, 3 : $v_{es} = \sum l \geq 4\, v_{es,ij}$, où les atomes i, j vérifient la relation ($l \geq 4$).

v_{es} est généralement calculé en affectant des charges atomiques partielles à chaque atome et en appliquant un potentiel coulombien.

MM2 n'utilise pas cette procédure mais associe un moment dipolaire avec chaque liaison puis calcule v_{es} comme somme des énergies potentielles d'interactions entre moments de liaisons. Chaque moment dipolaire étant localisé au centre, et dirigé le long de cette liaison. Les valeurs des moments de liaisons de certains types de liaisons ont été choisies pour ajuster les moments dipolaires expérimentaux de petites molécules.

L'équation (61) décrit un modèle de charge, basé essentiellement sur les dipôles centrés, utilisé dans MM2 [24].

$$v_{es} = \frac{\mu_i \mu_j}{kr^3}(\cos \chi - 3 \cos \alpha_i \cos \alpha_j) \qquad (61)$$

χ et α_i, α_j désignent respectivement les angles entre les dipôles et les angles entre chaque dipôle et le vecteur de connexion.

* Interactions de Van der Waals : la plupart des champs de forces utilisent le potentiel 12- 6 de Lennard Jones.

MM2 utilise le potentiel de Buckingham, avec un terme attractif proportionnel à r^{-6} et un terme répulsif proportionnel à $e^{-\alpha r}$ où α est un paramètre :

$$v_{vdw} = A\, e^{-\alpha r} - \frac{B}{r^6} \qquad (62)$$

* Liaisons conjuguées : MM2 utilise une procédure générale qui consiste à effectuer des calculs semi- empiriques sur les électrons π pour en tirer les ordres de liaisons, qui sont ensuite utilisés pour affecter des longueurs de liaisons et des constantes de force de référence pour les liaisons conjuguées.

III-5-2. Champ de force MM+

C'est une extension du Champ de force MM2, avec l'ajout de quelques paramètres additionnels. MM+ est un champ de force robuste, il a l'aptitude de prendre en considération les paramètres négligés dans d'autres champs de force et peut donc s'appliquer pour des molécules plus complexes tels que les composés inorganiques.

Le tableau suivant [18] compare les trois techniques computationnelles majeures évoquées.

VIII. Développement de modèles QSAR/QSPR

Nous avons utilisé le logiciel de modélisation moléculaire HyperChem 6.03 [62] pour représenter les molécules puis, à l'aide de la méthode semi-empirique AM1, PM3 [13], on a obtenu les géométries finales. Tous les calculs ont été menés dans le cadre du formalisme RHF [63] sans interaction de configuration. Les structures moléculaires ont été prèoptimisées à l'aide de l'algorithme Polak- Ribiere avec pour critère une racine du carré moyen du gradient égale à 0,001kcal/mol. Les géométries ainsi optimisées ont été transférées dans le logiciel informatique Dragon version 5.3 [64] pour le calcul de plus de 1200 descripteurs 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.

VIII.1. Sélection d'un sous- ensemble de variables par algorithme génétique (GA- VSS)

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 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.

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 le présent travail, la sélection des descripteurs a été réalisée par algorithme génétique, dans la version MOBY DIGS de Todeschini [65], en maximisant Q^2_{LOO}.

VIII.2. Méthodes utilisées pour le développement de modèles QSAR/QSPR

L'application pratique des gammes de descripteurs moléculaires dans le développement de modèles QSAR/QSPR n'est pas une tâche aisée [64]. Tout d'abord, un très grand nombre (~10 000) de descripteurs moléculaires, de différentes complexités et de conceptions diverses ont été imaginés et proposés au cours des (60 dernières) années. Ensuite, pendant ce temps, aucune règle stricte n'a été établie, ni même proposée, pour la sélection de descripteurs adaptés parmi la myriade de descripteurs disponibles. Ce choix a souvent été basé sur l'intuition chimique des chercheurs, ou en se pliant à la tradition.

Une autre difficulté dans la sélection des descripteurs QSAR/QSPR découle de la non standardisation des gammes de descripteurs. Les gammes empiriques des constantes d'induction, de résonance et d'effet stérique des constituants, ou les échelles empiriques d'effets de solvant comportent des erreurs intrinsèques liées aux erreurs respectives des mesures expérimentales. Par ailleurs, les méthodes quanto- mécaniques appliquées aux calculs des descripteurs moléculaires et aux distributions de charges liés aux OM sont souvent basées sur différents paramètres semi-empiriques, ou l'utilisation de différents ensembles de base dans les calculs ab- initio. Naturellement, un descripteur construit à l'aide de différentes méthodes expérimentales ou théoriques, pour divers composés, ne peut être utilisé pour le calcul d'un modèle QSAR/QSPR unique. Une approche systématique pour la sélection de gammes de descripteurs pour le calcul de modèles QSAR/QSPR est basée sur la discrimination statistique entre de larges ensembles de descripteurs.

Dans ce qui suit nous passerons en revue diverses approches utilisées pour le développement des " meilleures" équations QSPR dans de grands espaces de descripteurs.

En dernier ressort, les modèles QSAR/QSPR peuvent être développés selon des modèles mathématiques différents, généralement en relation avec l'analyse statistique multivariée. Le premier modèle, et le plus largement utilisé, consiste en une équation (multi) linéaire obtenue par

régression des données expérimentales en fonction d'un ensemble de descripteurs pré-sélectionnés (ou d'un seul), en utilisant la méthode des moindres carrés ordinaires (MCO). Dans quelques cas, les modèles physiques ou chimiques connus du phénomène étudié laissent prévoir certaines formes mathématiques non linéaires (exponentielles ou logarithmiques) de la dépendance entre les données expérimentales et les descripteurs moléculaires. Les modèles QSAR/QSPR peuvent alors être établis à l'aide de la technique de régression par les moindres carrés non linéaires. D'autres modèles ont été développés en utilisant l'analyse factorielle ou l'analyse en composantes principales. L'intérêt de ces méthodes est qu'elles évacuent le problème de multicolinéarité inhérent aux méthodes de régression linéaires. Cependant, l'interprétation des équations QSAR/QSPR est alors entravée par la nature formelle des facteurs ou des composantes principales. Une alternative aux méthodes très classiques de régression linéaire multiple (RLM) est l'analyse en composantes principales (ACP), de même que la technique de régression par les moindres carrés partiels (MCP ou PLS) [66-71].

On a également appliqué les méthodes modernes de l'intelligence artificielle au développement de modèles QSAR/QSPR [72-76]. Ces méthodes comprennent: les réseaux de neurones (RNA), les algorithmes génétiques (GA), et d'autres méthodes globales d'optimisation.

Nous présenterons dans ce qui suit une courte vue d'ensemble des différentes méthodes mathématiques utilisées pour développer nos modèles.

VIII.2. 1. La régression linéaire multiple :

L'étude d'un phénomène peut, le plus souvent, être schématisé de la manière suivante: 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.

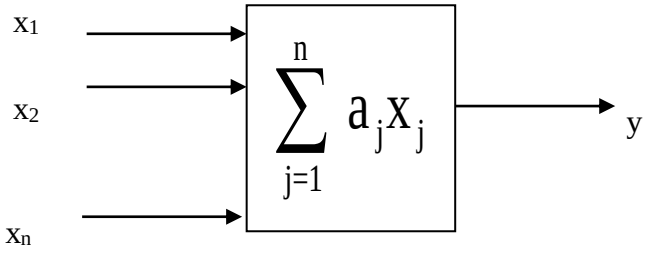
La régression est une des méthodes les plus connues et les plus appliquées en statistique pour l'analyse de données quantitatives. Elle est utilisée pour établir une liaison entre une variable quantitative et une ou plusieurs autres variables quantitatives, sous la forme d'un modèle. Si on s'intéresse 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 [77].

La régression multi-linéaire (MLR, pour Multiple Linear Regression) [78] 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 :

$$y = a_0 + a_1x_1 + a_2x_2 + \ldots + a_nx_n \qquad (71)$$

où les a_i sont les coefficients de la régression.



La détermination de l'équation (71) 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.

$$\hat{y}_1 = a_0 + a_1x_{1,1} + a_2x_{2,1} + \ldots + a_nx_{n,1} + \varepsilon_1$$

$$\hat{y}_2 = a_0 + a_1x_{1,2} + a_2x_{2,2} + \ldots + a_nx_{n,2} + \varepsilon_2 \qquad (72)$$

$$\hat{y}_p = a_0 + a_1x_{1,p} + a_2x_{2,p} + \ldots + a_nx_{n,p} + \varepsilon_p$$

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 :

$$\begin{pmatrix} y_1 \\ . \\ . \\ . \\ y_p \end{pmatrix} = \begin{pmatrix} 1 & x_{1,1} & \dots & x_{n,1} \\ . & . & . & . \\ . & . & . & . \\ . & . & . & . \\ 1 & x_{1,p} & \dots & x_{1,p} \end{pmatrix} \begin{pmatrix} a_0 \\ . \\ . \\ . \\ a_p \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ . \\ . \\ . \\ \varepsilon_p \end{pmatrix} \quad (73)$$

soit de manière condensée :

$$\mathbf{Y} = \mathbf{X}\,\mathbf{A} + \boldsymbol{\varepsilon} \quad (74)$$

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 xi, 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 xi est supposée linéaire.

La valeur prédite de la variable dépendante est alors :

$$\hat{y}_i = \hat{a}_0 + \hat{a}_1 x_{1,i} + \dots + \hat{a}_n x_{n,i} \quad (75)$$

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

$$\varepsilon_i = y_i - \hat{y}_i \quad (76)$$

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.

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

$$= \min (Y-X\hat{A})^{T} (Y-X\hat{A}) \quad (77)$$

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

$$\mathbf{A} = \mathbf{(X^{T}X)^{-1}X^{T}Y} \quad (78)$$

Bien entendu, la régression multi-linéaire 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 [79].

VIII.2.2. Réseaux de Neurones Artificiels

Les réseaux de neurones ont été étudiés depuis les années 40 [80]. Les idées de base de cette technique viennent de la recherche cognitive, d'où vient le nom ″réseaux de neurones″.

La technique inspirait beaucoup de chercheurs à cette époque, mais beaucoup de l'intérêt disparaît après un article de Minsky et Papert [81]. Finalement relancée au début des années 80 après un quasi-oubli d'une vingtaine d'années. La cause de l'intérêt soudain était l'apparition de nouvelles architectures de réseaux de neurones.

VIII. 2.2.1 Le neurone artificiel :

L'élément de base d'un réseau de neurones est, bien entendu, le neurone artificiel. Un neurone (figure 8) contient deux éléments principaux :

- Un ensemble de poids associés aux connections du neurone, et
- Une fonction d'activation (Figure 9).

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

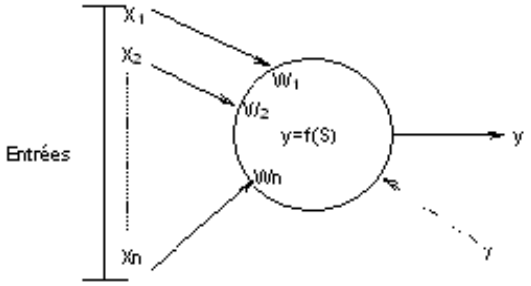
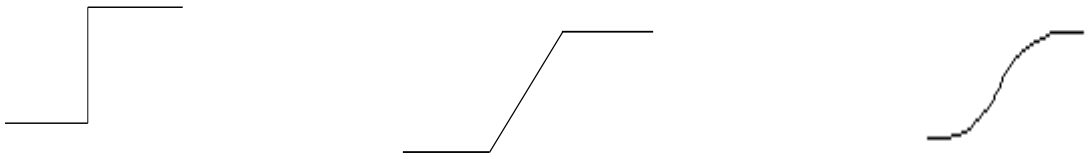


Figure 8 : *le neurone artificiel générique.*

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



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

Figure 9 : *Fonctions d'activation.*

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

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

VIII. 2.2.2. Propriétés des réseaux de neurones :

Un réseau de neurones se compose de neurones qui sont interconnectés de façon à ce que la sortie d'un neurone puisse être l'entrée d'un ou plusieurs autres neurones. Ensuite il y a des entrées de l'extérieur et des sorties vers l'extérieur [82].

Rumelbart et al. [82] donnent huit composants principaux d'un réseau de neurones:

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

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

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 [82], c'est-à-dire que plusieurs neurones sont plus ou moins actifs et contribuent à une décision. Une autre possibilité est d'avoir des représentations localisées, ce qui permet d'identifier le rôle de chaque neurone plus facilement. Les réseaux de neurones ont quand même une tendance à produire des présentations distribuées.

VIII.2.2.3. Les différents types de réseaux de neurones

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

- Le réseau de Hopfield (et sa version incluant l'apprentissage, la machine de Boltzmann).

- Les cartes auto-organisatrices de Kohonen .
- Les réseaux à fonction radiale que l'on nomme aussi RBF (pour " Radial Basic Functions ").
- Les réseaux multicouches ou perceptron multicouches PMC

Le réseau de Hopfield [83] est un réseau avec des sorties binaires où tous les neurones sont interconnectés avec des poids symétriques. C'est-à-dire que le poids du neurone N_i au neurone N_j est égal au poids du neurone N_j au neurone N_i. Les poids sont donnés par l'utilisateur. Les poids et les états des neurones permettent de définir l'«énergie» du réseau.

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

Les cartes auto-organisatrices de Kohonen [84] sont utilisées pour faire des classifications automatiques des vecteurs d'entrées.

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

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

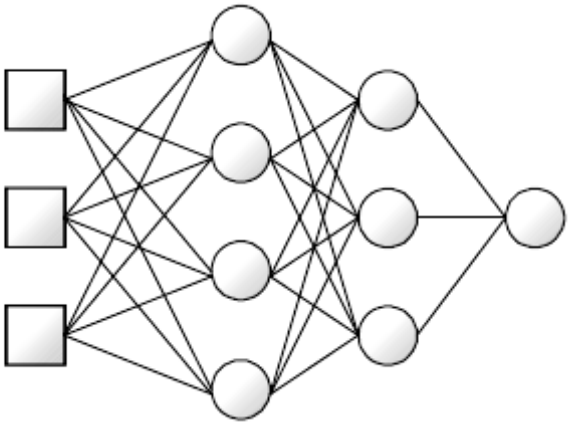
VIII.2.2.4. Les réseaux multicouches ou perceptron multicouches (PMC)

Les réseaux multicouches (PMC) (figure 10) se composent des entrées, une couche de sortie et zéro ou plusieurs couches cachées [82]. Les connexions sont permises seulement d'une couche inférieure (plus proche des entrées) vers une couche supérieure (plus proche de la couche de sortie). Il est aussi interdit d'avoir des connexions entre des neurones de la même couche.

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

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

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



Les entrées Couches cachées Couche de sortie

Figure 10 : *Structure générale du perceptron multicouches*

Il a été prouvé [85] qu'il existe toujours un réseau de neurones de ce type avec trois couches seulement (les entrées, couche de sortie et une couche cachée) qui peut approximer une fonction f : $[0\text{-}1]^n \Rightarrow R^n$ avec n'importe quelle précision $\varepsilon > 0$ désirée. Un problème consiste à trouver combien de neurones cachés sont nécessaires pour obtenir cette précision. Un autre problème est de s'assurer a priori qu'il est possible d'apprendre cette fonction.

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

VIII.2.2.5. Apprentissage :

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

Il existe trois types d'apprentissages principaux. Ce sont l'apprentissage supervisé, l'apprentissage non-supervisé, et l'apprentissage par tentative (graded training en anglais) [85].

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

Dans le cas de l'apprentissage non-supervisé le réseau décide lui-même quelles sont les bonnes sorties. Cette décision guidée par un but interne au réseau qui exprime une configuration idéale à atteindre par rapport aux exemples introduits. Les cartes auto- organisatrices de Kohonen sont un exemple de ce type de réseau [84].

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

Il existe plusieurs règles pour chaque type d'apprentissage. L'apprentissage supervisé est le type le plus utilisé. Pour ce type d'apprentissage la règle la plus exploitée est celle de Widrow-Hof. D'autres règles d'apprentissage sont par exemple la règle de Hebb, la règle de perceptron, la règle de Grossberg etc …. [82, 85, 86].

VIII.2.2.5.1. L'apprentissage de Widrow-Hof :

La règle d'apprentissage de Widrow-Hof est une règle qui permet d'ajuster les poids d'un réseau de neurones pour diminuer à chaque étape l'erreur commise par ce réseau de neurones (à condition que le facteur d'apprentissage soit bien choisi).
Un poids est modifié en utilisant la formule suivante :

$$w_{k+1} = w_k - \alpha \delta_k x_k \tag{79}$$

Où :

w_k est le poids à l'instant k ;

w_{k+1} le poids à l'instant k+1 ;

α est le facteur d'apprentissage ;

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

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

Ainsi, si δ_k et x_k sont positifs tous les deux, alors le poids doit être augmenté. L'ampleur du changement dépend avant tout de la grandeur de δ_k mais aussi de celle de x_k . Le coefficient

α sert à diminuer les changements pour éviter qu'ils deviennent trop grands, ce qui peut entraîner des oscillations du poids.

Deux versions améliorées de cet apprentissage existent, la version « par lois » et la version « par inertie » (momentum en anglais) [85], dont l'une utilise plusieurs exemples pour calculer la moyenne des changements requis avant de modifier le poids et l'autre empêche que le changement du poids au moment k ne devienne beaucoup plus grand qu'au moment k-1.

VIII.2.2.5.2. L'apprentissage par rétro-propagation du gradient (Levenberg-Marquardt back-propagation)

L'algorithme d'apprentissage par rétro-propagation du gradient (figure 11) est un algorithme itératif qui a pour objectif de trouver le poids des connexions minimisant l'écart commis par le réseau sur l'ensemble d'apprentissage. Cette minimisation par une méthode du gradient conduit à l'algorithme d'apprentissage de rétro-propagation.

La procédure d'apprentissage se décompose en deux étapes. Pour commencer, les valeurs d'entrées sont présentées au réseau, qui propage ensuite ces valeurs jusqu'à la couche de sortie et donne ainsi la réponse au réseau. A la deuxième étape les bonnes sorties correspondantes sont présentées aux neurones de la couche de sortie qui calculent l'écart, modifient leurs poids et rétro-propagent l'erreur jusqu'aux entrées pour permettre aux neurones cachés de modifier leurs poids de la même façon. Le principe de modification des poids est normalement l'apprentissage de Widrow-Hoff.

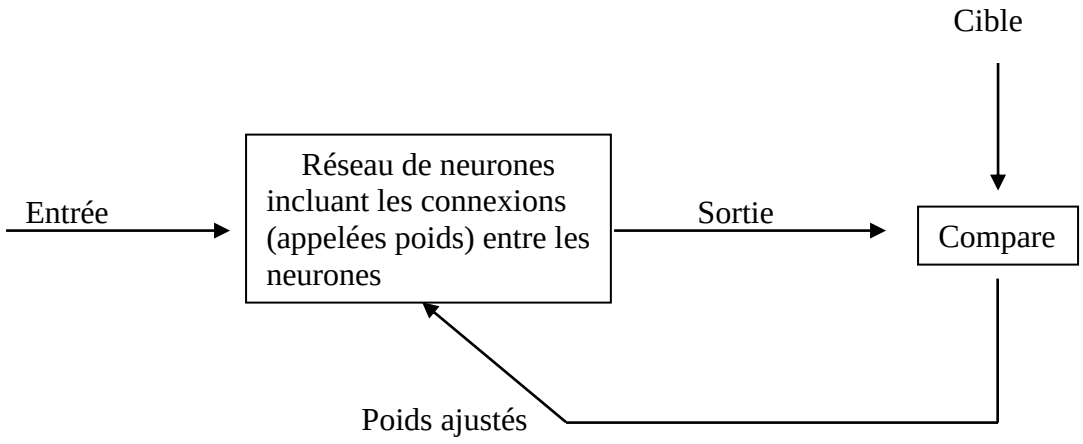


Figure11: *Apprentissage par un algorithme de rétro-propagation*

Généralement pour le calcul de l'écart on utilise l'erreur quadratique moyenne *MSE* (*Mean Square Error*) définie par la relation :

$$MSE = \frac{\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2}{n} \qquad (80)$$

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

VIII.2.2.6. Critères d'arrêt

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

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

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

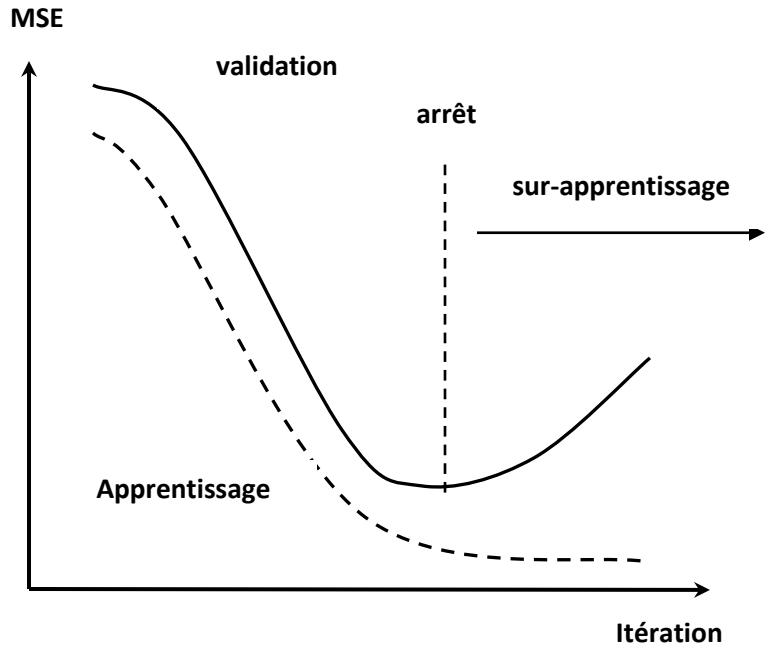


Figure 12 : *Illustration de l'arrêt précoce*

VIII.2.2.7. Construction d'un modèle

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

Les quatre grandes étapes de la création d'un réseau de neurones sont détaillées ci-après :

VIII.2.2.7.1. Construction de la base de données

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

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

- Un ensemble de 125 composés pour l'apprentissage du réseau de neurones.
- Un deuxième de 40 composés pour la validation.
- Et un troisième de 40 composés choisis aléatoirement de l'ensemble d'apprentissage pour le test.

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

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

$$x_{norm} = 2 \times \frac{\left(x_j - x_{\min}\right)}{\left(x_{\max} - x_{\min}\right)} - 1 \qquad (81)$$

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

VIII.2.2.7.2. Définition de la structure du réseau

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

VIII.2.2.7.3. Nombre de couches et de neurones cachés

Les entrées et la couche de sortie mises à part, il faut décider du nombre de couches cachées. Sans couche cachée, le réseau n'offre que de faibles possibilités d'adaptation. Néanmoins, il a été démontré qu'un Perceptron Multicouches avec une seule couche cachée pourvue d'un nombre suffisant de neurones, peut approximer n'importe quelle fonction avec la précision souhaitée [87].

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

VIII.2.2.7.4. Présentation de l'environnement utilisé

Dans cette optique, le logiciel MATLAB [88], qui contient un module consacré au développement de réseaux de neurones, a été retenu ; un PC Dell P4 avec une Ram de 512 et une vitesse de 3,4 GHZ a été utilisé.

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

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

La fonction sigmoïde de transfert, tangente hyperbolique, a été adoptée comme fonction d'activation pour les couches cachées et de sortie.

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

L'ajustement des modèles QSPR peut être déterminé par le coefficient de détermination multiple R^2 et la racine de l'erreur quadratique moyenne RMSE (Root Mean-Squared Error).

Ces paramètres sont calculés sur l'ensemble de calibrage et ils sont utilisés pour décider si le modèle possède la qualité prédictive reflétée dans le R^2. L'utilisation de la RMSE montre l'erreur entre la moyenne des valeurs expérimentales et prédites.

- Le coefficient de détermination multiple :

$$R^2 = 1 - \frac{SCE}{SCT} = 1 - \frac{\sum_1^n (y_i - \hat{y}_i)^2}{\sum_1^n (y_i - \bar{y})^2} \qquad (82)$$

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

- La racine de l'erreur quadratique moyenne de calibrage (désignée également par SDEC) :

$$SDEC = \sqrt{\frac{\sum_{i=1}^{n} \left(y_i - \hat{y}_i \right)^2}{n}} \qquad (83)$$

IX – 1 Robustesse du modèle

La stabilité du modèle a été explorée en utilisant la validation croisée, cette dernière est considérée comme une validation interne qui consiste à mesurer sa capacité à corréler la propriété avec les descripteurs quand on modifie légèrement les données (suppression d'une ou plusieurs données). Il existe plusieurs méthodes de validation croisée : LOO (*Leave One Out*) [89] et LMO (*Leave Many Out*) [90].

Dans le cas du Leave One Out (LOO), une seule observation du jeu d'entraînement est retirée et les coefficients de la régression sont optimisés sur les n-1 autres données. La propriété prédite $\hat{y}_{(i)}$ est recalculée à partir de cette nouvelle équation pour le composé isolé. Cette manipulation est effectuée pour les n hydrocarbures du jeu d'entraînement, puis le coefficient de prédiction noté Q^2 est calculé à l'aide de l'équation suivante :

$$Q^2_{LOO} = \frac{SCT - PRESS}{SCT} \qquad (84)$$

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

$$PRESS = \sum_1^n \left(y_i - \hat{y}_{(i)}\right)^2 \qquad (85)$$

Et le SDEP par :

$$\sigma_N = SDEP = \sqrt{\frac{\sum_{i=1}^{n}\left(y_i - \hat{y}_{(i)}\right)^2}{n}} = \sqrt{\frac{PRESS}{n}} \qquad (86)$$

Contrairement à R^2, qui augmente avec le nombre de paramètres de la régression, le facteur Q^2_{LOO} affiche une courbe avec maximum (ou avec palier), obtenu pour un certain nombre de variables explicatives, puis décroît par la suite de façon monotone. Ce fait confère une grande importance au coefficient Q^2_{LOO}. Une valeur de $Q^2_{LOO} > 0{,}5$ est, généralement, considérée comme satisfaisante, et une valeur supérieure à 0,9 est excellente [91].

Dans le cas du Leave Many Out (LMO), un groupe de molécules du jeu d'entraînement est retiré au lieu d'une seule observation. Une faible valeur de Q^2 implique que le modèle n'est pas robuste et ne sera pas prédictif, mais la réciproque n'est pas nécessairement vraie [92]. En effet, le modèle est considéré comme robuste quand les différents coefficients de prédiction Q^2 ont des valeurs très proches et quand la différence entre les Q^2 et le R^2 est faible.

IX – 2 Test de randomisation

Ce test permet de mettre en évidence des corrélations dues au hasard. Il consiste à générer un vecteur « propriété considérée » par permutation aléatoire des composantes du vecteur réel

(Figure 13). On calcule alors sur le vecteur obtenu (considéré comme vecteur expérimental réel) un modèle QSPR, selon la méthode habituelle. Ce procédé est répété plusieurs fois (100 dans notre cas).

Deux méthodes semblent exister : celle qui considère la permutation des descripteurs également [93-94] et celle qui ne le fait pas [95]. Dans ce travail, pour une raison pratique (comme la difficulté à automatiser la sélection des descripteurs) la sélection des descripteurs n'a pas été prise en compte lors de la randomisation.

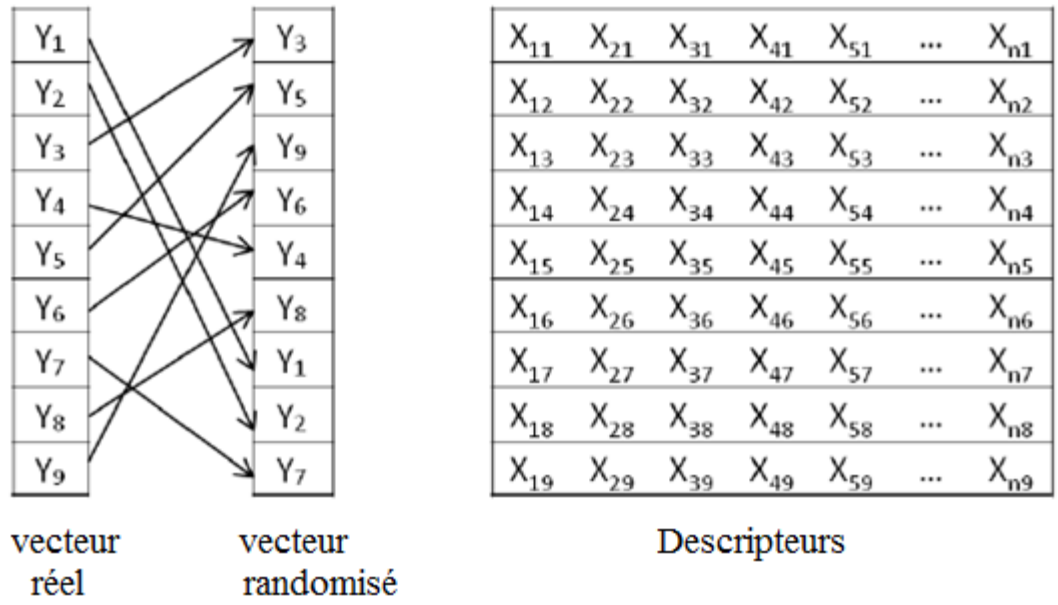


Figure 13 : Illustration de la méthode du test de randomisation

IX – 3 Validation externe

La meilleure façon d'estimer la véritable puissance prédictive d'un modèle QSPR est de comparer les valeurs prédites et observées d'un ou de plusieurs composés « ensemble de validation » qui ne sont pas utilisés dans le développement du modèle [82, 83].

La mesure de la prédictivité la plus utilisée est le $R^2_{CV,ext}$ ou Q^2_{ext} défini par la relation suivante :

$$\boldsymbol{R}^2_{\mathrm{CV,\,ext}} = 1 - \frac{\sum_{i=1}^{n_{ext}} \left(y_i - \hat{y}_i\right)^2}{\sum_{i=1}^{n_{ext}} \left(y_i - \bar{y}_{TR}\right)^2} \qquad (87)$$

De même, l'autre évaluateur (RMSE) de la prédictibilité peut être calculé pour le jeu de validation selon la relation :

$$\boldsymbol{SDEP_{ext}} = \sqrt{\frac{\sum_{i=1}^{n_{ext}} \left(\boldsymbol{y_i} - \hat{\boldsymbol{y}}_{(i)} \right)^2}{\boldsymbol{n_{ext}}}} \qquad (88)$$

L'article de Chirico et Gramatica [96] répertorie différents coefficients de calcul de la prédictivité présentés ci-après.

Le coefficient Q^2F1 proposé par Tropsha [92, 97, 98] n'est pas une vraie validation externe car il fait intervenir des informations sur le jeu de calibrage.

$$\boldsymbol{Q^2_{F1}} = 1 - \frac{\sum_{i=1}^{n_{ext}} (\hat{y}_i - y_i)^2}{\sum_{i=1}^{n_{ext}} (y_i - \bar{y}_{TR})^2} \qquad (89)$$

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

En 2008, une autre mesure de la prédictivité, proposée par Schüürmann [99], est le Q^2F2 qui se différencie de Q^2F1 par le fait que la moyenne utilisée au dénominateur est celle du jeu de validation et non celle du jeu d'entraînement : il s'agit donc bien d'une validation externe car aucune donnée du jeu d'entraînement n'est nécessaire. De plus, Q^2F1 est plus optimiste car supérieur ou égal à Q^2F2 et par conséquent accepte plus facilement les modèles. Le risque d'avoir un modèle non prédictif accepté est moins grand avec Q^2F2.

$$\boldsymbol{Q^2_{F2}} = 1 - \frac{\sum_{i=1}^{n_{ext}} (\hat{y}_i - y_i)^2}{\sum_{i=1}^{n_{ext}} (y_i - \bar{y}_{EXT})^2} \qquad (90)$$

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

En 2009, le coefficient Q^2F3 a été proposé par Consonni [100] afin de supprimer le biais introduit par la distribution des données. De plus, selon Consonni, l'absence d'information sur le jeu d'entraînement est un désavantage. En effet, il a été observé que la valeur de Q^2F3 est identique quelle que soit la distribution du jeu de validation. Il semble également être insensible

au nombre de composés. En effet, la valeur de Q^2F3 ne change pas avec la taille du jeu de validation, contrairement à Q^2F2 dont la valeur augmente avec le nombre de composés. Cependant, tout comme Q^2F1, ce coefficient n'est pas une vraie validation externe car il fait intervenir des informations sur le jeu d'entraînement.

$$Q^2_{F3} = 1 - \frac{\left[\sum_{i=1}^{n_{EXT}} (\hat{y}_i - y_i)^2\right] / n_{EXT}}{\left[\sum_{i=1}^{n_{TR}} (y_i - \bar{y}_{TR})^2\right] / n_{TR}} \quad (91)$$

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

Le dernier coefficient CCC [101, 102] mesure à la fois la précision (distance par rapport à l'équation) et la justesse (c'est-à-dire à quel point la ligne de la régression dévie de la droite x=y dite « concordance line »). Il s'agit d'une validation externe car aucune information du jeu d'entraînement n'est nécessaire

$$CCC = \frac{2\sum_{i=1}^{n} (x_i - \bar{x}_i)(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x}_i)^2 + \sum_{i=1}^{n} (y_i - \bar{y})^2 + n(\bar{x} - \bar{y})^2} \quad (92)$$

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

Une validation externe supplémentaire selon [96] est appliquée uniquement à l'ensemble de validation. Selon les critères recommandés de Tropsha et *al.*, un modèle QSPR prédictif, doit remplir les conditions suivantes:

1) $Q^2_{EXT} > 0.5$ (93-a)

2) $R^2 > 0.6$ (93-b)

3) $(R^2 - R_0^2)/R^2 < 0.1$ et $0.85 < k < 1.15$ (93-c)

$(R^2 - R'^2_0)/R^2 < 0.1$ et $0.85 < k' < 1.15$ (93-d)

où

$$R = \frac{\sum(y_i - \bar{y})(\tilde{y}_i - \bar{\tilde{y}})}{\sqrt{\sum(y_i - \bar{y})^2 \sum(\tilde{y}_i - \bar{\tilde{y}})^2}} \quad (94\text{-a})$$

$$R_0^2 = 1 - \frac{\sum(y_i - y_i^{r_0})^2}{\sum(y_i - \bar{y})^2} \quad (94\text{-b}) \qquad R_0'^2 = 1 - \frac{\sum(\tilde{y}_i - \tilde{y}_i^{r_0})^2}{\sum(\tilde{y}_i - \bar{\tilde{y}})^2} \quad (94\text{-c})$$

$$k = \frac{\sum(y_i \tilde{y}_i)}{\sum(\tilde{y}_i)^2} \quad (94\text{-d}) \qquad k' = \frac{\sum(y_i \tilde{y}_i)}{\sum(y_i)^2} \quad (94\text{-e})$$

R est le coefficient de corrélation entre les valeurs calculées et expérimentales dans l'ensemble de test; R_0^2 (valeurs calculées par rapport à celles observées) et $R_0'^2$ (valeurs observées par rapport à celles 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 à celles calculées, respectivement; $y_i^{r_0}$ et $\tilde{y}_i^{r_0}$ sont définis respectivement par : $y_i^{r_0} = k\,\tilde{y}$ et, $\tilde{y}_i^{r_0} = k'\,y$; les sommations portent sur tous les échantillons de l'ensemble de test.

La validation est en évolution permanente avec l'utilisation de nouveaux coefficients. De manière générale, les coefficients R^2 et Q^2 doivent avoir des valeurs proches de 1 (de préférence supérieures à 0,6) et leur différence doit être faible pour considérer le modèle comme robuste. Cependant, l'évaluation des coefficients doit se faire au regard de la taille de la base de données (notamment pour R^2) et de l'ordre de grandeur de l'incertitude expérimentale (RMSE). Mais d'autres paramètres sont pris en considération pour le choix du modèle comme la possibilité d'interprétation des descripteurs.

PARTİE APPLİCATİON

Nous traiterons séparément chacune des propriétés physiques visées dans notre travail (Teb, Tc, Pc), en adoptant systématiquement les mêmes ensembles d'estimation et de validation.

Nous testerons les deux approches hybrides : algorithme génétique / régression linéaire multiple (AG/RLM) ; algorithme génétique / réseau de neurones artificiel (AG/RNA).

Les descripteurs moléculaires sélectionnés par AG/RLM seront également exploités pour le modèle non linéaire.

I-Modélisation de la température critique

I-1- Introduction

Le RNA est une approche non linéaire (exponentielle, logarithmique, polynomiale...) pour calculer le modèle mathématique qui explique au mieux la variabilité d'une propriété en fonction des descripteurs moléculaires.

Le modèle le plus performant est approché en opérant des essais préliminaires. Ainsi l'architecture du RNA est modifiée progressivement en jouant sur le nombre de couches cachées, ou sur celui des neurones cachés et/ou sur le nombre de cycles d'entraînements c'est-à-dire le nombre d'itérations ; la fonction d'activation exploitée est de type sigmoïde.

La base de données utilisée a été scindée aléatoirement en deux parties, l'une pour le calibrage des modèles (MLR et RNA) et la détermination des paramètres statistiques qui représente 75 % de la taille totale de la base de données et l'autre pour la validation (~25%). Les critères de performance sont calculés aussi bien en mode de calibrage qu'en mode de validation.

Rappelons qu'un troisième ensemble de test (~25%) interne sera choisi aléatoirement pour l'apprentissage du réseau de neurones. Les variables étant caractérisées par des valeurs différentes, nous les avons normalisées afin de ramener la plage d'évolution des valeurs prises par les variables à l'intérieur d'un intervalle standardisé, fixé a priori. Elle est souhaitable car elle évite au système de se paramétrer sur une plage de valeurs particulières, ignorant ainsi les valeurs extrêmes. Pour notre cas, nous avons normalisé les données entre -1 et +1.

I-2-Résultats et discussion

I-2-1- Régression linéaire multiple

Sélection des descripteurs

Le modèle retenu doit présenter une corrélation suffisante et, en même temps, protéger contre toute surparamétrisation, ce qui conduirait à une perte du pouvoir de prédiction pour les échantillons externes à l'ensemble de calibrage.

D'un point de vue statistique le rapport entre le nombre d'échantillons (n) et le nombre de descripteurs (p) ne doit pas être trop faible. Habituellement un descripteur pour, au moins, cinq (5) observations est recommandé.

La valeur optimale de p a été obtenue par la méthode du "point de brisure" qui consiste à analyser l'amélioration de la corrélation avec l'augmentation de la dimension du modèle. La représentation des valeurs de R^2 en fonction du nombre de descripteurs (Figure 14) fait ressortir un comportement asymptotique, et l'amélioration de la corrélation devient moins significative après un certain rang ($\Delta R^2 < 0{,}02\text{-}0{,}03$) ; en ce point (le "point de brisure"), le modèle est considéré comme optimal, c'est-à-dire représentant le meilleur compromis corrélation / paramétrisation.

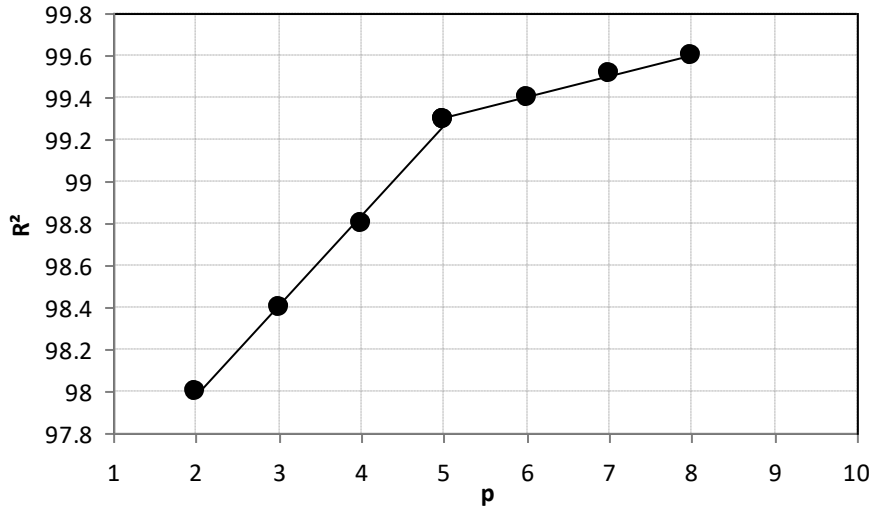


Figure 14: Variation du R^2 en fonction de nombre de descripteurs pour Tc**.**

La figure 14 met en évidence que, vraisemblablement, parmi les descripteurs pouvant être en relation avec la température critique, 5 sont mieux adaptés pour la modélisation par MLR.

Parmi les modèles obtenus par algorithme génétique, notre choix s'est porté sur celui caractérisé par les valeurs de Q^2= 99,25 % et R^2= 99,30 %. Les descripteurs entrant dans le modèle, leurs valeurs, classes et de brèves définitions sont donnés dans les tableaux I et II [cf : annexes].

Les cinq descripteurs ont été obtenus en utilisant le logiciel Dragon. On trouvera plus d'informations concernant ces descripteurs dans le guide de l'utilisateur du logiciel Dragon [63] et les références afférentes.

Equation de régression :

L'équation (95) reproduit la régression établie:

$$Tc = 65{,}00(\pm 7{,}78) + 198{,}70(\pm 2{,}23)\ \mathbf{piPC01} + 53{,}54(\pm 3{,}21)\ \mathbf{R2e} + 237{,}94(\pm 26{,}83)\mathbf{R4m} - 53{,}71(\pm 5{,}46)\ \mathbf{SIC2} - 8{,}83(\pm 0{,}97)\ \mathbf{GGI1} \quad (95)$$

Avec les paramètres statistiques suivants :

R^2= 99,30 Q^2_{LOO} = 99,17 EQMC= 8,51 EQMP = 9,19

n= 125 s = 8,72 F= 3275,13

n_{ext}= 40 Q^2_{ext} = 99,42 $EQMP_{ext}$ = 7,65

Les statistiques calculées établissent la pertinence du modèle. En effet, la valeur de R^2 signifie que 99,30 % de la variabilité de Tc est expliquée par les 5 descripteurs sélectionnés. Par contre, les valeurs des écarts quadratiques moyens de prédiction (EQMP) et de calcul (EQMC) sont proches mais sont quelque peu supérieures à 5 K, erreur admise pour les Tc calculées. En outre, le modèle est très hautement significatif (grande valeur du paramètre de Fisher : F=3275,13). Le coefficient de prédiction Q^2, supérieur à 99 %, indique un modèle robuste, c'est-à-dire dont les paramètres ne changent pas beaucoup lorsqu'on utilise d'autres ensembles de calibrage extraits de la population totale.

Selon les valeurs du test t (| t |), on peut classer les descripteurs sélectionnés dans ce modèle d'après leur pourcentage de contribution qui se présente dans l'ordre: piPC01 > R2e > SIC2 > GGI1 > R4m. Les valeurs des VIF (< 5) suggèrent que ces descripteurs sont faiblement corrélés les uns avec les autres. Ainsi, le modèle peut être considéré comme une équation de régression optimale.

Descripteur	X	Dx	T	probabilité-t	VIF
Constante	65,00	7,78	8,36	0,000	
piPC01	198,70	2,23	89,23	0,000	1,48
R2e	53,54	3,21	16,70	0,000	1,66
SIC2	-53,71	5,46	-9,84	0,000	1,26
GGI1	-8,83	0,97	-9,07	0,000	2,33
R4m	237,94	26,83	8,87	0,000	3,35

Les valeurs de la température critique mesurées, calculées et prédites pour les deux ensembles (calibrage et validation), ainsi que les valeurs des leviers et des erreurs standardisées sont regroupées dans le tableau 3.

Tableau - 3: Valeurs : des Tc expérimentales, calculées, prédites, de h_{ii}, et e_{istd}

N	**Composé**	**Tc- Exp**	**Tc-Calc**	**hii**	**ei**	**ei std**
1	Méthane	190,50	182,74	0,363	-7,76	-1,12
2	Ethane	305,40	287,84	0,138	-17,56	-2,17
3	Propane	369,80	353,92	0,068	-15,88	-1,89
4	Butane	425,20	414,26	0,039	-10,94	-1,28
5	Isopentane	460,40	452,68	0,025	-7,72	-0,9
6	2,2-Diméthylbutane	488,80	482,53	0,046	-6,27	-0,74
7	2,3-Diméthylbutane	500,00	504,92	0,022	4,92	0,57
8	2-Méthylpentane	497,70	496,94	0,015	-0,76	-0,09
9	3- Méthylpentane	504,50	495,96	0,014	-8,54	-0,99
10	Hexane	507,70	503,45	0,018	-4,25	-0,49
11	2,2-Diméthylpentane	520,50	525,49	0,024	4,99	0,58
12	2,4-Diméthylpentane	519,80	536,03	0,018	16,23	1,88
13	2,2,3-Triméthylbutane	531,20	524,01	0,043	-7,19	-0,84
14	2-Méthylhexane	530,40	527,56	0,013	-2,84	-0,33
15	3-Méthylhexane	535,30	532,05	0,015	-3,25	-0,38
16	3-Ethylpentane	540,70	546,98	0,049	6,28	0,74

Tableau - 3: (suite)

N	Composé	Tc- Exp	Tc-Calc	hii	ei	ei std
17	Heptane	540,30	536,14	0,015	-4,16	-0,48
18	2,2,4-Triméthylpentane	544,00	558,86	0,045	14,86	1,74
19	2,2,3,3-Tétraméthylbutane	567,80	556,1	0,088	-11,7	-1,41
20	2,2-Diméthylhexane	549,90	549,55	0,032	-0,35	-0,04
21	2,5-Diméthylhexane	550,10	553,31	0,025	3,21	0,37
22	2,4-Diméthylhexane	553,60	560,91	0,019	7,31	0,85
23	2,2,3-Triméthylpentane	563,50	560,01	0,038	-3,49	-0,41
24	3,3-Diméthylhexane	562,10	557,57	0,026	-4,53	-0,53
25	2,3,3-Triméthylpentane	573,60	557,97	0,039	-15,63	-1,83
26	3-Ethyl-3-méthylpentane	576,60	570,79	0,055	-5,81	-0,69
27	3-Ethylhexane	565,50	573,16	0,042	7,66	0,9
28	3-Méthylheptane	563,70	564,07	0,016	0,37	0,04
29	2,2,4,4-Tétraméthylpentane	574,70	587,11	0,105	12,41	1,5
30	2,2-Diméthylheptane	576,80	571,22	0,047	-5,58	-0,66
31	2,2,3,4-Tétraméthylpentane	592,70	596,76	0,076	4,06	0,48
32	2,2,3,3-Tétraméthylpentane	607,70	587,88	0,076	-19,82	-2,36
33	2-Méthyloctane	587,00	579,5	0,017	-7,5	-0,87
34	Nonane	594,60	590,54	0,018	-4,06	-0,47
35	2,2,5,5-Tétraméthylhexane	581,60	582,86	0,204	1,26	0,16
36	Décane	617,70	614,25	0,022	-3,45	-0,4
37	Undécane	638,80	635,02	0,026	-3,78	-0,44
38	Dodécane	658,20	654,52	0,031	-3,68	-0,43
39	Tridécane	676,00	672,04	0,036	-3,96	-0,46
40	Tétradécane	693,00	688,88	0,041	-4,12	-0,48
41	Hexadécane	722,00	718,41	0,051	-3,59	-0,42
42	Octadécane	748,00	744,42	0,062	-3,58	-0,42
43	2,2,4-Triméthylhexane	573,70	579,96	0,044	6,26	0,73

Tableau - 3: (suite)

N	Composé	Tc- Exp	Tc-Calc	hii	ei	ei std
44	3,3-Diethylpentane	610,00	618,19	0,172	8,19	1,03
45	Pentadecane	707,00	704,05	0,046	-2,95	-0,35
46	Eicosane	767,00	767,11	0,073	0,11	0,01
47	E-But-2-ène	428,60	444,06	0,023	15,46	1,79
48	Z-But-2-ène	435,60	452,24	0,023	16,64	1,93
49	Pent-1-ène	464,80	468,25	0,042	3,45	0,4
50	2-Méthyl but-1-ène	465,00	462,99	0,036	-2,01	-0,23
51	E-pent-2-ène	475,00	472,68	0,034	-2,32	-0,27
52	Z-pent-2-ène	476,00	480,3	0,039	4,3	0,5
53	2-Méthyl but-2-ène	470,00	479,39	0,018	9,39	1,09
54	Hex-1-ène	504,10	506,1	0,035	2	0,23
55	Hept-1-ène	537,30	537,18	0,028	-0,12	-0,01
56	Propylène	364,90	374,04	0,048	9,14	1,07
57	Isobutylène	417,90	417,64	0,042	-0,26	-0,03
58	3-Méthyl but-1-ène	450,00	465,74	0,022	15,74	1,83
59	Z-Hex-2-ène	518,00	507,73	0,039	-10,27	-1,2
60	E-Hex-2-ène	516,00	511,66	0,041	-4,34	-0,51
61	Z-Hex-3-ène	517,00	518,93	0,014	1,93	0,22
62	E-Hex-3-ène	519,90	524,8	0,015	4,9	0,57
63	2-Méthyl pent-2-ène	518,00	510,76	0,018	-7,24	-0,84
64	Z-3-Méthyl pent-2-ène	518,00	509,5	0,021	-8,5	-0,99
65	E-3-Méthyl pent-2-ène	521,00	519,66	0,026	-1,34	-0,16
66	Z-4-Méthyl pent-2-ène	490,00	512,05	0,018	22,05	2,55
67	E-4-Méthyl pent-2-ène	493,00	515,98	0,019	22,98	2,66
68	2,3-Diméthyl but-2-ène	524,00	524,97	0,037	0,97	0,11
69	2,3-Diméthyl but-1-ène	501,00	506,58	0,02	5,58	0,65
70	3,3-Diméthyl but-1-ène	490,00	493,17	0,048	3,17	0,37

Tableau - 3: (suite)

N	Composé	Tc- Exp	Tc-Calc	hii	ei	ei std
71	2,3,3-Triméthyl but-1-ène	533,00	529,19	0,043	-3,81	-0,45
72	Non-1-ène	592,00	591,1	0,021	-0,9	-0,1
73	Undec-1-ène	637,00	635,45	0,023	-1,55	-0,18
74	Dodec-1-ène	657,00	654,79	0,026	-2,21	-0,26
75	Tridec-1-ène	674,00	672,57	0,03	-1,43	-0,17
76	Hexadec-1-ène	717,00	718,84	0,045	1,84	0,22
77	Octadec-1-ène	739,00	744,93	0,056	5,93	0,7
78	Cyclobutane	460,00	471,03	0,061	11,03	1,31
79	Cyclopentane	511,70	517,32	0,064	5,62	0,67
80	Méthylcyclopentane	532,70	526,98	0,033	-5,72	-0,67
81	Cyclohexane	553,50	561,98	0,054	8,48	1
82	Méthylcyclohexane	572,20	576,76	0,023	4,56	0,53
83	Ethylcyclopentane	569,50	577,17	0,022	7,67	0,89
84	E-1,4-Diméthylcyclohexane	587,70	598,24	0,022	10,54	1,22
85	Cyclooctane	647,20	645,74	0,125	-1,46	-0,18
86	Cyclopropane	397,80	416,16	0,064	18,36	2,18
87	1,1-Diméthylcyclopentane	547,00	547,31	0,04	0,31	0,04
88	Z-1,2-Diméthylcyclopentane	564,80	555,3	0,016	-9,5	-1,1
89	E-1,2-Diméthylcyclopentane	553,20	561,39	0,023	8,19	0,95
90	E-1,2-Diméthylcyclohexane	596,00	602,32	0,028	6,32	0,73
91	Z-1,3-Diméthylcyclohexane	591,00	595,29	0,023	4,29	0,5
92	Ethylcyclohexane	609,00	611,82	0,032	2,82	0,33
93	1,1,2-Triméthylcyclopentane	579,50	574,56	0,031	-4,94	-0,58
94	1,1,3-Triméthylcyclopentane	569,50	570,49	0,054	0,99	0,12
95	Propylcyclopentane	603,00	604,36	0,022	1,36	0,16
96	Isopropylcyclopentane	601,00	592,21	0,022	-8,79	-1,02
97	Propylcyclohexane	639,00	636,02	0,035	-2,98	-0,35

Tableau - 3: (suite)

N	Composé	Tc- Exp	Tc-Calc	hii	ei	ei std
98	1,E-3,5-Triméthylcyclohexane	602,20	619,42	0,034	17,22	2,01
99	Butylcyclohexane	667,00	654,29	0,027	-12,71	-1,48
100	Isobutylcyclohexane	659,00	646,29	0,03	-12,71	-1,48
101	sec-Butylcyclohexane	669,00	659,52	0,079	-9,48	-1,13
102	tert-Butylcyclohexane	659,00	647,98	0,055	-11,02	-1,3
103	Hexylcyclopentane	660,10	660,68	0,027	0,58	0,07
104	Heptylcyclopentane	679,00	678,59	0,031	-0,41	-0,05
105	Decylcyclopentane	723,80	724,31	0,042	0,51	0,06
106	Decylcyclohexane	750,00	748,72	0,044	-1,28	-0,15
107	Tridecylcyclopentane	761,00	763,54	0,055	2,54	0,3
108	1-Cyclopentyltetradecane	772,00	774,4	0,06	2,4	0,28
109	1-Cyclopentylpentadecane	780,00	785,15	0,064	5,15	0,61
110	Buta-1,3-diène	425,00	428,95	0,114	3,95	0,48
111	Buta-1,2-diène	443,70	452,16	0,04	8,46	0,99
112	Penta-1,4-diène	478,00	489,73	0,035	11,73	1,37
113	3-Méthyl buta-1,2-diène	496,00	488,21	0,03	-7,79	-0,91
114	E-penta-1,3-diène	496,00	486,63	0,043	-9,37	-1,1
115	Penta-1,2-diène	503,00	496,78	0,04	-6,22	-0,73
116	Propadiène	393,00	379,9	0,193	-13,1	-1,67
117	Deca-1,3-diène	615,00	618,13	0,039	3,13	0,37
118	1-Méthylcyclopentène	542,15	529,28	0,046	-12,87	-1,51
119	Cyclohexène	560,50	554,43	0,029	-6,07	-0,71
120	1-Ethylcyclopentène	576,15	575,36	0,031	-0,79	-0,09
121	Cyclopentène	506,00	511,15	0,039	5,15	0,6
122	Acétylène	308,30	320,66	0,318	12,36	1,72
123	Propyne	402,40	412,04	0,042	9,64	1,13
124	But-2-yne	488,70	482,98	0,029	-5,72	-0,67

Tableau - 3: (suite)

N	Composé	Tc- Exp	Tc-Calc	hii	ei	ei std
125	Vinylacétylène	455,00	432,03	0,147	-22,97	-2,85
126*	Isobutane	408,20	407,17	0,06	-1,03	-0,12
127*	Pentane	469,70	462,45	0,024	-7,25	-0,84
128*	3,3-Diméthylpentane	536,40	520,79	0,026	-15,61	-1,81
129*	2,3-Diméthylpentane	537,40	533,51	0,018	-3,89	-0,45
130*	2,3,4-Triméthylpentane	566,50	579,20	0,06	12,7	1,5
131*	2,3-Diméthylhexane	563,50	561,58	0,016	-1,92	-0,22
132*	3-Ethyl-2-méthylpentane	567,10	575,12	0,071	8,02	0,95
133*	2-Méthylheptane	559,70	554,38	0,014	-5,32	-0,61
134*	3,4-Diméthylhexane	568,90	568,18	0,034	-0,72	-0,08
135*	4-Méthylheptane	561,80	564,07	0,016	2,27	0,26
136*	Octane	568,90	565,41	0,016	-3,49	-0,4
137*	2,2,5-Triméthylhexane	568,00	566,61	0,083	-1,39	-0,17
138*	2,3,3,4-Tétraméthylpentane	607,70	597,22	0,079	-10,48	-1,25
139*	2,2,3,3-Tétraméthylhexane	623,20	610,75	0,071	-12,45	-1,48
140*	3,3,5-Triméthylheptane	609,70	604,21	0,046	-5,49	-0,65
141*	2,2-Diméthylpropane	433,80	431,05	0,12	-2,75	-0,34
142*	2,2,3-Triméthylhexane	588,00	582,78	0,032	-5,22	-0,61
143*	Heptadecane	733,00	731,63	0,056	-1,37	-0,16
144*	Nonadecane	756,00	756,17	0,067	0,17	0,02
145*	Ethylène	282,30	304,46	0,181	22,16	2,81
146*	But-1-ène	419,60	424,63	0,042	5,03	0,59
147*	Oct-1-ène	566,70	565,88	0,023	-0,82	-0,1
148*	E-Oct-2-ène	580,00	565,59	0,031	-14,41	-1,68
149*	Dec-1-ène	615,00	614,39	0,021	-0,61	-0,07
150*	Tétradec-1-ène	689,00	689,09	0,035	0,09	0,01
151*	Pentadec-1-ène	704,00	704,54	0,04	0,54	0,06

Tableau - 3: (suite et fin)

N	Composé	Tc- Exp	Tc-Calc	hii	ei	ei std
152*	Cycloheptane	604,20	603,41	0,059	-0,79	-0,09
153*	1,1-Diméthylcyclohexane	591,00	590,45	0,017	-0,55	-0,06
154*	Z-1,2-Diméthylcyclohexane	606,00	604,70	0,033	-1,3	-0,15
155*	E-1,3-Diméthylcyclohexane	598,00	587,75	0,016	-10,25	-1,19
156*	Z-1,4-Diméthylcyclohexane	598,00	598,24	0,021	0,24	0,03
157*	1-Ethyl-1-méthylcyclopentane	592,00	590,11	0,018	-1,89	-0,22
158*	Isopropylcyclohexane	640,00	632,74	0,05	-7,26	-0,85
159*	Octylcyclopentane	694,00	695,26	0,034	1,26	0,15
160*	Nonylcyclopentane	710,50	710,52	0,038	0,02	0
161*	Dodecylcyclopentane	750,00	751,47	0,049	1,47	0,17
162*	Hexa-1,5-diène	507,00	527,51	0,032	20,51	2,39
163*	2-Méthyl buta-1,3-diène	484,00	477,79	0,038	-6,21	-0,73
164*	But-1-yne	463,70	456,58	0,039	-7,12	-0,83
165*	Pent-1-yne	493,50	497,88	0,04	4,38	0,51

*Composés de validation

Qualité de l'ajustement

La qualité de l'ajustement à été vérifiée en représentant les valeurs calculées Tc-Calc avec notre modèle (colonne3 tableau 3) en fonction des valeurs observées ou expérimentales Tc-Exp (colonne 4 tableau 3). La figue 15 montre une faible dispersion autour de la droite d'ajustement (qui peut être assimilée à la première bissectrice) définie par l'équation (96).

$$\text{Tc-Calc} = 3{,}99 + 0{,}99\ \text{Tc- Exp} \qquad (96)$$

$s = 8{,}58 \qquad R^2 = 99{,}30\ \% \qquad R^2_{ajust} = 99{,}30\ \%$

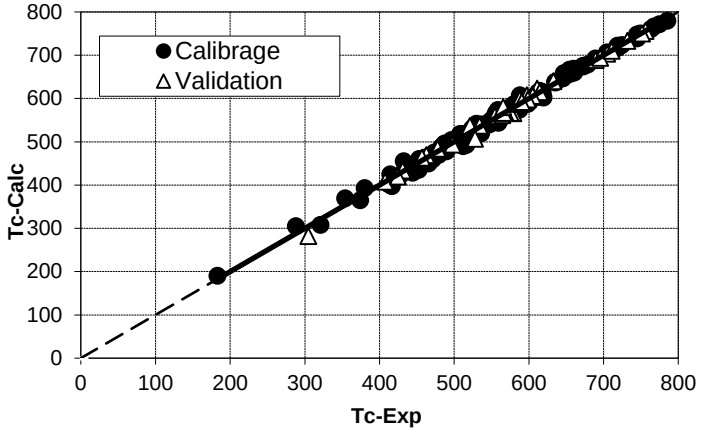


Figure 15: Graphe des valeurs Tc calculées et prédites en fonction des valeurs observées

La représentation des valeurs estimées en fonction de celles observées (figure 15) conduit à l'ordonnée à l'origine (a) et à la pente (b) qui sont établies par la méthode des moindres carrés, également les limites de confiance supérieures et inferieures (a_{sup} / a_{inf} et b_{sup} / b_{inf}) sont récapitulés dans le tableau suivant, compte tenu du nombre (N-2) degrés de liberté.

Ces limites de confiance serviront pour tester si la droite $y = a + bx$ (Eq :96) ne s'écartent pas de façon significative de l'origine [11,07 > a=3,99> -3,08], et si l'hypothèse d'une pente très peu différente de l'unité est vérifiée [1,01 > b=0,99> 0,98].

Le domaine d'application a été discuté à l'aide du diagramme de Williams (figure16) qui représente les résidus de prédiction standardisés en fonction des valeurs des leviers (hi), ième terme diagonal de la matrice de projection : $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\ \boldsymbol{X}'$ où $\boldsymbol{X}$ est la matrice des valeurs observées des variables explicatives et $\boldsymbol{X}'$ sa transposée, ces valeurs sont utiles pour le calcul des résidus caractéristiques. La valeur critique pour déterminer les points leviers correspond à $h^* = \frac{3 \times 6}{125} = \frac{3p}{n} = 0{,}144$. On constate que tous les h_{ii} sont inférieurs à cette valeur critique 0.144, à l'exception de ceux des composés 1, 2, 35, 44, 116, 122 et 125.

La figure 16 qui reproduit la variation des résidus standardisés e_{istd}, en fonction des leviers *hii* des composés ne laisse pas apparaître de points aberrants.

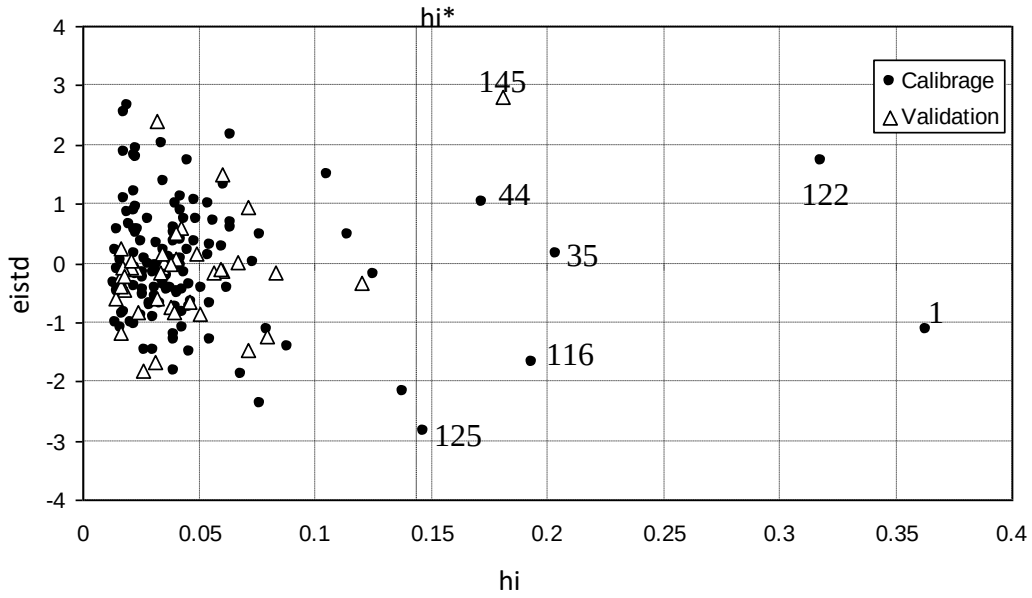


Figure 16: *Diagramme de Williams*

Test de randomisation

Les modèles QSPR, à cause (souvent) de leur complexité et de la sophistication des outils de chimiométrie employés, peuvent constituer une source de corrélations fortuites.

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 températures critiques ont été générés par permutation des positions des composantes du vecteur réel:

$$y = (y_1, y_2, ..., y_{27})' \xrightarrow{RND} y_{RND} = (y_8, y_5, ..., y_2)'$$

et utilisés comme sources d'observations pour des modèles QSPR dans les conditions optimales établies (5 paramètres) .

La figure 17 qui représente le graphe des coefficients statistiques Q^2 et R^2 permet de comparer les résultats obtenus pour les modèles randomisés (cercles noircis) au modèle réel de départ (astérisque).

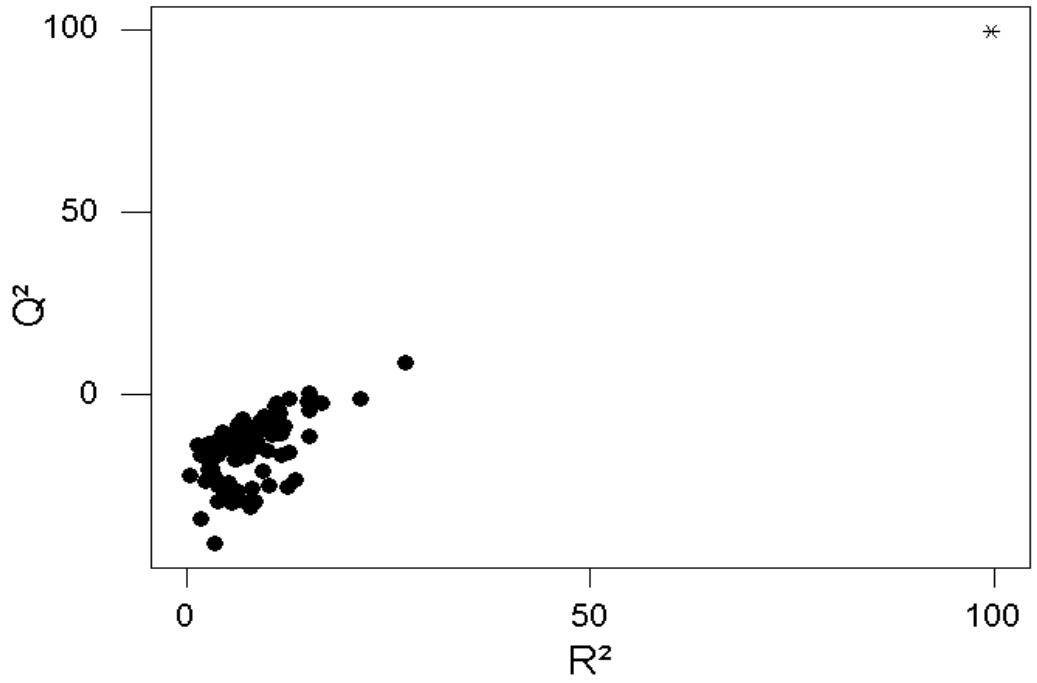


Figure 17 : *Test de randomisation associé au modèle QSPR. Les cercles noircis représentent les températures critiques ordonnées de façon aléatoire, et l'astérisque correspond au modèle réel.*

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

Si de petites valeurs de Q^2_{LOO} indiquent des modèles peu robustes, caractérisés par des faibles capacités prédictives internes, le contraire n'est pas nécessairement vrai. En fait, si une forte valeur de Q^2 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 sur-estimation de la capacité prédictive du modèle, lorsqu'il est appliqué à des composés réellement externes.

Les autres paramètres statistiques qui confirment cette bonne capacité prédictive sont regroupés ci-après :

Paramètre	R^2ext	Q^2_{F1}	Q^2_{F2}	Q^2_{F3}	CCC
Valeur	0,9928	0,9940	0,9936	0,9942	0,9962

Le pouvoir prédictif d'un modèle QSPR peut être vérifié à l'aide des critères de Tropsha (cf : IX-3). Ces résultats montrent le grand pouvoir de prédiction du modèle MLR

1- Q^2_{ext} = 0,9942 > 0,5

2- R^2= 0,9928 > 0,6

3- 0,85 < k =1,0025 < 1,15

4- 0,85 < k'=0,9973 < 1,15

On remarque que tous ces critères sont vérifiés.

I-2-2- Régression par les réseaux de neurones artificiels RNA

Choix des paramètres optimaux

Rappelons que les deux ensembles (calibrage et validation) et les descripteurs sont ceux utilisés pour le modèle RLM. Les descripteurs sont utilisés pour la configuration du réseau de neurones, qui est perfectionnée en phase d'apprentissage ; les paramètres de fonctionnement sont déterminés de façon à obtenir une bonne adéquation entre les valeurs simulées et les données d'apprentissage, combinée à une généralisation correcte de ces simulations.

Choix du nombre de couches cachées

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

Choix du nombre de neurones dans la couche cachée

Le choix de ce paramètre est très important. On fixe a priori un nombre d'itérations (50 par exemple), puis on discrétise l'ensemble des valeurs possibles pour le nombre de neurones cachés (par exemple entre 2 et 20 avec un pas de 2). On fixe la valeur du nombre de neurones de la couche cachée par la valeur minimale de la racine de l'erreur quadratique moyenne RMSE.

Le graphe RMSE = f(nombre de neurones) de la figure suivante permet de visualiser cet impact et de fixer à 8 le nombre de neurones de la couche cachée.

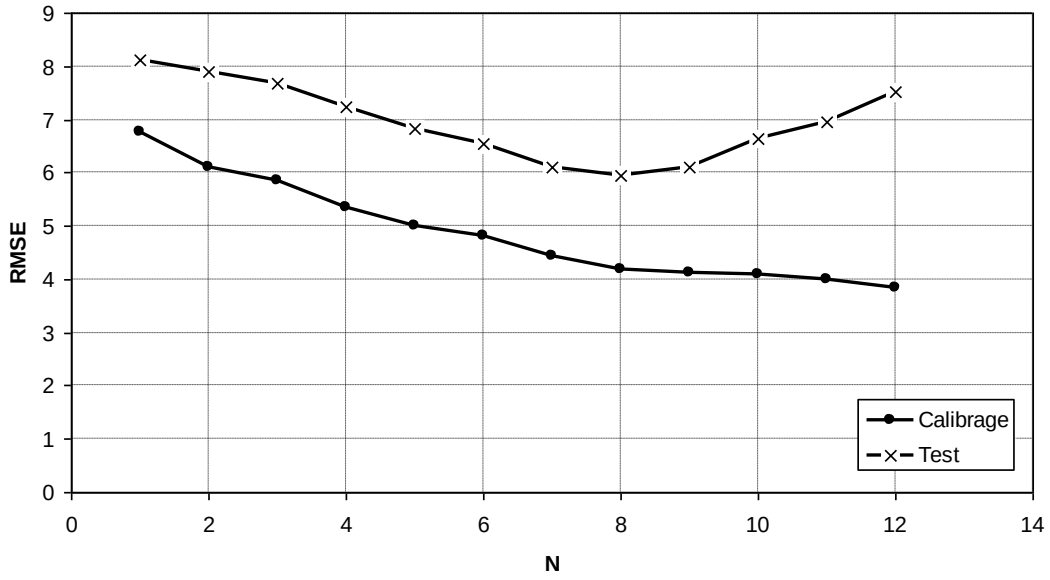


Figure 18 : Choix du nombre de neurones de la couche cachée.

Choix du nombre d'itérations

Après avoir déterminé le nombre de neurones dans la couche cachée, on opère de la même façon pour le nombre d'itérations. La figure 19 qui illustre cette opération, montre que la valeur minimale de la racine de l'erreur quadratique moyenne RMSE autant pour le calibrage que pour le test correspond à l'itération 80.

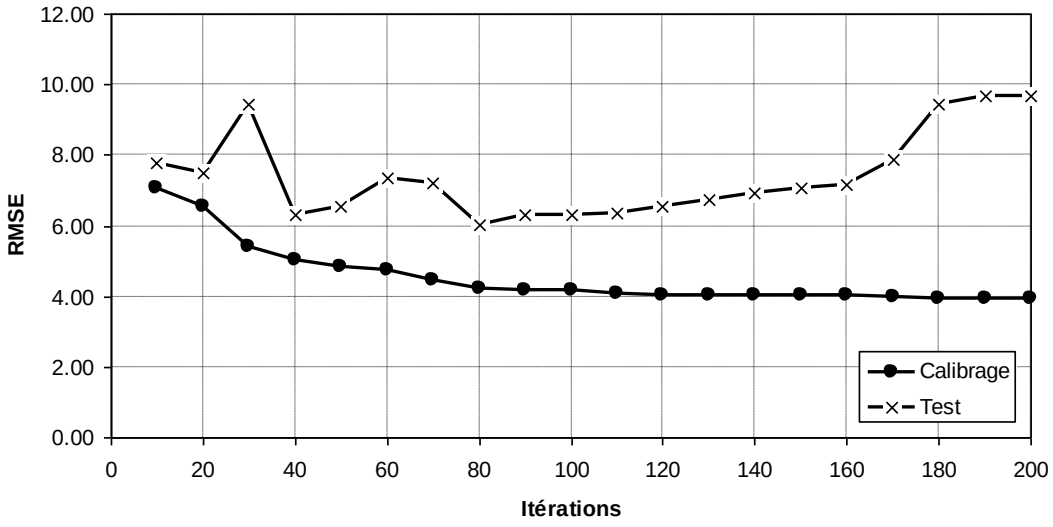


Figure 19 : Choix du nombre d'itérations

Choix de la fonction de transfert

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

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

Choix des paramètres d'apprentissage

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

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

L'apprentissage du réseau de neurones représente un fragile équilibre entre tous ces paramètres, d'où la difficulté pour l'atteindre. Une fois cet apprentissage achevé, le réseau de neurones devient un outil viable et peut être utilisé pour la simulation de nouvelles données. Le tableau 4 précise la structure optimale du réseau de neurones.

Tableau - 4: Structure optimale du réseau de neurones.

Nombre d'entrées	05 (les descripteurs)
Nombre de sorties	01 (la température critique)
Nombre de couches cachées	Une couche cachée
Nombre de neurones dans la couche cachée	08
Nombre d'itérations	80
Algorithme d'apprentissage	Rétro propagation du gradient de l'erreur
Fonctions d'apprentissage	Tangente hyperbolique

Tous les paramètres statistiques obtenus sont présentés dans le tableau 5

Tableau - 5: Valeurs des paramètres statistiques

Ensemble de Calibrage (125) et de			Test (40)	Ensemble de Validation (40)	
R^2 (%)	s	RMSEcal	RMSEtst	R^2cv (%)	RMSEval
99,83	4,19	4,17	5,95	99,59	6,31

La valeur du cœfficient de détermination R^2= 99,93 % explique très bien la variabilité de Tc en fonction des descripteurs choisis ; la petite valeur de la racine de l'erreur quadratique moyenne RMSE ou EQMC = 4,17, indique un modèle très hautement significatif.

Qualité de l'ajustement

La qualité de l'ajustement à été vérifiée en représentant les valeurs calculées Tc-Calc (colonne3 tableau 6) à l'aide de notre modèle en fonction des valeurs observées ou expérimentales Tc-Exp (colonne 4 tableau 6). La figue 20 montre une faible dispersion autour de la droite d'ajustement (qui peut être assimilée à la première bissectrice) définie par l'équation (97).

$$Tc\text{-}Calc = 0{,}01 + 1{,}00\ Tc\text{-}Exp \qquad (97)$$

$$s = 4{,}20 \qquad R^2 = 99{,}80\ \% \qquad R^2_{ajust} = 99{,}80\ \%$$

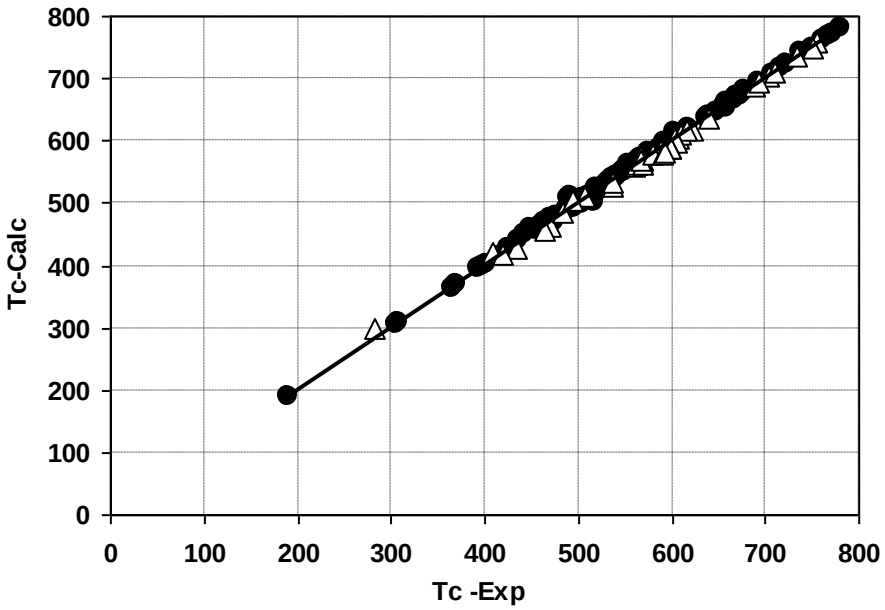
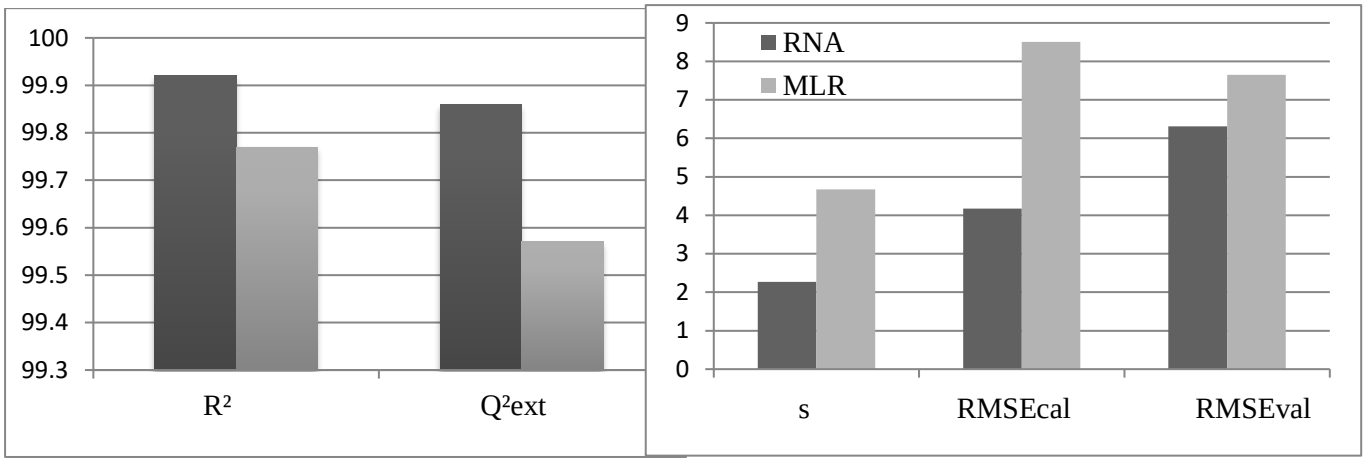


Figure 20: Graphe des valeurs Tc calculées et prédites en fonction des valeurs observées.

D'après les limites de confiance, la pente et l'ordonnée à l'origine se situés dans l'intervalle de ces limites.

D'après l'histogramme suivant on peut constate que le modèle non linéaire est nettement supérieur au modèle linéaire, surtout les erreurs qui sont diminués presque de la moitié à savoir le s et les RMSEs.



La distribution des résidus ordinaires (figure 21) en fonction des valeurs estimées est tout à fait équilibrée et acceptable, puisque 7 résidus de l'ensemble de calibration (soit 5,6 %) sont supérieurs à 2 fois l'erreur standard s ($\geq 2s = 2 \times 4{,}19$), par contre 8 composés de l'ensemble de validation (soit 22,5%) sont supérieures. Notons, enfin, que les 2 plus importants résidus de l'ensemble de calibration et validation valent -16,74 pour le composé 67 (E-4-Méthyl pent-2-ène) et 18,08 pour le composé 145 (Ethylène) respectivement (tableau 6).

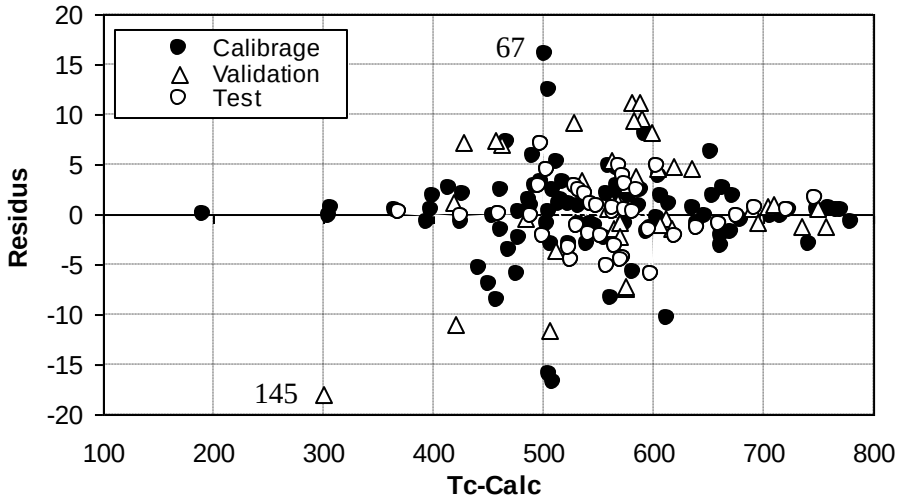


Figure 21 : Variation des résidus en fonction des valeurs calculées.

Tableau - 6: Valeurs des Tc expérimentales, calculées, prédites, et des résidus

N	Composé	Tc- Exp	Tc-Calc	ei
1	Méthane	190,50	190,47	0,03
2	Ethane	305,40	305,51	-0,11
3	Propane	369,80	369,68	0,12
4	Butane	425,20	425,33	-0,13
5	Isopentane	460,40	460,43	-0,03
6	2,2-Diméthylbutane	488,80	488,94	-0,14
7	2,3-Diméthylbutane	500,00	497,19	2,81
8	2-Méthylpentane	497,70	499,88	-2,18
9	3- Méthylpentane	504,50	497,50	7,00
10	Hexane	507,70	503,40	4,30
11	2,2-Diméthylpentane	520,50	525,19	-4,69
12	2,4-Diméthylpentane	519,80	523,14	-3,34
13	2,2,3-Triméthylbutane	531,20	528,41	2,79
14	2-Méthylhexane	530,40	531,52	-1,12
15	3-Méthylhexane	535,30	532,80	2,50
16	3-Ethylpentane	540,70	542,71	-2,01
17	Heptane	540,30	538,21	2,09

Tableau - 6 :(suite)

N	Composé	Tc- Exp	Tc-Calc	Ei
18	2,2,4-Triméthylpentane	544,00	542,96	1,04
19	2,2,3,3-Tétraméthylbutane	567,80	572,19	-4,38
20	2,2-Diméthylhexane	549,90	549,13	0,77
21	2,5-Diméthylhexane	550,10	552,25	-2,15
22	2,4-Diméthylhexane	553,60	558,70	-5,10
23	2,2,3-Triméthylpentane	563,50	563,20	0,30
24	3,3-Diméthylhexane	562,10	565,31	-3,21
25	2,3,3-Triméthylpentane	573,60	568,77	4,83
26	3-Ethyl-3-méthylpentane	576,60	572,77	3,83
27	3-Ethylhexane	565,50	570,14	-4,64
28	3-Méthylheptane	563,70	563,10	0,60
29	2,2,4,4-Tétraméthylpentane	574,70	574,31	0,39
30	2,2-Diméthylheptane	576,80	573,84	2,96
31	2,2,3,4-Tétraméthylpentane	592,70	598,63	-5,93
32	2,2,3,3-Tétraméthylpentane	607,70	602,99	4,71
33	2-Méthyloctane	587,00	584,64	2,36
34	Nonane	594,60	596,28	-1,68
35	2,2,5,5-Tétraméthylhexane	581,60	581,35	0,25
36	Décane	617,70	619,86	-2,16
37	Undécane	638,80	640,21	-1,41
38	Dodécane	658,20	659,12	-0,92
39	Tridécane	676,00	676,17	-0,17
40	Tétradécane	693,00	692,48	0,52
41	Hexadécane	722,00	721,64	0,36
42	Octadécane	748,00	746,43	1,57
43	2,2,4-Triméthylhexane	573,70	574,77	-1,07
44	3,3-Diethylpentane	610,00	608,13	1,87
45	Pentadecane	707,00	707,25	-0,25
46	Eicosane	767,00	766,52	0,48

Tableau - 6 :(suite)

N	Composé	Tc- Exp	Tc-Calc	ei
47	E-But-2-ène	428,60	426,56	2,04
48	Z-But-2-ène	435,60	441,07	-5,47
49	Pent-1-ène	464,80	468,34	-3,54
50	2-Méthyl but-1-ène	465,00	462,62	2,38
51	E-pent-2-ène	475,00	467,89	7,11
52	Z-pent-2-ène	476,00	478,49	-2,49
53	2-Méthyl but-2-ène	470,00	475,94	-5,94
54	Hex-1-ène	504,10	507,16	-3,06
55	Hept-1-ène	537,30	540,25	-2,95
56	Propylène	364,90	364,57	0,33
57	Isobutylène	417,90	415,29	2,61
58	3-Méthyl but-1-ène	450,00	458,64	-8,64
59	Z-Hex-2-ène	518,00	512,79	5,21
60	E-Hex-2-ène	516,00	515,01	0,99
61	Z-Hex-3-ène	517,00	515,61	1,39
62	E-Hex-3-ène	519,90	522,82	-2,92
63	2-Méthyl pent-2-ène	518,00	505,70	12,30
64	Z-3-Méthyl pent-2-ène	518,00	502,02	15,98
65	E-3-Méthyl pent-2-ène	521,00	517,72	3,28
66	Z-4-Méthyl pent-2-ène	490,00	506,09	-16,09
67	E-4-Méthyl pent-2-ène	493,00	509,74	-16,74
68	2,3-Diméthyl but-2-ène	524,00	522,92	1,08
69	2,3-Diméthyl but-1-ène	501,00	497,79	3,21
70	3,3-Diméthyl but-1-ène	490,00	489,20	0,80
71	2,3,3-Triméthyl but-1-ène	533,00	532,18	0,82
72	Non-1-ène	592,00	593,73	-1,73
73	Undec-1-ène	637,00	636,35	0,65
74	Dodec-1-ène	657,00	655,16	1,84
75	Tridec-1-ène	674,00	672,24	1,76

Tableau - 6 :(suite)

N	Composé	Tc- Exp	Tc-Calc	ei
76	Hexadec-1-ène	717,00	717,21	-0,21
77	Octadec-1-ène	739,00	741,93	-2,93
78	Cyclobutane	460,00	461,51	-1,51
79	Cyclopentane	511,70	509,34	2,36
80	Méthylcyclopentane	532,70	533,61	-0,91
81	Cyclohexane	553,50	555,81	-2,31
82	Méthylcyclohexane	572,20	567,59	4,61
83	Ethylcyclopentane	569,50	567,58	1,92
84	E-1,4-Diméthylcyclohexane	587,70	586,83	0,87
85	Cyclooctane	647,20	647,32	-0,12
86	Cyclopropane	397,80	397,30	0,50
87	1,1Diméthylcyclopentane	547,00	548,13	-1,13
88	Z-1,2-Diméthylcyclopentane	564,80	559,90	4,90
89	E-1,2-Diméthylcyclopentane	553,20	561,68	-8,48
90	E-1,2-Diméthylcyclohexane	596,00	597,35	-1,35
91	Z-1,3-Diméthylcyclohexane	591,00	588,66	2,34
92	Ethylcyclohexane	609,00	605,19	3,81
93	1,1,2-Triméthylcyclopentane	579,50	577,60	1,90
94	1,1,3-Triméthylcyclopentane	569,50	566,62	2,88
95	Propylcyclopentane	603,00	603,33	-0,33
96	Isopropylcyclopentane	601,00	592,93	8,07
97	Propylcyclohexane	639,00	640,02	-1,02
98	1,E-3,5-Triméthylcyclohexane	602,20	612,67	-10,47
99	Butylcyclohexane	667,00	664,37	2,63
100	Isobutylcyclohexane	659,00	662,19	-3,19
101	sec-Butylcyclohexane	669,00	670,83	-1,83
102	tert-Butylcyclohexane	659,00	652,73	6,27
103	Hexylcyclopentane	660,10	662,40	-2,30
104	Heptylcyclopentane	679,00	679,61	-0,61

Tableau - 6:(suite)

N	Composé	Tc- Exp	Tc-Calc	ei
105	Decylcyclopentane	723,80	723,35	0,45
106	Decylcyclohexane	750,00	749,65	0,35
107	Tridecylcyclopentane	761,00	760,44	0,56
108	1-Cyclopentyltetradecane	772,00	771,60	0,40
109	1-Cyclopentylpentadecane	780,00	780,78	-0,78
110	Buta-1,3-diène	425,00	425,83	-0,83
111	Buta-1,2-diène	443,70	450,73	-7,03
112	Penta-1,4-diène	478,00	477,74	0,26
113	3-Méthyl buta-1,2-diène	496,00	493,21	2,79
114	E-penta-1,3-diène	496,00	490,26	5,74
115	Penta-1,2-diène	503,00	503,91	-0,91
116	Propadiène	393,00	393,82	-0,82
117	Deca-1,3-diène	615,00	614,01	0,99
118	1-Méthylcyclopentène	542,15	543,22	-1,07
119	Cyclohexène	560,50	558,52	1,98
120	1-Ethylcyclopentène	576,15	582,00	-5,85
121	Cyclopentène	506,00	505,86	0,14
122	Acétylène	308,30	307,65	0,65
123	Propyne	402,40	400,58	1,82
124	But-2-yne	488,70	487,24	1,46
125	Vinylacétylène	455,00	455,24	-0,24
126*	Isobutane	408,20	419,28	-11,08
127*	Pentane	469,70	462,68	7,02
128*	3,3-Diméthylpentane	536,40	527,30	9,10
129*	2,3-Diméthylpentane	537,40	534,07	3,33
130*	2,3,4-Triméthylpentane	566,50	573,94	-7,44
131*	2,3-Diméthylhexane	563,50	562,35	1,15
132*	3-Ethyl-2-méthylpentane	567,10	574,30	-7,20
133*	2-Méthylheptane	559,70	559,05	0,65

Tableau - 6 :(suite)

N	Composé	Tc- Exp	Tc-Calc	ei
134*	3,4-Diméthylhexane	568,90	567,67	1,23
135*	4-Méthylheptane	561,80	563,10	-1,30
136*	Octane	568,90	569,61	-0,71
137*	2,2,5-Triméthylhexane	568,00	562,53	5,47
138*	2,3,3,4-Tétraméthylpentane	607,70	603,10	4,60
139*	2,2,3,3-Tétraméthylhexane	623,20	618,42	4,78
140*	3,3,5-Triméthylheptane	609,70	610,17	-0,47
141*	2,2-Diméthylpropane	433,80	426,64	7,16
142*	2,2,3-Triméthylhexane	588,00	584,27	3,73
143*	Heptadecane	733,00	734,19	-1,19
144*	Nonadecane	756,00	757,20	-1,20
145*	Ethylène	282,30	300,38	-18,08
146*	But-1-ène	419,60	418,33	1,27
147*	Oct-1-ène	566,70	568,99	-2,29
148*	E-Oct-2-ène	580,00	577,08	2,92
149*	Dec-1-ène	615,00	616,43	-1,43
150*	Tétradec-1-ène	689,00	688,41	0,59
151*	Pentadec-1-ène	704,00	703,23	0,77
152*	Cycloheptane	604,20	605,27	-1,07
153*	1,1-Diméthylcyclohexane	591,00	579,86	11,14
154*	Z-1,2-Diméthylcyclohexane	606,00	597,81	8,19
155*	E-1,3-Diméthylcyclohexane	598,00	588,42	9,58
156*	Z-1,4-Diméthylcyclohexane	598,00	586,83	11,17
157*	1-Ethyl-1-méthylcyclopentane	592,00	582,61	9,39
158*	Isopropylcyclohexane	640,00	635,35	4,65
159*	Octylcyclopentane	694,00	694,87	-0,87
160*	Nonylcyclopentane	710,50	709,42	1,08
161*	Dodecylcyclopentane	750,00	749,47	0,53
162*	Hexa-1,5-diène	507,00	510,58	-3,58

Tableau - 6:(suite et fin)

N	Composé	Tc- Exp	Tc-Calc	ei
163*	2-Méthyl buta-1,3-diène	484,00	484,37	-0,37
164*	But-1-yne	463,70	456,25	7,45
165*	Pent-1-yne	493,50	505,17	-11,67

*Composés de validation

II-Modélisation de la température d'ébullition

II-1- Introduction

La température d'ébullition (Teb) peut être définie comme le point auquel un liquide saturé pur a une pression de vapeur de 760 mm Hg. La Teb peut être utilisée pour évaluer beaucoup de propriétés physiques et physicochimiques comme par exemple la température critique, l'enthalpie de vaporisation ou la pression de vapeur [90-91]. Ainsi, la connaissance précise de Teb est très importante pour l'industrie chimique. La mesure directe du point d'ébullition de composés organiques peut être coûteuse, laborieuse et même dangereuse au chercheur ou à l'environnement si le composé a quelques propriétés dangereuses. Il est donc, essentiel de développer des méthodes fiables pour évaluer la température d'ébullition.

II-2-Résultats et discussion

II-2-1- Régression linéaire multiple

Sélection des descripteurs

On commencera par le choix de la taille de modèle, qui sera illustré par le graphe de la variation des coefficients de détermination par rapport au nombre de descripteurs. On remarque que la valeur de R^2 cesse d'augmente de façon significative après le modèle à cinq descripteurs, qui est la taille optimale adoptée.

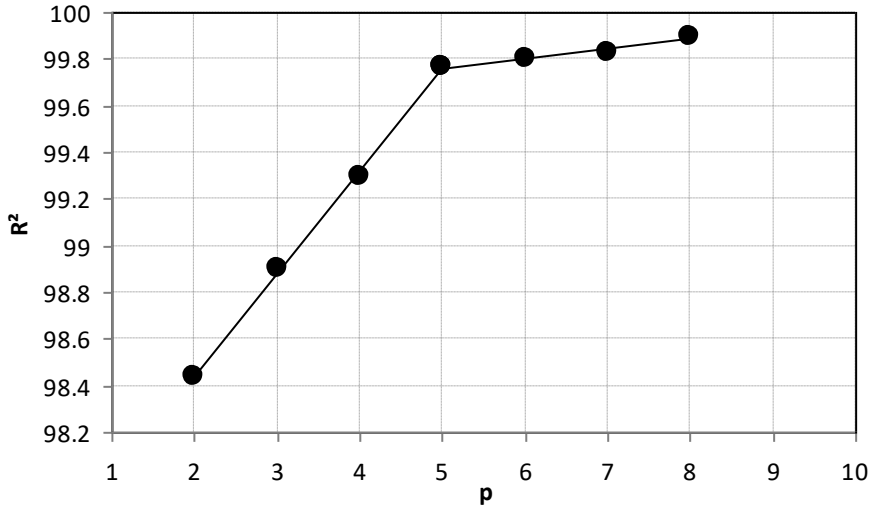


Figure 22: Variation du R² en fonction de nombre de descripteurs pour Teb**.**

Les 5 descripteurs du modèle choisis et leurs classes ainsi que de brèves définitions sont regroupés dans les annexes.

Ainsi l'équation de régression est la suivante :

Teb = - 55,37(±4,05) + 158,05(±1,15) **VEv1** – 36,96(±2,07) **MAXDN** + 8,41(±1,14) **HATS5u** - 238,77(±29,45) **H6m** + 344,09(±35,37) **R1p+** (98)

R^2= 99,77% Q^2_{LOO} = 99,70% EQMC = 4,75 EQMP = 4,91

s = 4,67 F = 9722,15

Q^2_{EXT} =99,75% $EQMP_{ext}$ = 4,53

La valeur de R² indique que 99,77 (%) de la variation totale est expliquée par le modèle, alors que la valeur élevée du rapport de la variance expliquée par le modèle à la variance résiduelle (F= 9722,15 ; p = 0,000) montre que l'équation (98) permet une très bonne prédiction de la température d'ébullition pour les125 composés de l'ensemble de calibrage avec une erreur standard inférieur à 5 (s = 4,67) ; la valeur élevée de Q^2_{LOO}, qui diffère peu de celle de R², renseigne sur la robustesse du modèle.

La contribution des descripteurs sélectionnés se présente dans l'ordre suivant: VEv1 > MAXDN > R1p+ > H6m > HATS5u. En plus, ces descripteurs ne sont pas corrélés les uns avec les autres puisque toutes les valeurs des VIF sont inférieures à 5.

Descripteur	x	Dx	T	probabilité-t	VIF
Constante	55,37	4,05	-13,69	0,000	
VEv1	158,05	1,15	137,43	0,000	2,95
MAXDN	-36,96	2,07	-17,86	0,000	1,22
R1p+	344,09	35,37	9,73	0,000	1,48
H6m	-238,77	29,45	-8,11	0,000	2,76
HATS5u	8,41	1,14	7,39	0,000	1,29

Les valeurs de la température d'ébullition expérimentales, calculées et prédites pour les deux ensembles, ainsi que les valeurs des leviers et des erreurs standardisées sont regroupées dans le tableau 7.

Tableau - 7: Valeurs des Teb expérimentales, calculées, prédites, ainsi que de h_{ii}, et e_{istd}

N	**Composé**	**Teb- Exp**	**Teb-Calc**	**hii**	**ei**	**ei std**
1	Méthane	111,60	102,67	0,368	8,93	2,15
2	Ethane	184,50	188,41	0,072	-3,91	-1,02
3	Propane	231,00	227,46	0,046	3,54	0,58
4	Butane	272,60	274,72	0,023	-2,12	-0,44
5	Isopentane	301,00	303,06	0,013	-2,06	-0,36
6	2,2-Diméthylbutane	322,90	324,57	0,025	-1,67	-0,30
7	2,3-Diméthylbutane	331,10	336,84	0,021	-5,74	-1,04
8	2-Méthylpentane	333,40	334,51	0,008	-1,11	-0,18
9	3- Méthylpentane	336,40	339,14	0,018	-2,74	-0,45
10	Hexane	341,90	341,92	0,018	-0,02	0,04
11	2,2-Diméthylpentane	352,30	352,08	0,027	0,22	0,09
12	2,4-Diméthylpentane	353,60	360,76	0,01	-7,16	-1,50
13	2,2,3-Triméthylbutane	354,00	356,67	0,033	-2,67	-0,36
14	2-Méthylhexane	363,20	362,34	0,009	0,86	0,21
15	E-1,2-Diméthylcyclopentane	365,00	371,67	0,014	-6,67	-1,29
16	3-Ethylpentane	366,60	369,85	0,018	-3,25	-0,54
17	Heptane	371,60	371,31	0,017	0,29	0,10
18	2,2,4-Triméthylpentane	372,40	377,25	0,034	-4,85	-1,00

Tableau -7: (suite)

N	Composé	Teb- Exp	Teb-Calc	hii	ei	ei std
19	2,2,3,3-Tetraméthylbutane	379,60	381,06	0,052	-1,46	-0,09
20	2,2-Diméthylhexane	380,00	377,81	0,032	2,19	0,48
21	2,5-Diméthylhexane	382,30	385,32	0,014	-3,02	-0,59
22	2,4-Diméthylhexane	382,60	389,37	0,009	-6,77	-1,35
23	2,2,3-Triméthylpentane	383,00	382,97	0,029	0,03	0,16
24	3,3-Diméthylhexane	385,10	382,56	0,02	2,54	0,67
25	2,3,3-Triméthylpentane	387,90	385,67	0,038	2,23	0,73
26	3-Ethyl-3-méthylpentane	391,40	390,17	0,028	1,23	0,50
27	3-Ethylhexane	391,70	395,63	0,01	-3,93	-0,71
28	3-Méthylheptane	392,10	392,72	0,008	-0,62	-0,05
29	2,2,4,4-Tetraméthylpentane	395,40	400,00	0,044	-4,60	-0,94
30	2,2-Diméthylheptane	405,80	402,25	0,038	3,55	0,78
31	2,2,3,4-Tetraméthylpentane	406,20	406,74	0,031	-0,54	0,07
32	2,2,3,3-Tetraméthylpentane	413,40	409,49	0,044	3,91	1,11
33	2-Méthyloctane	416,40	414,10	0,015	2,30	0,51
34	Nonane	424,00	422,88	0,023	1,12	0,24
35	2,2,5,5-Tetraméthylhexane	410,60	419,05	0,057	-8,45	-1,81
36	Decane	447,30	447,38	0,028	-0,08	0,02
37	Undecane	469,10	469,86	0,033	-0,76	-0,12
38	Dodecane	489,50	491,80	0,038	-2,30	-0,45
39	Tridecane	508,60	510,38	0,034	-1,78	-0,30
40	Tetradecane	526,70	527,50	0,029	-0,80	-0,09
41	Hexadecane	560,00	559,43	0,048	0,57	0,37
42	Octadecane	589,50	588,97	0,107	0,53	0,48
43	2,2,4-Triméthylhexane	399,70	403,65	0,03	-3,95	-0,77
44	3,3-Diethylpentane	419,30	416,23	0,026	3,07	0,90
45	Pentadecane	543,80	543,68	0,034	0,12	0,21
46	Eicosane	617,00	616,42	0,204	0,58	0,64

Tableau-7: (suite)

N	Composé	Teb- Exp	Teb-Calc	hii	ei	ei std
47	E-But-2-ène	274,00	288,01	0,063	-14,01	-2,93
48	Z-But-2-ène	276,90	287,36	0,055	-10,46	-2,16
49	Pent-1-ène	303,10	308,02	0,012	-4,92	-1,02
50	2-Méthyl but-1-ène	304,30	306,93	0,016	-2,63	-0,45
51	E-pent-2-ène	309,50	308,16	0,016	1,34	0,37
52	Z-pent-2-ène	310,10	306,19	0,012	3,91	0,86
53	2-Méthyl but-2-ène	311,70	313,41	0,037	-1,71	-0,18
54	Hex-1-ène	336,60	341,85	0,01	-5,25	-1,10
55	Hept-1-ène	366,80	371,76	0,01	-4,96	-1,00
56	Propylène	225,50	232,59	0,049	-7,09	-1,68
57	Isobutylène	266,20	260,74	0,047	5,46	1,04
58	3-Méthyl but-1-ène	293,30	300,14	0,032	-6,84	-1,41
59	Z-Hex-2-ène	342,00	339,44	0,009	2,56	0,58
60	E-Hex-2-ène	341,00	341,32	0,01	-0,32	-0,05
61	Z-Hex-3-ène	339,60	334,66	0,009	4,94	1,08
62	E-Hex-3-ène	340,30	335,91	0,008	4,39	0,92
63	2-Méthyl pent-2-ène	340,50	339,59	0,01	0,91	0,27
64	Z-3-Méthyl pent-2-ène	340,90	341,69	0,034	-0,79	0,01
65	E-3-Méthyl pent-2-ène	343,60	343,74	0,03	-0,14	0,17
66	Z-4-Méthyl pent-2-ène	329,60	327,97	0,017	1,63	0,40
67	E-4-Méthyl pent-2-ène	331,70	329,79	0,019	1,91	0,41
68	2,3-Diméthyl but-2-ène	346,40	347,21	0,064	-0,81	0,11
69	2,3-Diméthyl but-1-ène	328,80	335,57	0,036	-6,77	-1,28
70	3,3-Diméthyl but-1-ène	314,40	322,72	0,057	-8,32	-1,72
71	2,3,3-Triméthyl but-1-ène	351,00	351,62	0,053	-0,62	0,09
72	Non-1-ène	420,00	423,22	0,014	-3,22	-0,63
73	Undec-1-ène	465,80	469,26	0,024	-3,46	-0,69
74	Dodec-1-ène	486,50	490,39	0,028	-3,89	-0,77

Tableau -7: (suite)

N	Composé	Teb- Exp	Teb-Calc	hii	ei	ei std
75	Tridec-1-ène	505,90	508,68	0,026	-2,78	-0,55
76	Hexadec-1-ène	558,00	556,76	0,047	1,24	0,52
77	Octadec-1-ène	588,00	585,10	0,118	2,90	1,04
78	Cyclobutane	285,70	280,21	0,043	5,49	1,01
79	Cyclopentane	322,40	314,88	0,044	7,52	1,48
80	Méthylcyclopentane	344,90	341,16	0,01	3,74	0,79
81	Cyclohexane	353,90	352,03	0,031	1,87	0,37
82	Méthylcyclohexane	374,10	372,19	0,01	1,91	0,46
83	Ethylcyclopentane	376,60	374,43	0,012	2,17	0,53
84	E-1,4-Diméthylcyclohexane	392,50	396,22	0,015	-3,72	-0,70
85	Cyclooctane	422,00	414,57	0,041	7,43	1,70
86	Cyclopropane	240,30	243,14	0,052	-2,84	-0,77
87	1,1Diméthylcyclopentane	361,00	363,05	0,011	-2,05	-0,40
88	Z-1,2-Diméthylcyclopentane	372,70	373,64	0,016	-0,94	-0,05
89	3-Méthylhexane	365,00	366,52	0,009	-1,52	-0,25
90	E-1,2-Diméthylcyclohexane	396,60	400,36	0,018	-3,76	-0,65
91	Z-1,3-Diméthylcyclohexane	393,30	399,33	0,012	-6,03	-1,22
92	Ethylcyclohexane	404,90	401,85	0,015	3,05	0,75
93	1,1,2-Triméthylcyclopentane	386,90	392,29	0,023	-5,39	-0,97
94	1,1,3-Triméthylcyclopentane	378,00	386,53	0,016	-8,53	-1,74
95	Propylcyclopentane	404,10	399,27	0,014	4,83	1,06
96	Isopropylcyclopentane	399,60	395,99	0,01	3,61	0,87
97	Propylcyclohexane	429,90	425,77	0,018	4,13	0,94
98	1,E-3,5-Triméthylcyclohexane	413,70	423,53	0,015	-9,83	-2,00
99	Butylcyclohexane	454,10	448,92	0,023	5,18	1,15
100	Isobutylcyclohexane	444,50	441,62	0,017	2,88	0,67
101	sec-Butylcyclohexane	452,50	445,57	0,013	6,93	1,62
102	tert-Butylcyclohexane	444,70	434,94	0,025	9,76	2,22

Tableau -7: (suite)

N	Composé	Teb- Exp	Teb-Calc	hii	ei	ei std
103	Hexylcyclopentane	476,30	473,00	0,037	3,30	0,72
104	Heptylcyclopentane	497,30	494,66	0,044	2,64	0,58
105	Décylcyclopentane	552,50	553,90	0,054	-1,40	-0,24
106	Decylcyclohexane	570,80	570,32	0,046	0,48	0,17
107	Tridecylcyclopentane	598,60	600,32	0,052	-1,72	-0,15
108	1-Cyclopentyltetradecane	599,00	614,06	0,074	-15,06	-3,06
109	1-Cyclopentylpentadecane	625,00	628,12	0,102	-3,12	-0,34
110	Buta-1,3-diène	268,70	268,98	0,023	-0,28	-0,15
111	Buta-1,2-diène	284,00	286,17	0,068	-2,17	-0,35
112	Penta-1,4-diène	299,10	296,69	0,028	2,41	0,46
113	3-Méthyl buta-1,2-diène	314,00	313,86	0,072	0,14	0,25
114	E-penta-1,3-diène	315,20	318,25	0,042	-3,05	-0,62
115	Penta-1,2-diène	318,00	309,33	0,038	8,67	1,85
116	Propadiène	238,70	231,37	0,046	7,33	1,43
117	Dec-1-ène	443,70	447,17	0,02	-3,47	-0,69
118	1-Méthylcyclopentène	348,95	347,70	0,016	1,25	0,33
119	Cyclohexène	356,10	350,84	0,017	5,26	1,11
120	1-Ethylcyclopentène	379,45	375,78	0,016	3,67	0,88
121	Cyclopentène	317,40	315,10	0,037	2,30	0,35
122	Acétylène	188,40	187,62	0,074	0,78	-0,02
123	Propyne	249,90	236,76	0,094	13,14	2,75
124	But-2-yne	300,10	288,43	0,123	11,67	2,97
125	Vinylacétylène	278,10	272,25	0,088	5,85	1,18
126*	Isobutane	261,40	258,11	0,049	3,29	0,57
127*	Pentane	309,20	312,31	0,02	-3,11	-0,60
128*	3,3-Diméthylpentane	359,20	357,39	0,026	1,81	0,57
129*	2,3-Diméthylpentane	362,90	366,12	0,018	-3,22	-0,50
130*	2,3,4-Triméthylpentane	386,60	392,98	0,016	-6,38	-1,21

Tableau -7: (suite)

N	Composé	Teb- Exp	Teb-Calc	hii	ei	ei std
131*	2,3-Diméthylhexane	388,80	390,19	0,011	-1,39	-0,16
132*	3-Ethyl-2-méthylpentane	388,80	391,96	0,016	-3,16	-0,48
133*	2-Méthylheptane	390,80	388,58	0,011	2,22	0,46
134*	4-Méthylheptane	390,90	392,72	0,008	-1,82	-0,30
135*	3,4-Diméthylhexane	390,90	393,71	0,019	-2,81	-0,40
136*	Octane	398,80	397,41	0,019	1,39	0,30
137*	2,2,5-Triméthylhexane	397,20	400,14	0,04	-2,94	-0,59
138*	2,3,3,4-Tetraméthylpentane	414,70	409,50	0,044	5,20	1,39
139*	2,2,3,3-Tetraméthylhexane	433,50	430,13	0,035	3,37	0,93
140*	3,3,5-Triméthylheptane	428,80	431,09	0,022	-2,29	-0,36
141*	2,2-Diméthylpropane	282,60	282,79	0,079	-0,19	-0,18
142*	2,2,3-Triméthylhexane	406,80	406,59	0,025	0,21	0,19
143*	Heptadecane	575,20	574,11	0,075	1,09	0,55
144*	Nonadecane	603,10	602,96	0,152	0,14	0,47
145	Ethylène	169,40	188,41	0,072	-19,01	-4,32
146*	But-1-ène	266,90	270,04	0,02	-3,14	-0,66
147*	Oct-1-ène	394,40	398,62	0,013	-4,22	-0,88
148*	E-Oct-2-ène	398,10	396,78	0,012	1,32	0,33
149*	Deca-1,3-diène	442,00	451,41	0,037	-9,41	-1,97
150	Tetradec-1-ène	524,30	525,85	0,023	-1,55	-0,24
151*	Pentadec-1-ène	541,50	541,20	0,029	0,30	0,21
152*	Cycloheptane	391,60	386,23	0,036	5,37	1,24
153*	1,1-Diméthylcyclohexane	392,70	390,60	0,014	2,10	0,51
154*	Z-1,2-Diméthylcyclohexane	402,90	401,47	0,018	1,43	0,43
155*	E-1,3-Diméthylcyclohexane	397,60	399,64	0,013	-2,04	-0,32
156*	Z-1,4-Diméthylcyclohexane	397,50	396,22	0,015	1,28	0,36
157*	1-Ethyl-1-méthylcyclopentane	394,70	395,29	0,015	-0,59	0,04
158*	Isopropylcyclohexane	427,70	420,91	0,013	6,79	1,56

Tableau -7: (suite et fin)

159*	Octylcyclopentane	516,90	516,02	0,051	0,88	0,21
160*	Nonylcyclopentane	535,30	534,65	0,049	0,65	0,19
161*	Dodecylcyclopentane	584,10	586,16	0,047	-2,06	-0,34
162*	Hexa-1,5-diène	332,60	337,28	0,008	-4,68	-0,94
163*	2-Méthyl buta-1,3-diène	307,20	305,29	0,044	1,91	0,49
164*	But-1-yne	281,20	274,20	0,121	7,00	1,55
165*	Pent-1-yne	313,30	309,60	0,089	3,70	0,78

*Composés de validation

Qualité de l'ajustement

La qualité de l'ajustement a été vérifiée en procédant à une validation croisée par « leave –one – out ». La figure (23), qui reproduit les valeurs prédites en fonction de celles observées, fait ressortir une faible dispersion caractéristique d'un bon ajustement, cela est d'ailleurs confirmé par la grande valeur de Q^2 (= 99,70%).

$$\text{Teb-Calc} = 0{,}94 + 0{,}99\text{Teb-Exp} \qquad (99)$$

$S = 4{,}60 \qquad R^2 = 99{,}80\ \% \qquad R^2_{ajust} = 99{,}80\ \%$

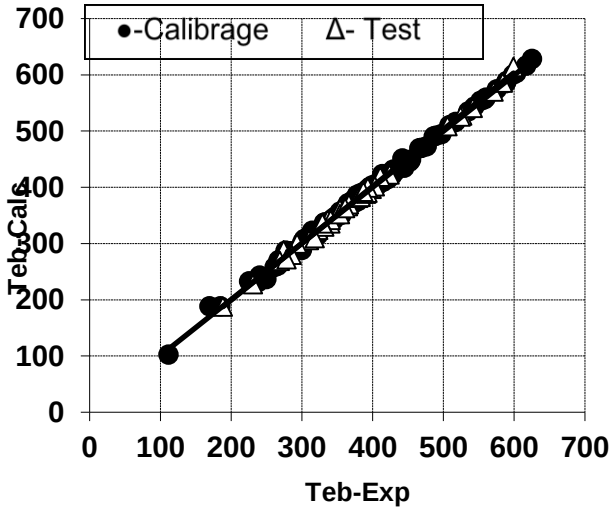


Figure 23: Graphe des valeurs de Teb calculées et prédites (Teb-Calc) en fonction des valeurs observées (Teb-Exp)

La droite de la méthode des moindres carrés (y=a+bx) ayant une pente très proche ou égale à l'unité est pratiquement superposée à la première bissectrice du repère choisi [1,01 > b=1,00> 0,99], et avec une ordonnée à l'origine faible [3,90 > a=0,94> - 2,02],.

La variation des résidus standardisés (e_{istd}) en fonction des valeurs de *hii* (diagramme de Williams) semble aléatoire avec une distribution tout à fait acceptable et équilibrée autour de zéro, puisque 3 résidus de l'ensemble de calibrage (soit 2,4 %) sont supérieurs à la valeur critique hi*=0.144, il s'agit des composés : 1, 50, et 51 qui sont des composés structurellement influents, mais un seul composé a une erreur standardisée hors les limites ± 3 pour l'ensemble de calibrage et un autre de l'ensemble de validation, ce sont les composés 52 et 161 respectivement.

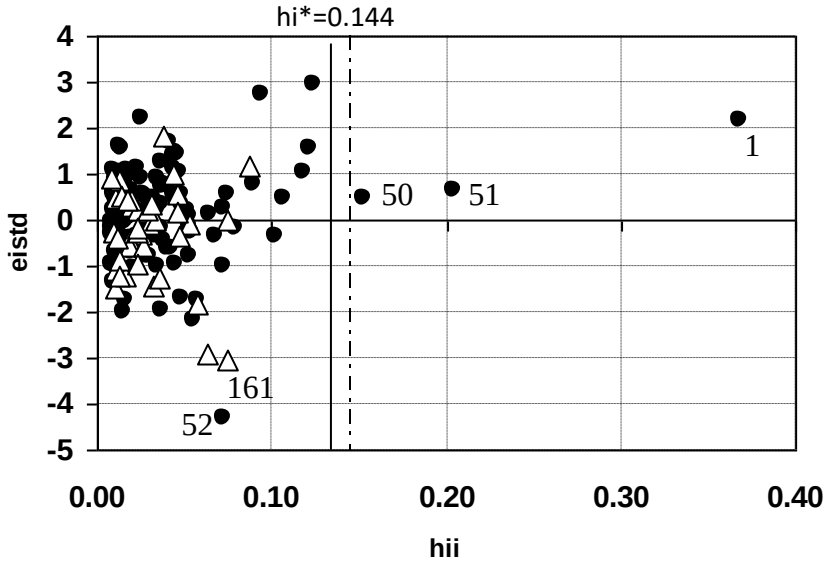


Figure 24: *Diagramme de Williams*

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.

La figure 25 qui représente le graphe des coefficients statistiques Q^2 et R^2 permet de comparer les résultats obtenus pour les modèles randomisés (gros points) au modèle de départ (cercle vide). Il est clair que les statistiques obtenues pour les vecteurs modifiés des températures d'ébullitions sont plus petites que celles du modèle QSPR réel, ce qui permet d'assurer qu'une relation réelle structure / température d'ébullition a été établie.

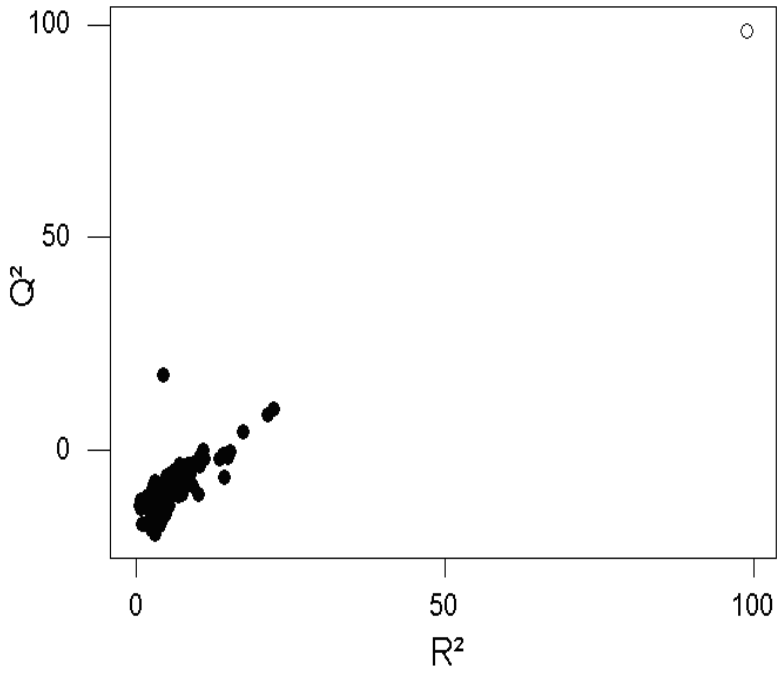


Figure 25 : *Test de randomisation*

Les autres paramètres statistiques qui montrent le pouvoir prédictif du modèle sont regroupés dans le tableau suivant :

Paramètre	R^2ext	Q^2_{F1}	Q^2_{F2}	Q^2_{F3}	CCC
Valeur	0,9976	0,9970	0,9969	0,9969	0,9985

En plus nous notons la vérification du test de G.et T.

1- Q^2_{ext} = 0,9970 > 0,5

2- R^2= 0,9976 > 0,6

3- 0,85 < k =0,9949 < 1,15

4- 0,85 < k'=1,0049 < 1,15

II-2-2- Régression par les réseaux de neurones artificiels RNA

Choix des paramètres optimaux

Pour améliorer la performance du modèle, plusieurs essais préliminaires ont été réalisés. Ainsi, l'architecture du réseau de neurones est modifiée progressivement jusqu'à l'obtention du modèle le plus performant, en jouant surtout sur le nombre de couches cachées, ou sur celui des neurones cachés (figure 26) et/ou sur le nombre de cycles d'entraînements ou nombre d'itérations (figure 27).

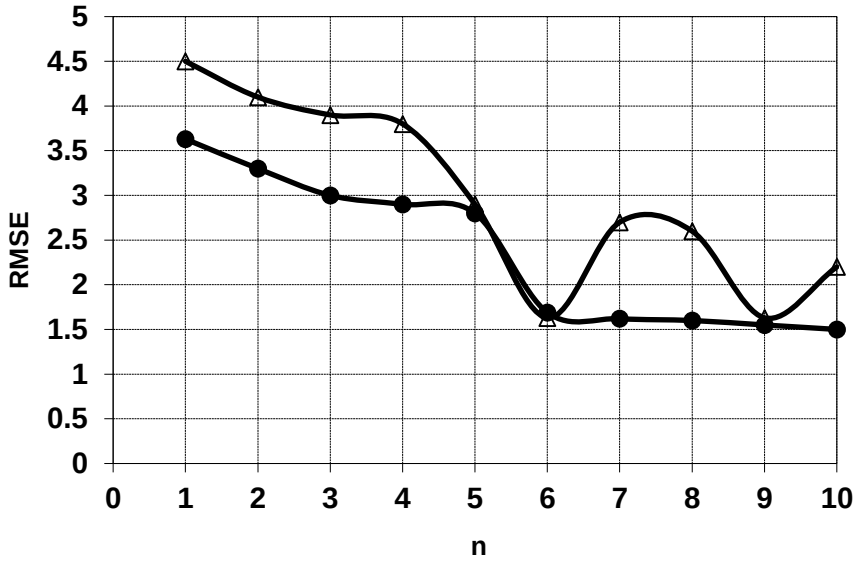


Figure 26: Choix du nombre de neurones de la couche cachée

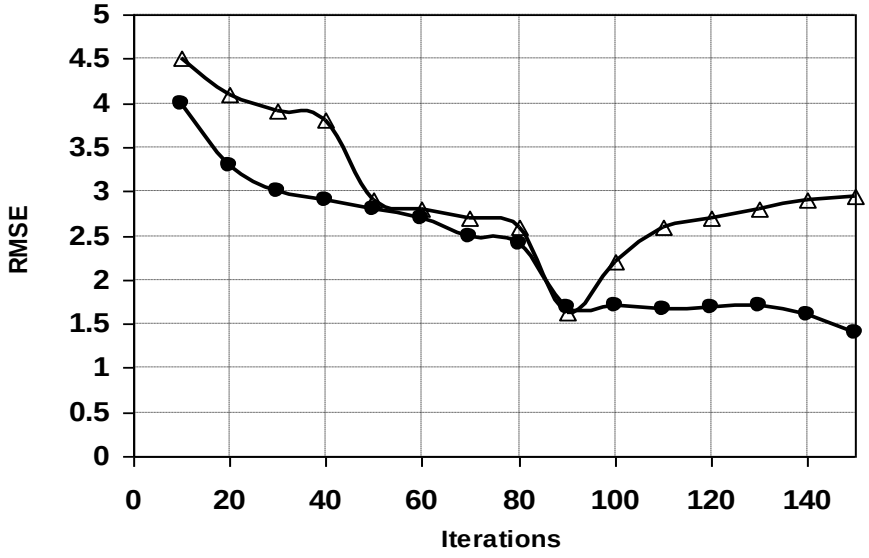


Figure 27: Choix du nombre d'itérations

Parmi les différentes configurations testées du réseau, nous avons retenu celle produisant la plus faible racine carrée de l'erreur quadratique moyenne (RMSE) sur l'ensemble de données de test (40 composés). Le tableau 8 montre l'architecture optimale avec six (06) neurones dans la couche cachée et 90 itérations

Tableau - 8: Structure optimale du réseau de neurones.

Nombre d'entrées	05 (les descripteurs)
Nombre de sorties	01 (la température d'ébullition)
Nombre de couches cachées	Une couche cachée
Nombre de neurones dans la couche cachée	06
Nombre d'itérations	90
Algorithme d'apprentissage	Rétro propagation du gradient de l'erreur
Fonctions d'apprentissage	Tangente hyperbolique

Les résultats obtenus (valeurs estimées ou calculées) sont comparés aux valeurs expérimentales (valeurs observées) (figure 28). La relation entre les valeurs observées et celles estimées montre la performance et la qualité du modèle développé par la méthode neuronale. Cette performance est évaluée par les paramètres statistiques réunis dans le tableau 9.

Tableau - 9 : Valeurs des paramètres statistiques

Ensemble de Calibrage :125			Tests :40	Ensemble de Validation :40	
R^2 (%)	s	RMSEcal	RMSEtst	R^2cv (%)	RMSEval
99,92	2,27	2,57	2,19	99,86	3,54

La valeur du cœfficient de détermination R^2= 99,92 % explique très bien la variabilité de Teb en fonction des descripteurs choisis ; la petite valeur de la racine de l'erreur quadratique moyenne RMSE ou EQMC = 2,57, indique un modèle très hautement significatif.

Qualité de l'ajustement

La qualité de l'ajustement à été vérifiée en représentant les valeurs calculées Teb-Calc (colonne 4 tableau 10) par notre modèle en fonction des valeurs observées ou expérimentales Teb-Exp (colonne 3 tableau 10). La figue 43 montre une faible dispersion autour de la droite d'ajustement (qui peut être assimilée à la première bissectrice) définie par l'équation (100).

$$\text{Teb-Calc} = 0{,}1 + 1{,}00\ \text{Teb-Exp} \qquad (100)$$

$s = 2{,}59 \qquad R^2 = 99{,}90\ \% \qquad R^2_{ajust} = 99{,}90\ \%$

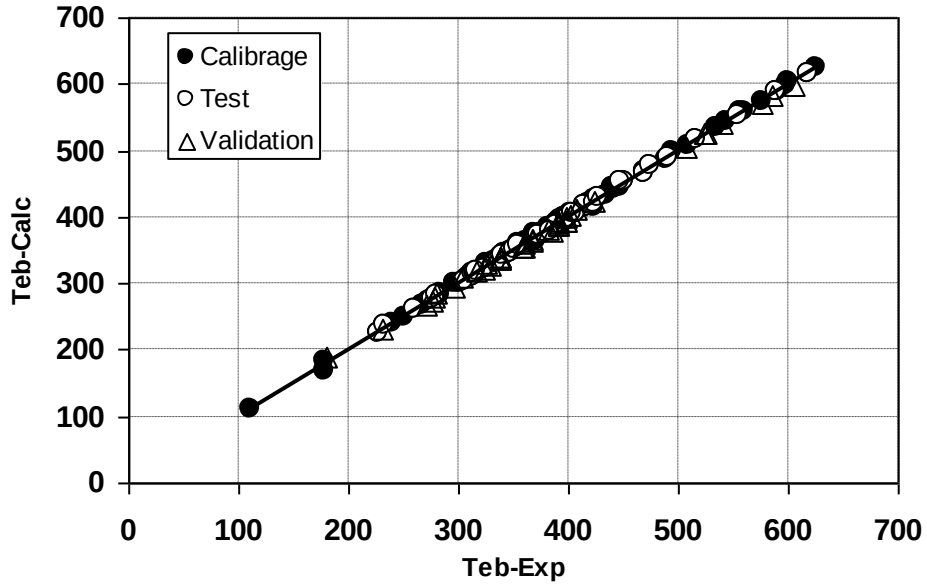
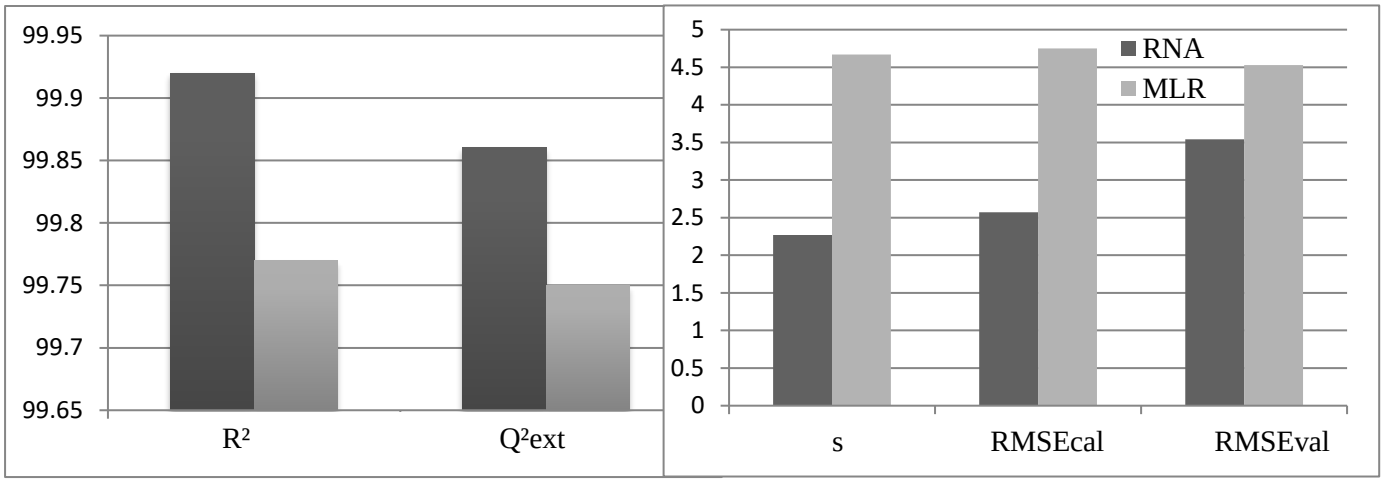


Figure 28 : Graphe des valeurs Teb calculées, Tests et prédites en fonction des valeurs observées

La comparaison des paramètres statistiques entre les deux méthodes (histogramme suivant) nous a permis de conclure que la méthode neuronale est meilleure que la MLR.



La variation des résidus en fonction des températures observées, représentée dans la figure suivante montre une distribution est acceptable

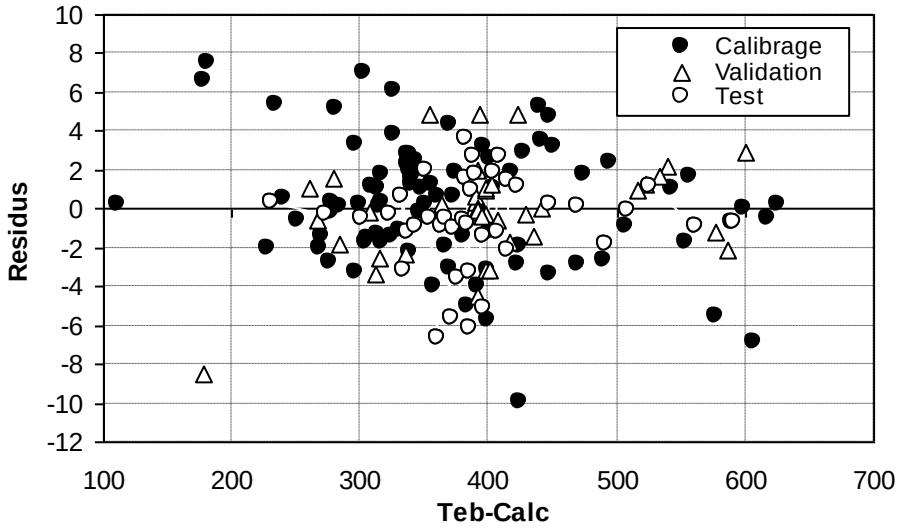


Figure 29: Variation des résidus en fonction des valeurs calculées.

Tableau - 10 Valeurs de Teb expérimentales, calculées, prédites, et des résidus

N	Composé	Teb-Exp	Teb-Calc	ei
1	Méthane	111,60	111,33	0,27
2	Ethane	184,50	177,86	6,64
3	Propane	231,00	230,70	0,30
4	Butane	272,60	272,87	-0,27
5	Isopentane	301,00	301,49	-0,49
6	2,2-Diméthylbutane	322,90	323,22	-0,32
7	2,3-Diméthylbutane	331,10	334,28	-3,18
8	2-Méthylpentane	333,40	332,74	0,66
9	3- Méthylpentane	336,40	337,59	-1,19
10	Hexane	341,90	342,78	-0,88
11	2,2-Diméthylpentane	352,30	350,32	1,98
12	2,4-Diméthylpentane	353,60	360,22	-6,62
13	2,2,3-Triméthylbutane	354,00	354,45	-0,45
14	2-Méthylhexane	363,20	364,13	-0,93
15	E-1,2-Diméthylcyclopentane	365,00	370,61	-5,61
16	3-Ethylpentane	366,60	367,11	-0,51
17	Heptane	371,60	372,64	-1,04
18	2,2,4-Triméthylpentane	372,40	375,98	-3,58
19	2,2,3,3-Tétraméthylbutane	379,60	385,78	-6,18
20	2,2-Diméthylhexane	380,00	380,57	-0,57
21	2,5-Diméthylhexane	382,30	383,09	-0,79
22	2,4-Diméthylhexane	382,60	385,83	-3,23
23	2,2,3-Triméthylpentane	383,00	381,48	1,52
24	3,3-Diméthylhexane	385,10	381,49	3,61
25	2,3,3-Triméthylpentane	387,90	386,93	0,97
26	3-Ethyl-3-méthylpentane	391,40	388,73	2,67
27	3-Ethylhexane	391,70	396,79	-5,09
28	3-Méthylheptane	392,10	390,28	1,82
29	2,2,4,4-Tétraméthylpentane	395,40	396,85	-1,45

Tableau – 10 : (suite)

N	Composé	Teb-Exp	Teb-Calc	ei
30	2,2-Diméthylheptane	405,80	403,91	1,89
31	2,2,3,4-Tétraméthylpentane	406,20	407,41	-1,21
32	2,2,3,3-Tétraméthylpentane	413,40	415,50	-2,10
33	2-Méthyloctane	416,40	414,98	1,42
34	Nonane	424,00	422,82	1,18
35	2,2,5,5-Tétraméthylhexane	410,60	407,94	2,66
36	Décane	447,30	447,04	0,26
37	Undécane	469,10	469,01	0,09
38	Dodécane	489,50	491,33	-1,83
39	Tridécane	508,60	508,67	-0,07
40	Tétradécane	526,70	525,53	1,17
41	Hexadécane	560,00	560,89	-0,89
42	Octadécane	589,50	590,23	-0,73
43	2,2,4-Triméthylhexane	399,70	400,86	-1,16
44	3,3-Diethylpentane	419,30	417,46	1,84
45	Pentadecane	543,80	542,79	1,01
46	Eicosane	617,00	617,51	-0,51
47	E-But-2-ène	274,00	276,70	-2,70
48	Z-But-2-ène	276,90	277,03	-0,13
49	Pent-1-ène	303,10	304,77	-1,67
50	2-Méthyl but-1-ène	304,30	305,83	-1,53
51	E-pent-2-ène	309,50	308,30	1,20
52	Z-pent-2-ène	310,10	303,06	7,04
53	2-Méthyl but-2-ène	311,70	312,96	-1,26
54	Hex-1-ène	336,60	338,88	-2,28
55	Hept-1-ène	366,80	369,84	-3,04
56	Propylène	225,50	227,51	-2,01
57	Isobutylène	266,20	268,18	-1,98
58	3-Méthyl but-1-ène	293,30	296,58	-3,28

Tableau – 10 :(suite)

N	Composé	Teb-Exp	Teb-Calc	ei
59	Z-Hex-2-ène	342,00	340,40	1,60
60	E-Hex-2-ène	341,00	339,78	1,22
61	Z-Hex-3-ène	339,60	337,33	2,27
62	E-Hex-3-ène	340,30	337,51	2,79
63	2-Méthyl pent-2-ène	340,50	337,71	2,79
64	Z-3-Méthyl pent-2-ène	340,90	338,95	1,95
65	E-3-Méthyl pent-2-ène	343,60	341,87	1,73
66	Z-4-Méthyl pent-2-ène	329,60	325,79	3,81
67	E-4-Méthyl pent-2-ène	331,70	325,61	6,09
68	2,3-Diméthyl but-2-ène	346,40	346,59	-0,19
69	2,3-Diméthyl but-1-ène	328,80	329,95	-1,15
70	3,3-Diméthyl but-1-ène	314,40	316,11	-1,71
71	2,3,3-Triméthyl but-1-ène	351,00	350,72	0,28
72	Non-1-ène	420,00	422,83	-2,83
73	Undec-1-ène	465,80	468,68	-2,88
74	Dodec-1-ène	486,50	489,19	-2,69
75	Tridec-1-ène	505,90	506,77	-0,87
76	Hexadec-1-ène	558,00	556,33	1,67
77	Octadec-1-ène	588,00	588,74	-0,74
78	Cyclobutane	285,70	280,58	5,12
79	Cyclopentane	322,40	323,86	-1,46
80	Méthylcyclopentane	344,90	342,43	2,47
81	Cyclohexane	353,90	357,88	-3,98
82	Méthylcyclohexane	374,10	369,72	4,38
83	Ethylcyclopentane	376,60	374,68	1,92
84	E-1,4-Diméthylcyclohexane	392,50	392,59	-0,09
85	Cyclooctane	422,00	423,96	-1,96
86	Cyclopropane	240,30	239,78	0,52
87	1,1-Diméthylcyclopentane	361,00	360,35	0,65

Tableau -10:(suite)

N	Composé	Teb-Exp	Teb-Calc	ei
88	Z-1,2-Diméthylcyclopentane	372,70	372,04	0,66
89	3-Méthylhexane	365,00	366,96	-1,96
90	E-1,2-Diméthylcyclohexane	396,60	399,80	-3,20
91	Z-1,3-Diméthylcyclohexane	393,30	399,01	-5,71
92	Ethylcyclohexane	404,90	403,87	1,03
93	1,1,2-Triméthylcyclopentane	386,90	390,93	-4,03
94	1,1,3-Triméthylcyclopentane	378,00	383,06	-5,06
95	Propylcyclopentane	404,10	401,49	2,61
96	Isopropylcyclopentane	399,60	396,39	3,21
97	Propylcyclohexane	429,90	427,00	2,90
98	1,E-3,5-Triméthylcyclohexane	413,70	423,66	-9,96
99	Butylcyclohexane	454,10	450,85	3,25
100	Isobutylcyclohexane	444,50	440,99	3,51
101	sec-Butylcyclohexane	452,50	447,70	4,80
102	tert-Butylcyclohexane	444,70	439,48	5,22
103	Hexylcyclopentane	476,30	474,56	1,74
104	Heptylcyclopentane	497,30	494,91	2,39
105	Decylcyclopentane	552,50	554,17	-1,67
106	Decylcyclohexane	570,80	576,32	-5,52
107	Tridecylcyclopentane	598,60	598,58	0,02
108	1-Cyclopentyltetradecane	599,00	605,82	-6,82
109	1-Cyclopentylpentadecane	625,00	624,74	0,26
110	Buta-1,3-diène	268,70	270,06	-1,36
111	Buta-1,2-diène	284,00	283,83	0,17
112	Penta-1,4-diène	299,10	295,80	3,30
113	3-Méthyl buta-1,2-diène	314,00	312,99	1,01
114	E-penta-1,3-diène	315,20	315,12	0,08
115	Penta-1,2-diène	318,00	316,27	1,73
116	Propadiène	238,70	233,37	5,33

Tableau -10:(suite)

N	Composé	Teb-Exp	Teb-Calc	ei
117	Dec-1-ène	443,70	447,07	-3,37
118	1-Méthylcyclopentène	348,95	347,91	1,04
119	Cyclohexène	356,10	354,81	1,29
120	1-Ethylcyclopentène	379,45	380,84	-1,39
121	Cyclopentène	317,40	317,06	0,34
122	Acétylène	188,40	180,83	7,57
123	Propyne	249,90	250,46	-0,56
124	But-2-yne	300,10	299,84	0,26
125	Vinylacétylène	278,10	277,76	0,34
126*	Isobutane	261,40	260,30	1,10
127*	Pentane	309,20	312,59	-3,39
128*	3,3-Diméthylpentane	359,20	354,37	4,83
129*	2,3-Diméthylpentane	362,90	362,64	0,26
130*	2,3,4-Triméthylpentane	386,60	391,11	-4,51
131*	2,3-Diméthylhexane	388,80	389,00	-0,20
132*	3-Ethyl-2-méthylpentane	388,80	388,81	-0,01
133*	2-Méthylheptane	390,80	390,48	0,32
134*	4-Méthylheptane	390,90	390,28	0,62
135*	3,4-Diméthylhexane	390,90	390,90	0,00
136*	Octane	398,80	397,86	0,94
137*	2,2,5-Triméthylhexane	397,20	397,45	-0,25
138*	2,3,3,4-Tétraméthylpentane	414,70	416,46	-1,76
139*	2,2,3,3-Tétraméthylhexane	433,50	434,94	-1,44
140*	3,3,5-Triméthylheptane	428,80	429,09	-0,29
141*	2,2-Diméthylpropane	282,60	284,38	-1,78
142*	2,2,3-Triméthylhexane	406,80	407,44	-0,64
143*	Heptadecane	575,20	576,45	-1,25
144*	Nonadecane	603,10	600,19	2,91
145*	Ethylène	169,40	177,86	-8,46

Tableau -10:(suite et fin)

N	Composé	Teb-Exp	Teb-Calc	ei
146*	But-1-ène	266,90	267,44	-0,54
147*	Oct-1-ène	394,40	397,52	-3,12
148*	E-Oct-2-ène	398,10	397,02	1,08
149*	Deca-1,3-diène	442,00	441,97	0,03
150*	Tétradec-1-ène	524,30	523,08	1,22
151*	Pentadec-1-ène	541,50	539,36	2,14
152*	Cycloheptane	391,60	391,67	-0,07
153*	1,1-Diméthylcyclohexane	392,70	390,67	2,03
154*	Z-1,2-Diméthylcyclohexane	402,90	401,64	1,26
155*	E-1,3-Diméthylcyclohexane	397,60	400,72	-3,12
156*	Z-1,4-Diméthylcyclohexane	397,50	392,59	4,91
157*	1-Ethyl-1-méthylcyclopentane	394,70	395,13	-0,43
158*	Isopropylcyclohexane	427,70	422,83	4,87
159*	Octylcyclopentane	516,90	515,94	0,96
160*	Nonylcyclopentane	535,30	533,65	1,65
161*	Dodecylcyclopentane	584,10	586,24	-2,14
162*	Hexa-1,5-diène	332,60	334,90	-2,30
163*	2-Méthyl buta-1,3-diène	307,20	307,42	-0,22
164*	But-1-yne	281,20	279,66	1,54
165*	Pent-1-yne	313,30	315,80	-2,50

*Composés de validation

III- Modélisation de la pression critique

III-1-Introduction

La pression critique (Pc) est définie comme la pression la plus élevée à laquelle les phases gazeuse et liquide, d'une substance donnée, peuvent encore coexister.

La pression critique est nécessaire dans les calculs thermodynamiques, et importante pour les ingénieurs de l'industrie chimique. Elle est encore importante dans de nombreux domaines de la recherche pharmaceutique, comme par exemple l'ingénierie des cristaux qui utilise les fluides supercritiques.

Dans cette partie les deux méthodes (MLR, RNA) ont été appliquées pour prédire la pression critique, comprise entre 10,20 et 61,39 bars, d'un ensemble hétérogène de 165 hydrocarbures.

III-2- Résultats et discussion

III-2-1- Régression linéaire multiple

Sélection des descripteurs

De la même façon que précédemment, on commencera par le choix de la taille de modèle, le graphe de la variation des coefficients de détermination par rapport au nombre de descripteurs nous a permis de voir qu'après le modèle à cinq descripteurs la valeur de R^2 cesse d'augmenter, ce que fixe la taille du modèle à 5.

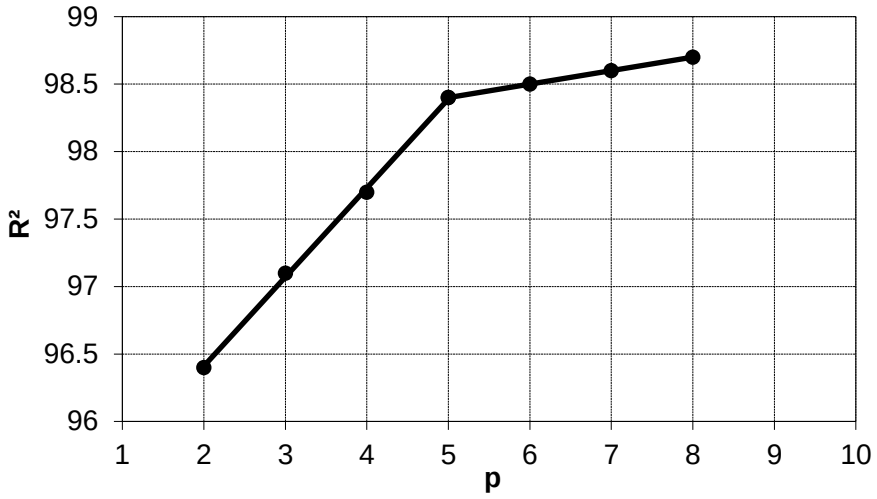


Figure 30: Variation du R^2 en fonction du nombre de descripteurs pour Pc

L'équation de régression établie est la suivante :

Pc = 72,71(±0,98) + 11,18(±0,61) **GNar** – 16,40(±0,35) **CIC0** – 7,12(±0,85)**BIC2**

- 2,18(±0,24) **ATS8v** + 3,35(±0,55)**H5e** (101)

Tous les paramètres statistiques pertinents du modèle proposé sont explicités ci-après

R^2= 98,40 Q^2_{LOO} = 98,16 EQMP = 1,37 EQMC= 1,30

n= 125 s= 1,34 F= 1430,14

n_{ext}= 40 Q^2ext = 98,21 EQMPext = 1,22

Le modèle obtenu (Eq: 101) peut expliquer environ 98 % de la variance expérimentale de la variable dépendante (Pc) en plus il présente un F de Fischer élevé (F = 1430,14 pour un p=0,000) et une faible erreur standard (s = 1,34) ce qui confirme que ce modèle explique la pression critique d'une manière statistiquement satisfaisante.

La contribution de chaque descripteur impliqué (valeurs du test | t |) s'établit dans l'ordre suivant: CIC0 > GNar > ATS8v > BIC2 > H5e. Tous ces descripteurs ne sont pas corrélés les uns avec les autres parce que tous les VIF correspondants sont inférieurs à 5. Les caractéristiques des descripteurs sélectionnés dans le meilleur modèle MLR sont regroupées dans le tableau suivant :

Descripteur	x	Dx	t	probabilité-t	VIF
Constante	72,71	0,98	73,76	0,000	
GNar	11,18	0,61	18,33	0,000	1,84
CIC0	-16,40	0,35	-47,33	0,000	4,46
BIC2	-7,12	0,85	-8,40	0,000	1,27
ATS8v	-2,18	0,24	-9,16	0,000	2,23
H5e	3,35	0,55	6,14	0,000	3,78

Tableau - 11 : Valeurs des Pc expérimentales, calculées, prédites,ainsi que de h_{ii}, et e_{istd}

N°	**Composé**	**Pc-Exp**	**Pc-Calc**	**hii**	**ei**	**eistd**
1	Méthane	46,04	46,47	0,429	-0,43	-0,42
2	Ethane	48,84	47,99	0,174	0,85	0,70
3	Propane	42,50	40,31	0,034	2,19	1,66
4	Butane	38,00	36,69	0,027	1,31	0,99
5	Isopentane	33,81	32,71	0,020	1,10	0,83
6	2,2-Diméthylbutane	30,90	30,11	0,024	0,79	0,59
7	2,3-Diméthylbutane	31,31	31,24	0,023	0,07	0,05
8	2-Méthylpentane	30,10	29,97	0,019	0,13	0,10
9	3- Méthylpentane	31,26	30,57	0,013	0,69	0,52
10	Hexane	30,10	30,81	0,032	-0,71	-0,54
11	2,2-Diméthylpentane	27,73	27,75	0,021	-0,02	-0,02
12	2,4-Diméthylpentane	27,37	27,89	0,023	-0,52	-0,39
13	2,2,3-Triméthylbutane	29,54	28,96	0,034	0,58	0,44
14	2-Méthylhexane	27,34	27,59	0,022	-0,25	-0,19
15	3-Méthylhexane	28,14	28,33	0,016	-0,19	-0,14
16	3-Ethylpentane	28,91	28,08	0,023	0,83	0,62
17	Heptane	27,40	28,48	0,041	-1,08	-0,82
18	2,2,4-Triméthylpentane	25,68	24,95	0,034	0,73	0,56
19	2,2,3,3-Tétraméthylbutane	28,70	28,27	0,092	0,43	0,34
20	2,2-Diméthylhexane	25,29	25,34	0,027	-0,05	-0,04
21	2,5-Diméthylhexane	24,87	25,83	0,024	-0,96	-0,73

Tableau – 11 : (suite)

N°	Composé	Pc-Exp	Pc-Calc	hii	ei	eistd
22	2,4-Diméthylhexane	25,56	25,68	0,021	-0,12	-0,09
23	2,2,3-Triméthylpentane	27,30	26,98	0,055	0,32	0,24
24	3,3-Diméthylhexane	26,54	27,06	0,041	-0,52	-0,39
25	2,3,3-Triméthylpentane	28,20	27,89	0,108	0,31	0,24
26	3-Ethyl-3-méthylpentane	28,08	27,45	0,043	0,63	0,48
27	3-Ethylhexane	26,08	25,73	0,027	0,35	0,26
28	3-Méthylheptane	25,46	26,51	0,019	-1,05	-0,79
29	2,2,4,4-Tétraméthylpentane	24,85	23,34	0,054	1,51	1,16
30	2,2-Diméthylheptane	23,50	23,76	0,029	-0,26	-0,20
31	2,2,3,4-Tétraméthylpentane	26,02	25,49	0,072	0,53	0,41
32	2,2,3,3-Tétraméthylpentane	27,41	27,02	0,187	0,39	0,32
33	2-Méthyloctane	23,10	24,32	0,024	-1,22	-0,92
34	Nonane	22,88	23,90	0,027	-1,02	-0,77
35	2,2,5,5-Tétraméthylhexane	21,86	21,95	0,046	-0,09	-0,07
36	Décane	21,04	21,76	0,026	-0,72	-0,54
37	Undécane	19,66	19,91	0,030	-0,25	-0,19
38	Dodécane	18,24	18,34	0,033	-0,10	-0,07
39	Tridécane	17,20	16,83	0,039	0,37	0,28
40	Tétradécane	14,40	15,48	0,045	-1,08	-0,82
41	Hexadécane	14,10	12,98	0,058	1,12	0,86
42	Octadécane	12,00	10,73	0,071	1,27	0,98
43	2,2,4-Triméthylhexane	23,70	23,59	0,029	0,11	0,08
44	3,3-Diethylpentane	26,70	25,05	0,031	1,65	1,25
45	Pentadecane	15,20	14,16	0,051	1,04	0,79
46	Eicosane	11,10	8,66	0,084	2,44	1,90
47	E-But-2-ène	39,85	41,70	0,021	-1,85	-1,39
48	Z-But-2-ène	41,97	42,75	0,036	-0,78	-0,59
49	Pent-1-ène	35,27	34,82	0,047	0,45	0,34
50	2-Méthyl but-1-ène	34,50	35,53	0,042	-1,03	-0,79

Tableau – 11 : (suite)

N°	Composé	Pc-Exp	Pc-Calc	hii	ei	eistd
51	E-pent-2-ène	36,60	35,30	0,036	1,30	0,99
52	Z-pent-2-ène	36,50	36,49	0,040	0,01	0,01
53	2-Méthyl but-2-ène	34,50	37,77	0,018	-3,27	-2,46
54	Hex-1-ène	31,70	31,95	0,036	-0,25	-0,19
55	Hept-1-ène	28,30	29,37	0,033	-1,07	-0,81
56	Propylène	46,00	44,43	0,059	1,57	1,21
57	Isobutylene	40,00	39,79	0,029	0,21	0,16
58	3-Méthyl but-1-ène	35,10	34,96	0,030	0,14	0,11
59	Z-Hex-2-ène	32,80	31,97	0,038	0,83	0,63
60	E-Hex-2-ène	32,70	32,83	0,039	-0,13	-0,10
61	Z-Hex-3-ène	32,80	33,07	0,025	-0,27	-0,20
62	E-Hex-3-ène	32,50	34,13	0,011	-1,63	-1,22
63	2-Méthyl pent-2-ène	32,80	32,92	0,022	-0,12	-0,09
64	Z-3-Méthyl pent-2-ène	32,80	34,05	0,067	-1,25	-0,96
65	E-3-Méthyl pent-2-ène	32,90	32,28	0,023	0,62	0,46
66	Z-4-Méthyl pent-2-ène	30,40	31,53	0,034	-1,13	-0,86
67	E-4-Méthyl pent-2-ène	30,40	32,12	0,022	-1,72	-1,30
68	2,3-Diméthyl but-2-ène	33,60	35,90	0,046	-2,30	-1,75
69	2,3-Diméthyl but-1-ène	32,40	31,77	0,023	0,63	0,48
70	3,3-Diméthyl but-1-ène	32,50	31,91	0,025	0,59	0,45
71	2,3,3-Triméthyl but-1-ène	28,90	29,65	0,022	-0,75	-0,57
72	Non-1-ène	23,40	24,32	0,026	-0,92	-0,69
73	Undec-1-ène	19,90	20,23	0,032	-0,33	-0,25
74	Dodec-1-ène	18,50	18,60	0,037	-0,10	-0,08
75	Tridec-1-ène	17,00	17,10	0,042	-0,10	-0,07
76	Hexadec-1-ène	13,30	13,21	0,062	0,09	0,07
77	Octadec-1-ène	11,30	10,95	0,075	0,35	0,27
78	Cyclobutane	49,80	49,51	0,126	0,29	0,23
79	Cyclopentane	45,08	44,37	0,109	0,71	0,56

Tableau – 11 : (suite)

N°	Composé	Pc-Exp	Pc-Calc	hii	ei	eistd
80	Méthylcyclopentane	37,84	36,87	0,032	0,97	0,74
81	Cyclohexane	40,70	40,17	0,105	0,53	0,42
82	Méthylcyclohexane	34,71	34,44	0,027	0,27	0,20
83	Ethylcyclopentane	33,97	34,34	0,027	-0,37	-0,28
84	E-1,4-Diméthylcyclohexane	29,70	31,80	0,029	-2,10	-1,59
85	Cyclooctane	35,60	34,86	0,085	0,74	0,58
86	Cyclopropane	54,90	56,07	0,167	-1,17	-0,96
87	1,1Diméthylcyclopentane	34,40	32,94	0,022	1,46	1,10
88	Z-1,2-Diméthylcyclopentane	34,40	32,79	0,022	1,61	1,21
89	E-1,2-Diméthylcyclopentane	34,40	33,43	0,019	0,97	0,73
90	E-1,2-Diméthylcyclohexane	29,60	31,08	0,021	-1,48	-1,11
91	Z-1,3-Diméthylcyclohexane	29,60	30,71	0,022	-1,11	-0,84
92	Ethylcyclohexane	30,00	32,25	0,029	-2,25	-1,70
93	1,1,2-Triméthylcyclopentane	29,40	29,75	0,016	-0,35	-0,26
94	1,1,3-Triméthylcyclopentane	28,30	28,83	0,025	-0,53	-0,40
95	Propylcyclopentane	30,00	31,28	0,031	-1,28	-0,97
96	Isopropylcyclopentane	30,00	31,76	0,039	-1,76	-1,34
97	Propylcyclohexane	28,00	29,65	0,029	-1,65	-1,25
98	1,E-3,5-Triméthylcyclohexane	28,00	27,63	0,024	0,37	0,28
99	Butylcyclohexane	31,50	27,81	0,035	3,69	2,80
100	Isobutylcyclohexane	31,20	26,74	0,029	4,46	3,37
101	sec-Butylcyclohexane	26,70	26,99	0,031	-0,29	-0,22
102	tert-Butylcyclohexane	26,60	26,58	0,032	0,02	0,01
103	Hexylcyclopentane	21,30	23,16	0,032	-1,86	-1,41
104	Heptylcyclopentane	19,40	20,85	0,041	-1,45	-1,10
105	Decylcyclopentane	15,20	16,07	0,053	-0,87	-0,67
106	Decylcyclohexane	13,50	15,05	0,058	-1,55	-1,19
107	Tridecylcyclopentane	12,00	12,37	0,069	-0,37	-0,29
108	1-Cyclopentyltetradecane	11,20	11,17	0,074	0,03	0,02

Tableau – 11 : (suite)

N°	Composé	Pc-Exp	Pc-Calc	hii	ei	eistd
109	1-Cyclopentylpentadecane	10,20	10,17	0,080	0,03	0,02
110	Buta-1,3-diène	43,30	46,00	0,034	-2,70	-2,05
111	Buta-1,2-diène	44,90	45,00	0,056	-0,10	-0,07
112	Penta-1,4-diène	37,90	40,15	0,024	-2,25	-1,69
113	3-Méthyl buta-1,2-diène	41,10	40,12	0,019	0,98	0,74
114	E-penta-1,3-diène	39,90	39,10	0,042	0,80	0,61
115	Penta-1,2-diène	40,70	39,42	0,041	1,28	0,97
116	Propadiène	54,70	53,64	0,076	1,06	0,82
117	Deca-1,3-diène	22,30	22,10	0,028	0,20	0,15
118	1-Méthylcyclopentène	38,94	38,90	0,046	0,04	0,03
119	Cyclohexène	43,40	40,75	0,050	2,65	2,02
120	1-Ethylcyclopentène	33,77	35,83	0,037	-2,06	-1,56
121	Cyclopentène	48,00	45,53	0,069	2,47	1,91
122	Acétylène	61,39	64,42	0,184	-3,03	-2,50
123	Propyne	56,28	51,87	0,082	4,41	3,43
124	But-2-yne	50,80	48,93	0,083	1,87	1,46
125	Vinylacétylène	49,60	49,82	0,088	-0,22	-0,17
126*	Isobutane	36,50	36,22	0,035	0,28	0,22
127*	Pentane	33,64	33,21	0,022	0,43	0,34
128*	3,3-Diméthylpentane	29,46	29,28	0,022	0,18	0,14
129*	2,3-Diméthylpentane	29,08	28,99	0,019	0,09	0,07
130*	2,3,4-Triméthylpentane	27,30	27,95	0,037	-0,65	-0,51
131*	2,3-Diméthylhexane	26,28	26,63	0,019	-0,35	-0,28
132*	3-Ethyl-2-méthylpentane	27,00	25,83	0,015	1,17	0,91
133*	2-Méthylheptane	24,84	25,68	0,018	-0,84	-0,65
134*	3,4-Diméthylhexane	26,92	27,48	0,031	-0,56	-0,44
135*	4-Méthylheptane	25,42	26,41	0,013	-0,99	-0,77
136*	Octane	24,93	26,75	0,025	-1,82	-1,42
137*	2,2,5-Triméthylhexane	23,30	23,25	0,024	0,05	0,04

Tableau – 11 : (suite et fin)

N°	Composé	Pc-Exp	Pc-Calc	hii	ei	eistd
138*	2,3,3,4-Tétraméthylpentane	27,16	26,78	0,088	0,38	0,31
139*	2,2,3,3-Tétraméthylhexane	25,10	24,62	0,088	0,48	0,39
140*	3,3,5-Triméthylheptane	23,17	23,61	0,053	-0,44	-0,35
141*	2,2-Diméthylpropane	32,00	32,60	0,062	-0,60	-0,48
142*	2,2,3-Triméthylhexane	24,90	24,62	0,029	0,28	0,22
143*	Heptadecane	13,00	11,78	0,047	1,22	0,96
144*	Nonadecane	11,10	9,65	0,056	1,45	1,15
145*	Ethylène	50,41	53,89	0,088	-3,48	-2,81
146*	But-1-ène	40,23	39,11	0,036	1,12	0,88
147*	Oct-1-ène	26,20	27,37	0,021	-1,17	-0,92
148*	E-Oct-2-ène	27,70	27,37	0,024	0,33	0,26
149*	Dec-1-ène	22,00	23,58	0,034	-1,58	-1,24
150*	Tétradec-1-ène	15,60	15,74	0,035	-0,14	-0,11
151*	Pentadec-1-ène	14,50	14,43	0,039	0,07	0,05
152*	Cycloheptane	38,10	37,60	0,073	0,50	0,40
153*	1,1-Diméthylcyclohexane	29,60	30,23	0,017	-0,63	-0,49
154*	Z-1,2-Diméthylcyclohexane	29,60	30,85	0,016	-1,25	-0,97
155*	E-1,3-Diméthylcyclohexane	29,70	29,24	0,032	0,46	0,36
156*	Z-1,4-Diméthylcyclohexane	29,70	31,68	0,021	-1,98	-1,55
157*	1-Ethyl-1-méthylcyclopentane	30,00	30,60	0,015	-0,60	-0,46
158*	Isopropylcyclohexane	28,30	29,06	0,021	-0,76	-0,59
159*	Octylcyclopentane	17,90	19,00	0,033	-1,10	-0,86
160*	Nonylcyclopentane	16,50	17,57	0,035	-1,07	-0,84
161*	Dodecylcyclopentane	12,90	13,47	0,046	-0,57	-0,45
162*	Hexa-1,5-diène	34,40	36,25	0,013	-1,85	-1,44
163*	2-Méthyl buta-1,3-diène	38,50	38,99	0,024	-0,49	-0,39
164*	But-1-yne	47,10	44,51	0,041	2,59	2,04
165*	Pent-1-yne	40,50	39,33	0,033	1,17	0,92

*Composés de validation

Qualité de l'ajustement

La qualité de l'ajustement est représentée par le graphe des valeurs calculées de l'ensemble de calibrage et les valeurs prédites de l'ensemble de validation en fonction de celles des valeurs Pc mesurées.

$$Pc\text{-}Calc = 0{,}5 + 0{,}98\ Pc\text{-}Exp \qquad (102)$$

$$s = 1{,}33 \qquad R^2 = 98{,}40\ \% \qquad R^2_{ajust} = 98{,}30\ \%$$

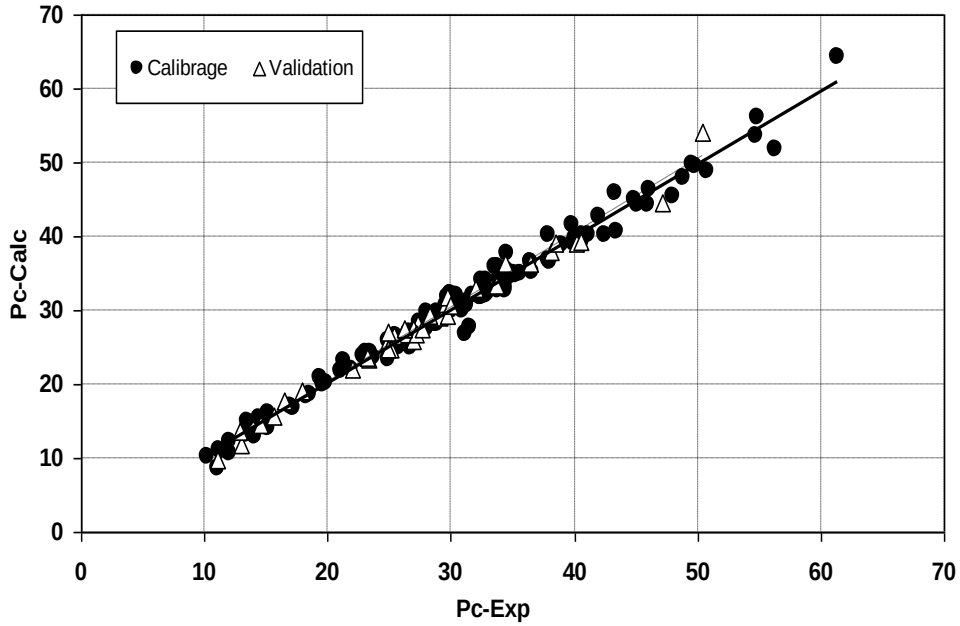


Figure 31: Graphe des valeurs Pc calculées (tests et prédites) en fonction des valeurs observées.

La figure ci-dessus montre que les points sont peu dispersés autour de la première bissectrice.

La représentation des valeurs estimées de la pression critique en fonction de celles observées conduit à l'équation (102) qui ne s'écarte pas de façon significative de l'ordonnée à l'origine puisque [-0,11 <a =0,5 <1,11] et possède une pente (b) très peu différente de l'unité puisque [0,96 < b=0,98 < 1,00].

Tous les résidus standardisés de prédiction sont compris entre les limites ± 3 à l'exception des résidus de trois composés de l'ensemble de calibrage (97, 98 et 120), tandis que trois autres composés (1, 86 et 119) présentent un bras de levier important ($h_i > h^* = 0{,}144$).

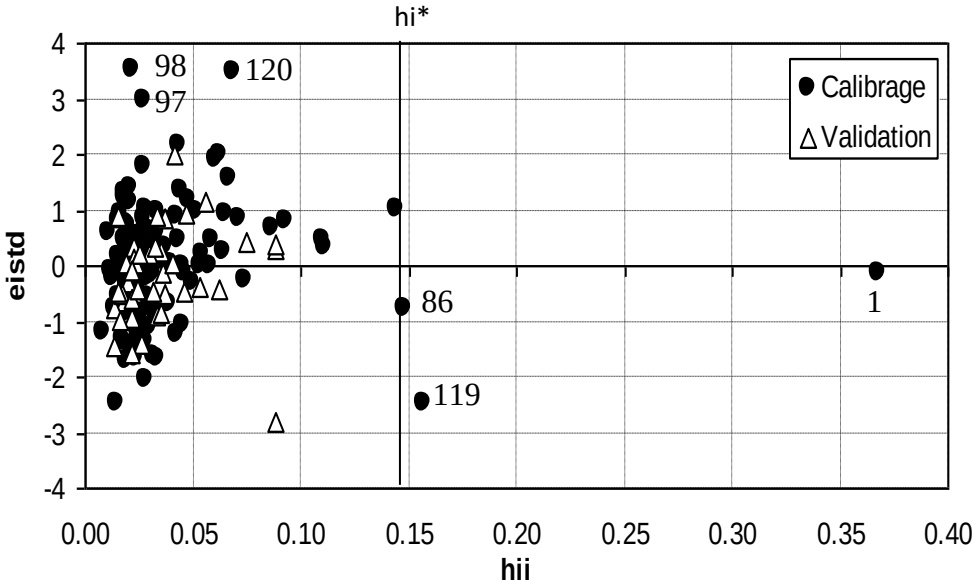


Figure 32: *Diagramme de Williams*

Par le test de randomisation (figure 33), nous nous sommes assuré qu'une relation structure-propriété réelle a été établie par notre modèle.

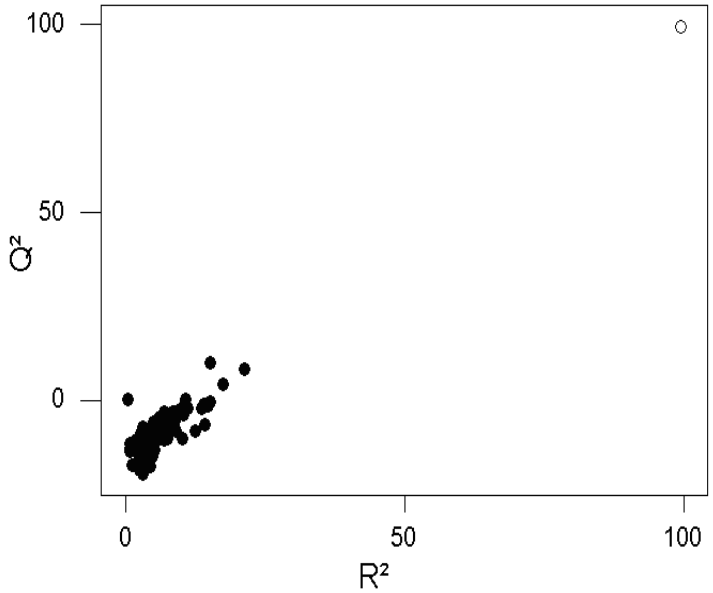


Figure 33: *Test de randomisation*

Encore une fois les autres paramètres statistiques montrent la bonne prédictivité du modèle, soit :

Paramètre	R^2ext	Q^2_{F1}	Q^2_{F2}	Q^2_{F3}	CCC
Valeur	0,9817	0,9843	0,9825	0,9876	0,9914

Avec:

1- Q^2_{ext} = 0,9875 > 0,5

2- R^2= 0,9878 > 0,6

3- 0,85 < k =0,9998 < 1,15

4- 0,85 < k'=0,9983 < 1,15

III-2-2- Régression par les réseaux de neurones artificiels RNA

Choix des paramètres optimaux

Pour fixer le nombre de neurones de la couche cachée le graphe de la figure 34 permet de constater que 6 est la valeur optimale pour une RMSE minimale, tandis que 60 époques sera le nombre optimal d'itérations mis en évidence sur la figure 35.

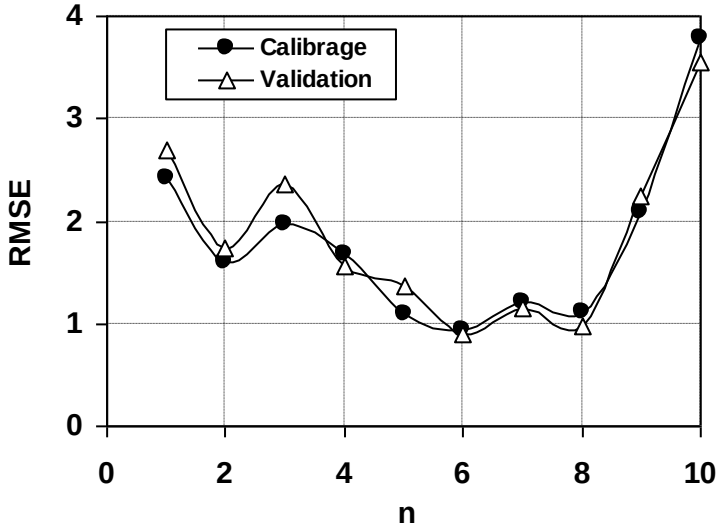


Figure 34: Choix du nombre de neurones de la couche cachée

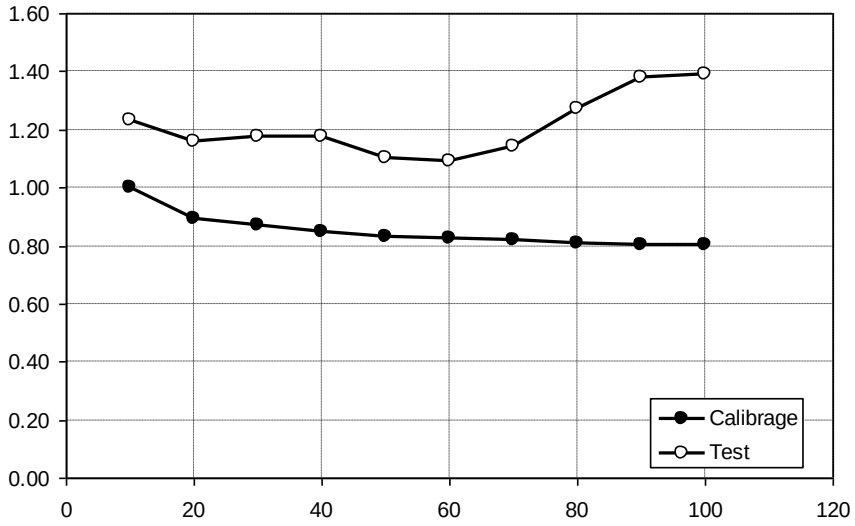


Figure 35: Choix du nombre d'itérations

Finalement, nous avons opté la structure du réseau explicitée dans le tableau 12.

Tableau - 12: Structure optimale du réseau de neurones.

Nombre d'entrées	05 (les descripteurs)
Nombre de sorties	01 (la pression critique)
Nombre de couches cachées	Une couche cachée
Nombre de neurones dans la couche cachée	06
Nombre d'itérations	60
Algorithme d'apprentissage	Rétro propagation du gradient de l'erreur
Fonctions d'apprentissage	Tangente hyperbolique

Tous les résultats du traitement sont présentés dans le tableau 13

Tableau - 13: Valeurs des paramètres statistiques

Ensemble de Calibrage (125) et de			Tests (40)	Ensemble de Validation (40)	
R^2 (%)	s	RMSEcal	RMSEtst	R^2cv (%)	RMSEval
99,34	0,8503	0,8597	0,8800	98,85	0,9876

$$Pc\text{-}Calc = 0{,}05 + 0{,}99\ Pc\text{-}Exp \qquad (103)$$

$S = 0{,}85 \qquad R^2 = 99{,}00\ \% \qquad R^2_{ajust} = 98{,}67\ \%$

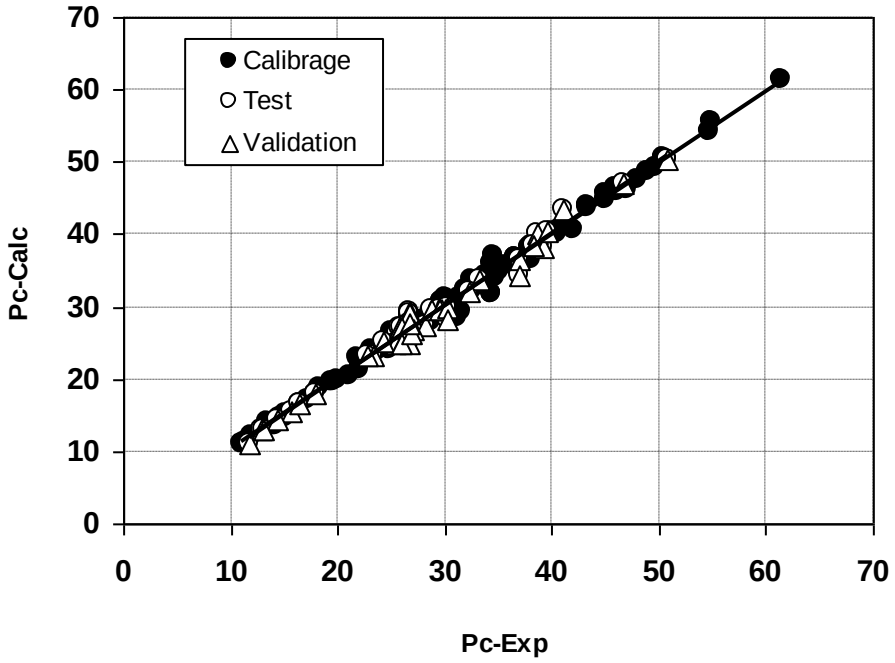
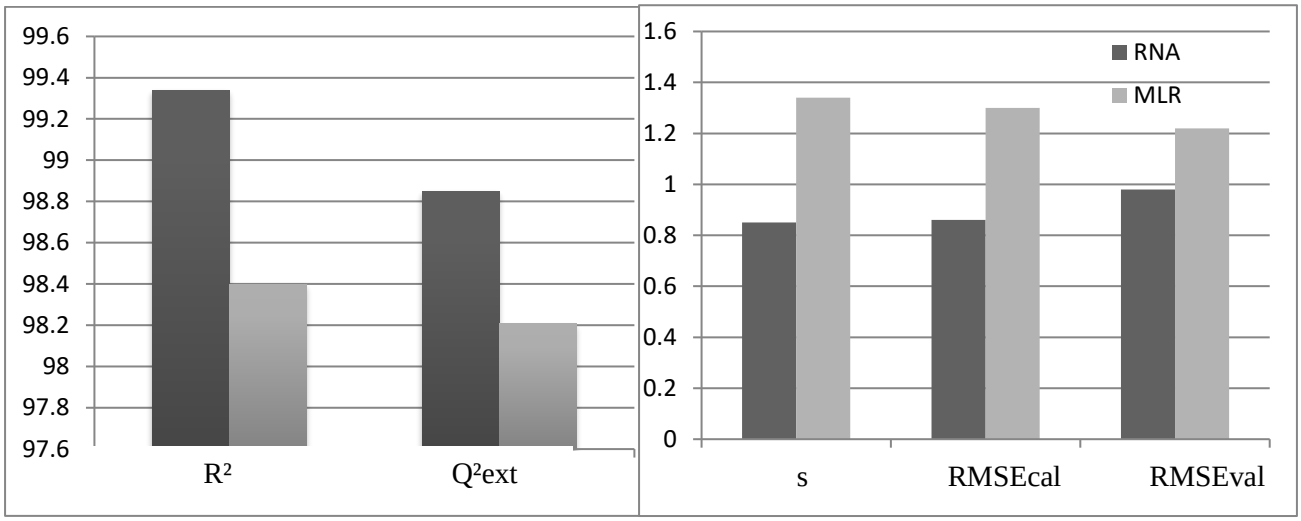


Figure 36 : Graphe des valeurs Pc calculées, tests et prédites, en fonction des valeurs observées

Les paramètres statistiques ont été comparés par l'histogramme suivant qui montre nettement la supériorité de la méthode non linéaire.



D'après la figure 37 on remarque qu'il y a 10 résidus supérieurs à 2s (1.7) soit 8 % de l'ensemble de calibrage et 3 résidus de l'ensemble de validation (soit 7,5 %).

Rappelons, également que les composés 112 (Penta-1,4-diène) et 145(Ethylène) présentent les plus grand résidus avec comme valeurs -2,37 et – 3,17 (tableau 14)

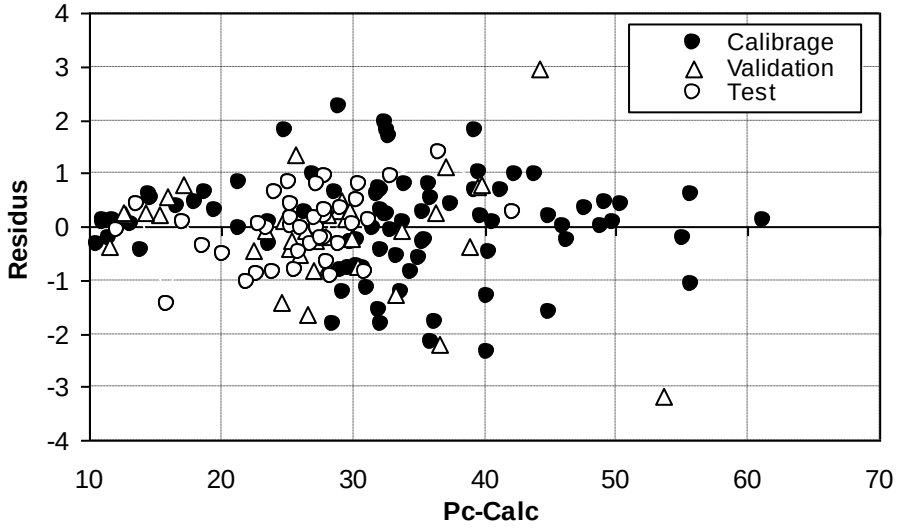


Figure 37: Variation des résidus en fonction des valeurs calculées.

Tableau - 14: Valeurs de Pc expérimentales, calculées, prédites, et des résidus.

N	Composé	Pc-Exp	Pc-Calc	ei
1	Méthane	46,04	46,04	0,00
2	Ethane	48,84	48,84	0,00
3	Propane	42,50	42,24	0,26
4	Butane	38,00	36,63	1,37
5	Isopentane	33,81	32,87	0,94
6	2,2-Diméthylbutane	30,90	30,43	0,47
7	2,3-Diméthylbutane	31,31	31,20	0,11
8	2-Méthylpentane	30,10	30,06	0,04
9	3- Méthylpentane	31,26	30,49	0,77
10	Hexane	30,10	30,95	-0,85
11	2,2-Diméthylpentane	27,73	27,96	-0,23
12	2,4-Diméthylpentane	27,37	28,03	-0,66
13	2,2,3-Triméthylbutane	29,54	29,20	0,34
14	2-Méthylhexane	27,34	27,36	-0,02
15	3-Méthylhexane	28,14	27,97	0,17
16	3-Ethylpentane	28,91	27,97	0,94
17	Heptane	27,40	28,35	-0,95
18	2,2,4-Triméthylpentane	25,68	25,27	0,41
19	2,2,3,3-Tétraméthylbutane	28,70	29,02	-0,32
20	2,2-Diméthylhexane	25,29	25,29	0,00
21	2,5-Diméthylhexane	24,87	25,68	-0,81
22	2,4-Diméthylhexane	25,56	25,41	0,15
23	2,2,3-Triméthylpentane	27,30	27,14	0,16
24	3,3-Diméthylhexane	26,54	26,89	-0,35
25	2,3,3-Triméthylpentane	28,20	27,89	0,31
26	3-Ethyl-3-méthylpentane	28,08	27,31	0,77
27	3-Ethylhexane	26,08	25,26	0,82
28	3-Méthylheptane	25,46	25,95	-0,49
29	2,2,4,4-Tétraméthylpentane	24,85	24,20	0,65

Tableau – 14: (suite)

N	Composé	Pc-Exp	Pc-Calc	ei
30	2,2-Diméthylheptane	23,50	23,55	-0,05
31	2,2,3,4-Tétraméthylpentane	26,02	26,05	-0,03
32	2,2,3,3-Tétraméthylpentane	27,41	27,63	-0,22
33	2-Méthyloctane	23,10	23,97	-0,87
34	Nonane	22,88	22,86	0,02
35	2,2,5,5-Tétraméthylhexane	21,86	22,76	-0,90
36	Décane	21,04	22,07	-1,03
37	Undécane	19,66	20,17	-0,51
38	Dodécane	18,24	18,60	-0,36
39	Tridécane	17,20	17,11	0,09
40	Tétradécane	14,40	15,87	-1,47
41	Hexadécane	14,10	13,69	0,41
42	Octadécane	12,00	12,09	-0,09
43	2,2,4-Triméthylhexane	23,70	23,61	0,09
44	3,3-Diethylpentane	26,70	24,92	1,78
45	Pentadecane	15,20	14,68	0,52
46	Eicosane	11,10	11,02	0,08
47	E-But-2-ène	39,85	40,33	-0,48
48	Z-But-2-ène	41,97	41,29	0,68
49	Pent-1-ène	35,27	35,55	-0,28
50	2-Méthyl but-1-ène	34,50	35,08	-0,58
51	E-pent-2-ène	36,60	35,83	0,77
52	Z-pent-2-ène	36,50	35,98	0,52
53	2-Méthyl but-2-ène	34,50	36,31	-1,81
54	Hex-1-ène	31,70	32,15	-0,45
55	Hept-1-ène	28,30	29,11	-0,81
56	Propylène	46,00	46,28	-0,28
57	isobutylène	40,00	39,32	0,68
58	3-Méthyl but-1-ène	35,10	35,39	-0,29

Tableau – 14: (suite)

N	Composé	Pc-Exp	Pc-Calc	ei
58	3-Méthyl but-1-ène	35,10	35,39	-0,29
59	Z-Hex-2-ène	32,80	32,10	0,70
60	E-Hex-2-ène	32,70	32,48	0,22
61	Z-Hex-3-ène	32,80	33,37	-0,57
62	E-Hex-3-ène	32,50	33,72	-1,22
63	2-Méthyl pent-2-ène	32,80	32,56	0,24
64	Z-3-Méthyl pent-2-ène	32,80	32,88	-0,08
65	E-3-Méthyl pent-2-ène	32,90	32,21	0,69
66	Z-4-Méthyl pent-2-ène	30,40	31,97	-1,57
67	E-4-Méthyl pent-2-ène	30,40	32,23	-1,83
68	2,3-Diméthyl but-2-ène	33,60	34,47	-0,87
69	2,3-Diméthyl but-1-ène	32,40	31,80	0,60
70	3,3-Diméthyl but-1-ène	32,50	32,21	0,29
71	2,3,3-Triméthyl but-1-ène	28,90	29,70	-0,80
72	Non-1-ène	23,40	23,74	-0,34
73	Undec-1-ène	19,90	19,61	0,29
74	Dodec-1-ène	18,50	18,05	0,45
75	Tridec-1-ène	17,00	16,63	0,37
76	Hexadec-1-ène	13,30	13,26	0,04
77	Octadec-1-ène	11,30	11,52	-0,22
78	Cyclobutane	49,80	49,73	0,07
79	Cyclopentane	45,08	44,88	0,20
80	Méthylcyclopentane	37,84	37,42	0,42
81	Cyclohexane	40,70	40,64	0,06
82	Méthylcyclohexane	34,71	33,94	0,77
83	Ethylcyclopentane	33,97	33,91	0,06
84	E-1,4-Diméthylcyclohexane	29,70	30,43	-0,73
85	Cyclooctane	35,60	35,32	0,28
86	Cyclopropane	54,90	55,12	-0,22

Tableau – 14: (suite)

N	Composé	Pc-Exp	Pc-Calc	ei
87	1,1-Diméthylcyclopentane	34,40	32,61	1,79
88	Z-1,2-Diméthylcyclopentane	34,40	32,44	1,96
89	E-1,2-Diméthylcyclopentane	34,40	32,72	1,68
90	E-1,2-Diméthylcyclohexane	29,60	29,91	-0,31
91	Z-1,3-Diméthylcyclohexane	29,60	29,89	-0,29
92	Ethylcyclohexane	30,00	31,15	-1,15
93	1,1,2-Triméthylcyclopentane	29,40	28,75	0,65
94	1,1,3-Triméthylcyclopentane	28,30	28,11	0,19
95	Propylcyclopentane	30,00	30,79	-0,79
96	Isopropylcyclopentane	30,00	30,28	-0,28
97	Propylcyclohexane	28,00	29,24	-1,24
98	1,E-3,5-Triméthylcyclohexane	28,00	27,04	0,96
99	Butylcyclohexane	31,50	31,52	-0,02
100	Isobutylcyclohexane	31,20	28,96	2,24
101	sec-Butylcyclohexane	26,70	28,54	-1,84
102	tert-Butylcyclohexane	26,60	26,33	0,27
103	Hexylcyclopentane	21,30	21,33	-0,03
104	Heptylcyclopentane	19,40	18,75	0,65
105	Decylcyclopentane	15,20	14,60	0,60
106	Decylcyclohexane	13,50	13,94	-0,44
107	Tridecylcyclopentane	12,00	11,87	0,13
108	1-Cyclopentyltetradecane	11,20	11,08	0,12
109	1-Cyclopentylpentadecane	10,20	10,54	-0,34
110	Buta-1,3-diène	43,30	44,90	-1,60
111	Buta-1,2-diène	44,90	43,93	0,97
112	Penta-1,4-diène	37,90	40,27	-2,37
113	3-Méthyl buta-1,2-diène	41,10	39,30	1,80
114	E-penta-1,3-diène	39,90	39,73	0,17
115	Penta-1,2-diène	40,70	39,68	1,02

Tableau – 14: (suite)

N	Composé	Pc-Exp	Pc-Calc	ei
116	Propadiène	54,70	55,78	-1,08
117	Deca-1,3-diène	22,30	21,46	0,84
118	1-Méthylcyclopentène	38,94	40,23	-1,29
119	Cyclohexène	43,40	42,43	0,97
120	1-Ethylcyclopentène	33,77	35,92	-2,15
121	cyclopentène	48,00	47,68	0,32
122	Acétylène	61,39	61,26	0,13
123	Propyne	56,28	55,67	0,61
124	But-2-yne	50,80	50,40	0,40
125	vinylacétylène	49,60	49,15	0,45
126*	Isobutane	36,50	36,22	0,28
127*	Pentane	33,64	33,70	-0,06
128*	3,3-Diméthylpentane	29,46	29,16	0,30
129*	2,3-Diméthylpentane	29,08	28,78	0,30
130*	2,3,4-Triméthylpentane	27,30	28,06	-0,76
131*	2,3-Diméthylhexane	26,28	26,37	-0,09
132*	3-Ethyl-2-méthylpentane	27,00	25,66	1,34
133*	2-Méthylheptane	24,84	25,25	-0,41
134*	3,4-Diméthylhexane	26,92	27,17	-0,25
135*	4-Méthylheptane	25,42	25,95	-0,53
136*	Octane	24,93	26,58	-1,65
137*	2,2,5-Triméthylhexane	23,30	23,38	-0,08
138*	2,3,3,4-Tétraméthylpentane	27,16	27,30	-0,14
139*	2,2,3,3-Tétraméthylhexane	25,10	25,36	-0,26
140*	3,3,5-Triméthylheptane	23,17	24,59	-1,42
141*	2,2-Diméthylpropane	32,00	33,26	-1,26
142*	2,2,3-Triméthylhexane	24,90	24,77	0,13
143*	Heptadecane	13,00	12,78	0,22
144*	Nonadecane	11,10	11,46	-0,36

Tableau – 14: (suite et fin)

N	Composé	Pc-Exp	Pc-Calc	ei
145*	Ethylène	50,41	53,58	-3,17
146*	But-1-ène	40,23	39,47	0,76
147*	Oct-1-ène	26,20	27,01	-0,81
148*	E-Oct-2-ène	27,70	27,90	-0,20
149*	Dec-1-ène	22,00	22,46	-0,46
150*	Tétradec-1-ène	15,60	15,39	0,21
151*	Pentadec-1-ène	14,50	14,24	0,26
152*	Cycloheptane	38,10	36,98	1,12
153*	1,1-Diméthylcyclohexane	29,60	29,44	0,16
154*	Z-1,2-Diméthylcyclohexane	29,60	29,84	-0,24
155*	E-1,3-Diméthylcyclohexane	29,70	29,21	0,49
156*	Z-1,4-Diméthylcyclohexane	29,70	30,43	-0,73
157*	1-Ethyl-1-méthylcyclopentane	30,00	29,70	0,30
158*	Isopropylcyclohexane	28,30	28,08	0,22
159*	Octylcyclopentane	17,90	17,12	0,78
160*	Nonylcyclopentane	16,50	15,93	0,57
161*	Dodecylcyclopentane	12,90	12,63	0,27
162*	Hexa-1,5-diène	34,40	36,59	-2,19
163*	2-Méthyl buta-1,3-diène	38,50	38,87	-0,37
164*	But-1-yne	47,10	44,14	2,96
165*	Pent-1-yne	40,50	39,70	0,80

CONCLUSION GENERALE

CONCLUSION GENERALE

Nous avons appliqué la méthodologie QSPR pour relier trois propriétés (température d'ébullition ; température et pression critiques) d'un mélange hétérogène d'hydrocarbures, comportant 1 à 18 atomes de carbone ; à des descripteurs moléculaires théoriques reflétant certaines particularités, à chaînes ouvertes (linéaires ou ramifiées) ou fermées (du cyclobutane au cyclooctane) dont plusieurs isomères [de chaîne ; de position (liaisons (s) multiples (s) ; doubles liaisons alternées ou cumulées…)].

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

Les 165 données de base ont été éclatées aléatoirement en deux ensembles disjoints, invariants pour tous les modèles :

- un ensemble principal de 125 composés utilisés pour le calcul et, éventuellement, les essais du modèle ;
- un ensemble de 40 composés pour la prédiction externe.

La taille d'un modèle est fixée par la valeur optimale en utilisant la variation de coefficient de détermination R^2 en fonction de la variation de descripteurs (point de brisure). La sélection des variables explicatives a été réalisée par algorithme génétique, dans la version MOBYDIGS de TODESCHINI, en maximisant Q^2_{L00}.

Les statistiques réunies ci-après permettent de faire des comparaisons, et de tirer plusieurs conclusions.

		Calibrage				Validation	
		(I)	(II)	(III)	(IV)	(V)	(VI)
		R^2 (%)	**S**	EQMC	**Q^2 (%)**	**Q^2_{ext} (%)**	**$EQMP_{ext}$**
Tc	(MLR)	99,30	8,72	8,51	99,17	99,42	7,65
	(RNA)	99,83	4,19	4,17	-	99,59	6,31
Teb	(MLR)	99,77	4,67	4,80	99,70	99,57	5,52
	(RNA)	99,92	2,27	2,57	-	99,86	3,54
Pc	(MLR)	98,40	1,34	1,30	98,16	98,21	1,22
	(RNA)	99,34	0,85	0,86	-	98,85	0,98

Les statistiques calculées établissent la pertinence de chaque modèle développé.

A chaque fois, la qualité de l'ajustement a été vérifiée en procédant à une validation croisée par "leave – one - out". Les grandes valeurs de Q^2 obtenues (proches de 100 % - colonne (IV)) reproduisent pratiquement celles du coefficient de détermination multiple correspondant (colonne (I)), ce qui fait ressortir la qualité de l'ajustement pour chaque modèle obtenu.

Le test de randomisation montre, dans tous les cas (Figures 17, 25, 33), que seul le vecteur réel des observations conduit à des valeurs élevées des statistiques R^2 et Q^2, ce qui prouve que les modèles obtenus ne sont pas aléatoires.

Les valeurs RMSE réunies dans les colonnes (III) et (VI) sont faibles et proches les unes des autres pour chaque modèle, ce qui permet, tout à la fois, de s'assurer de la bonne capacité prédictive et de la possibilité d'extension suffisante de chaque modèle.

Ainsi, les propriétés physicochimiques (Teb ; Tc ; Pc) peuvent être prédites à partir de leur structure moléculaire en utilisant la régression multilinéaire, ou la modélisation non linéaire par les réseaux de neurones, qui semble, d'ailleurs, plus performante.

Ce travail sera étendu à d'autres caractéristiques en introduisant des paraffines à hauts poids moléculaires et, dans une étape ultérieure, en diversifiant la structure des ensembles de données qui seront choisis les plus étendus possibles.

Le choix aléatoire de l'ensemble d'essai pouvant influencer négativement (extrapolation) la capacité prédictive du modèle, nous appliquerons une approche multivariée (qui dérive de l'analyse en composantes principales) pour extraire un ensemble réduit représentatif, à partir d'un ensemble de données plus large.

Les limitations des modèles doivent être clairement définies, et l'existence des points aberrants analysée avec soins.

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

REFERENCES BIBLIOGRAPHIES

REFERENCES BIBLIOGRAPHIQUES

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

2- D. R. Hatree, (1928) ″The Wave Mechanics of an Atom tvith a Non-Voulomb Ventral″, Proc. Cambridge. Phill. Soc. Vol, 24, pp, 328.

3- V. Fock, (1930)″Näherungsmethode zur lösung des quantenmechanischen mehrkörperproblems″, Z. Physik. Vol, 61, pp, 126.

4- J. C. Slater, (1930)″ The self consistent field and the structure of atoms″, Phys. Rev. Vol, 32, pp, 339.

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

6- R. Pariser, R. G. Parr, (1953) ″A Semi Empirical Theory of the Electronic Spectra and Electronic Structure of Complex Unsaturated Molecules I″, J. Chem. Phys. Vol, 21, pp, 466-477.

7- R. Pariser, R. G. Parr, (1953) A Semi Empirical Theory of the Electronic Spectra and Electronic Structure of Complex Unsaturated Molecules II, J. Chem. Phys. Vol, 21, pp, 767-776.

8- J. A. Pople, (1953) ″Electron interaction in unsaturated hydrocarbons″, Trans. Faraday Soc. Vol, 49, pp, 1375-1385.

9- J. A. Pople, D. P. Santry, G. A. Segal, (1965) ″Approximate Self Consistent Molecular Orbital Theory. I. Invariant Procedures″. J. Chem. Phys. Vol, 43, pp, 5129.

10- A. Pople, D. L. Beveridge, P. A. Dobosh, (1967) ″Approximate Self Consistent Molecular Orbital Theory. V. Intermediate Neglect of Differential Overlap″, J. Chem. Phys. Vol, 47, pp, 2026-2033.

11- M. J. S. Dewar, W. Thiel, (1977) ″Ground states of molecules. The MNDO method. Approximations and Parameters″, J. Am. Chem. Soc., Vol, 99, pp, 4899-4917 ; MNDO: W. Theil, (1998) Encyclopedia of Computational Chemistry, P. v. R. Schleyer, N. L. Allinger, T. Clarck, J. Gasteiger, P. A. Kollman, H. F. Schaefer III, P. R. Schreiner (Eds.), Wiley, Chichester, Vol. 3, pp, 1599.

12- K. Y. Burstein, A.N. Isaev, (1984) ″MNDO calculations on hydrogen bonds. Modified function for core-core repulsion″, Theor. Chim. Acta. Vol, 64, pp, 397- 401.

13- M. J. S. Dewar, E. G. Zoebisch, E. F. Healy, J. J. P. Stewart, (1985)″Development and use of quantum mechanical molecular models. 76. AM1: a new general purpose quantum mechanical molecular model″, J. Am. Chem. Soc, Vol, 107, pp, 3902- 3909. AM1: A. J. Holder, (1998) Encyclopedia of Computational Chemistry, P. v. R. Schleyer, N. L. Allinger, T. Clark, J. Gasteiger, P. A. Kollman, H. F. Schaefer III, P. R. Schreiner (Eds.), Wiley, Chichester, Vol, 1, pp, 8- 11.

14- J. J. P. Stewart, (1989) ″Optimization of parameters for semi empirical methods I. Method″, J. Comput. Chem., Vol, 10, pp, 209- 220 ; 221- 264; PM3 : J. J. P. Stewart, (1998) Encyclopedia of Computational Chemistry, P. v. R. Schleyer, N. L. Allinger, T. Clark, J. Gasteiger, P. A. Kollman, H. F. Schaefer III, P. R. Schreiner (Eds.), Wiley, Chichester, Vol, 3, pp, 2080-2086.

15- F. Jensen, (1998) ″Introduction to Computational Chemistry″, Wiley, pp, 94- 96.

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

17- A. D. Daniels, J. M. Millam, G. E. Scuseria, (1997) ″Semi empirical methods with conjugate gradient density matrix search to replace diagonalization for molecular systems containing thousands of atoms″, J. Chem. Phys., Vol, 107,pp, 425.

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

19- C. A. Coulson, H. C. Longuet- Higgins, (1947) ″The Electronic Structure of Conjugated Systems. I. General theory ″, Proc. Roy. Soc. (London) A 191, p, 39.

20- R. S. Mulliken, (1962) ″Criteria for the Construction of Good Self Consistent Field Molecular Orbital Wave Functions, and the Significance of LCAOMO Population Analysis″, J. Chem. Phys., Vol, 36, p, 3428.

21- W. Kutzelnigg, G. Delre, G. Bertheir. (1971) ″σ and π Electrons in Theoretical Organic Chemistry″, Springer Verlag, Berlin.

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

23- D. B. Boyd, K. B. Lipkowitz, (2000) ″eds. Reviews in Computational Chemistry, History of the Gordon Conferences on Computational Chemistry″,Wiley- VCH, New York, pp, 399- 439.

24- N. L. Allinger, (1977) ″Conformational analysis. 130. MM2. A hydrocarbon force field utilizing V1 and V2 torsional terms″. J. Am. Chem. Soc., Vol, 99, pp, 8127- 8134.

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

26- N. L. Allinger, Y. H. Yuk, J. –H. Lii, (1989) ″Molecular Mxhanics. The MM3 Force Field for Hydrocarbon 3. 1″, J. Am. Chem. Soc., Vol, 111, pp, 8551- 8565.

27- N. L. Allinger. K. Chem, J. A. Katzenellbogen, S. R. Wilson, G. M. Anstead, (1996) ″Hyperconjugative Effects on Carbon-Carbon Bond Lengths in Molecular Mechanics (MM4) ″, J. Comput. Chem., Vol, 17, pp, 747- 755.

28- A. D. Mac Kerell, Jr., D. Bashford, M. Bellott, R. L. Dumbrack, Jr., J. D. Evaseck, M. J. Field, S. Fischer, J. Gao, H. Gao, S. He, D. Joseph- Mac Carthy, L. Kuchnir, K. Kuczera, F. T. K. Lau, C. Mattos, S. Michmick, T. Nego, D. T. Nguyen, B. Prodhom, W. E. Reiher III, B. Roux, M. Schlemkrich, J. C. Smith, R. Stote, J. Straub, M. Watanabe, J. Wiorkiewicz-Kuczera, D. Yin, M. Karplus, (1998)″All-atom empirical potential for molecular modeling and dynamics studies of proteins″, J. Phys. Chem. B, Vol,102, pp, 3586- 3616.

29- B. R. Brooks et al., (1983) ″CHARMM: A Program for Macromolecular Energy, Minimization, and Dynamics Calculations″, J. Comput. Chem., Vol, 4, pp, 187.

30- A. D. Mackerell et al., (1995) ″An all-atom empirical energy function for the simulation of nucleic acids″, J. Am. Chem. Soc., Vol, 117, pp, 11946.

31- A. D. Mackerell et al., (1998) ″All-Atom Empirical Potential for Molecular Modeling and Dynamics Studies of Proteins″. J. Phys. Chem. B, Vol, 102, pp, 3586.

32- F. A. Momany, R. Rone, (1992) ″Validation of the General Purpose QUANTAa3.2/ CHARMm ® Force Field″, J. Comput. Chem., Vol, 13, pp, 888.

33- T. A. Halgren, (1996) ″Merck Molecular Force Field I″, J. Comput. Chem., Vol, 17, pp, 490, 520, 553, 616.

34- T. A. Halgren, R. B. Nachbar, (1996) ″Merck Molecular Force Field. IV.Conformational Energies and Geometries for MMFF94″, J. Comput. Chem., Vol, 17, pp, 587-615.

35- A. R. Leach, (2001) ″Molecular modeling- Principles and applications″, Person, Prentice Hall, Second Edition, England, Chap.4.

36- B. J. Alder, T. E. Wainwright, (1957) ″Phase Transition for a Hard Sphere System″, J. Chem. Phys., Vol, 27, pp, 1208.

37- A. Rahman, (1964)″Correlations in motion of atoms in liquid argon″, Phys. Rev., 136, A405.

38- A. Rahman, F. H. Stillinger, (1971) ″Molecular Dynamics Study of Liquid Water″, J. Chem. Phys., Vol, 5, pp, 3336.

39- J. A. Mc Cammon, S. C. Harvey, (1987) ″Dynamics of Proteins and Nucleic Acids″, Cambridge Univ. Press.

40- A. R. Leach, (2007)″Introduction to Chemoinformatics″ ; Springer.

41- K. Roy, S. Kar, R. N. Das, (2015) ″A premier on QSAR / QSPR Modeling- Fundamental Concepts″, Springer Breifs in Molecular Science- DOI 10. 1007 / 978- 3- 319- 17 281- 1.

42- R. Todeschini, V. Consonni, (2000) ″Handbook of molecular descriptors″. Wiley VCH, Weinhein

43- D. J. Livingstone. (2000) ″The Characterization of chemical structures using molecular properties. A survey. I″. Chem. Inf. Comput. Sci., Vol, 40, pp, 195- 209.

44- A. F. A. Cros, (1863) ″Action de l'alcool amylique sur l'organisme″, Thèse de doctorat, faculté de médecine.

45- A. C. Crum-Brown and T. R. Fraser, (1868) ″On the Connection Between Chemical Constitution and Physiological Action, Part I: On the Physiological Action of the Salts of the Ammonium Bases, Derived from Strychnia, Brucia, Thebia, Codeia, Morphia, Nicotia, Earth an, Trans, Roy, Soc., Vol, 25, pp, 151-203.

46- M. C. Richet, (1893) ″Noté sur le rapport entre la toxicité et les propriétés physiques des corps″, Comptes rendus des séances de la Société de biologie et de ses filiales, Paris, Vol, 45, pp, 775–6.

47- H. Meyer, (1899) ″Zur Theorie der Alkoholnarkose. Erste Mittheilung. Welche Eigenschaft der Anästhetica bedingt ihre narkotische Wirkung″, Archiv für experimentelle Pathologie und Pharmakologie, Vol, 42, pp, 109–118.

48- E. Overton , (1901)″Studien über die Narkose zugleich ein Beitrag zur allgemeinen″ Pharmakologie, Ed. G. Fischer, Jena.

49- R. L. Lipnick, (1986)″Charles Ernest Overton: narcosis studies and a contribution to general pharmacology″, Trends in Pharmacological Sciences, Vol, 7, pp, 161–164.

50- H. Fühner and E. Neubauer, 1907″ămolyse durch Substanzen homologen Reihen″, Archiv für experimentelle Pathologie und Pharmakologie, Vol, 56, pp,333–345.

51- O. R. Hansen, 1962″Hammett Series with Biological Activity″, Acta Chemica Scandinavica, Vol, 16, pp, 1593–1600.

52- C. Hansch and T. Fujita, (1964) ″ p-σ-π Analysis. A Method for the Correlation of Biological Activityand Chemical Structure″, Journal of the American Chemical Society, Vol, 86(8), pp, 1616–1626.

53- S. M. Free and J. W. Wilson, (1964) ″A Mathematical Contribution to Structure-Activity Studies″, Journal of Medicinal Chemistry, Vol, 7(4), pp, 395–399.

54- C. Hansch and E, J, Lien, (1971) ″Structure-activity relationships in antifungial agents. A survey″, Journal of Medicinal Chemistry, Vol,14(8), pp, 653–670.

55- S. Y. Tham and S. Agatonovic-Kustrin, (2002) ″Application of the artificial neural network in quantitative structure-gradient elution retention relationship of phenylthiocarbamyl amino acids derivatives″, Journal of Pharmaceutical and Biomedical Analysis, Vol, 28(3), pp, 581-590.

56- R. D. Cramer, D. E. Patterson, and J. D. Bunce, (1988) ″Comparative molecular field analysis″, J. Am. Chem. Soc., Vol, 110(18), pp, 5959- 5967.

57- G. Klebe, U. Abraham, and T. Mietzner, (1994) ″Molecular similarity indices in a comparative analysis (CoMSIA) of drug molecules to correlate and predict their biological activity″, Journal of Medicinal Chemistry, Vol, 37(24), pp, 4130-4146.

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

59- G. E. P. Box and D. R. Cox, (1964) ″An analysis of distributions″, Journal of the royal statistical society, Series B, Vol, 26(2), pp, 211-243.

60- P. Armitage, G. Berry, (1994) ″Statistical Methods in Medical Research″, 3rd ed., Blackwell.

61- R.C. Reid, J.M. Prausnitz, B.E. Poling, (1987) ″The Properties of Gases & liquids″, Fourth Edition, Mc Graw-Hill Book Company, New York.

62- Hyperchem ^{TM} Release 6,03 for windows, Molecular Modeling System (2000).

63- I. N. Levine, (2000)"Quantum Chemistry", 5thed, New Jersey: Prentice Hall.

64- R. Todeschini, V. Consonni, M. Pavan, (2005) DRAGON, Software for the Calculation of Molecular Descriptors, Release 5.3 for windows, Milano.

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

66- B. Kowalski, R. Gerlach, H. Wold, (1982) ″Systems under indirect observation″, (K, Jörenskoget H, Wold, eds,), North Holland, Amsterdam, pp, 191-206..

67- L. Erikson, E. Johannson, N. Kettaneh- Wold, (2001) ″Multi and megavariate data analysis- principles and applications″, Umetrics Academy, Umeå.

68- S. Wold, (1984) ″Chemometrics: Mathematics and Statistics in Chemistry″, Reidel, Dordrecht, The Netherlands.

69- S. Wold, A. Ruhe, H. Wold, W. Dunn, (1984) ″The Collinearity Problem in Linear Regression. The Partial Least Squares (PLS) Approach to Generalized Inverses",SIAMJ, Sci, Stat, Comput, Vol, 5, pp, 735- 743.

70- P. Gelada, B. R. Kowalski, (1986) ″Partial least- squares regression: tutorial″, Anal, Chim, Acta, Vol, 185, pp, 1- 17.

71- A. Höskuldsson, (1988) ″PLS regression methods″, J, Chemometrics, Vol, 2, pp, 211- 228.

72- J. A. Burns, G. M. Whiteside, (1993) ″Feed- forward neural networks in chemistry: mathematical systems for classification and pattern recognition″, Chem, Rev, Vol, 93(8), pp, 2583- 2601.

73- L. S. Anker, P. C. Jurs, (1992) ″Prediction of carbon-13 nuclear magnetic resonance chemical shifts by artificial neural networks″, Anal, Chem, Vol, 64, pp, 1157- 1164.

74- T. Aoyama, Y. Suzuki, H. Ichikawa, (1990) ″Neural networks applied to quantitative structure-activity relationship analysis″, J, Med, Chem, Vol, 33, pp, 2583- 2590.

75- T. A. Andrea, (1991) ″Applications of neural networks in quantitative structure-activity relationships of dihydrofolate reductase inhibitors″, J, Med, Chem, Vol, 34, pp, 2824 – 2836.

76- P. C. Jurs, (1996) ″Computer Software Applications in Chemistry″, Second Edition, J, Wiely.

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

78- Le jeune M, (2004) ″Statistiques : la théorie et ses applications″, Springer-Verlag, Paris.

79- G. Fayet, (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.

80- Mc Culloch-Pitts, (1943) ″a logical Calculus at the ideas imminent in Nervous Activity″. Bulletin at math. Biophysics.Vol. 5, pp, 115-133.

81- M. Minsky, S. Papert, (1969) ″Perceptrons″. Massachusetts: MIT press.

82- D. E. Rumelbart, J. L. Mcclelland et al.,(1988) ″Parallel Distributed processing″. Massachusetts: MIT press,Vol, 1, pp, 547.

83- J. J. Hopfield. (1982) ″Neural Networks and physical systems with emergent collective computational abilities″. Proceedings of the National Academy of sciences. USA.

84- T. Kohonen. (1988) ″Self-organization and associative memory″. Bulletin: Springer-Verlag. 984.

85- R. Hecht-Nielson. (1991) ″Neurocomputing″. Addison-Wesly Publishing Company, New York.

86- F. Fogelmen-Soulié. (1988) ″Méthodes connexionnistes pour l'apprentissage″. Actes des journées Nationales sur l'intelligence Artificielle. Paris: Teknea. pp, 275-293.

87- K. Hornik, (1991) ″Approximation capabilities of multilayer feedforward networks″, Neural Networks, Vol, 4, pp, 251-257.

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

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

90- P. Gramatica. (2007) ″Principles of QSAR Models Validation: Internal and External″. Qsar & Combinatorial Science, Vol, 26, pp, 694–701.

91- L. Eriksson, J. Jaworska, A. P. Worth, M.T.D.Cronin, R. M. Mc Dowell, P. Gramatica, (2003) ″Methods for reliability and uncertainty assessment and for applicability evaluations of

classification-and regression-based QSARs″. Environmental health perspectives, Vol, 111(10), pp, 1361-1375.

92- A. Golbraikh. and A.Tropsha. (2002) ″Beware of Q(2) ″, Journal of Molecular Graphics & Modelling, Vol, 20, pp, 269–276.

93- J. C. Dearden, M. T. D. Cronin and K. L. E. Kaiser. (2009) ″How Not to Develop a Quantitative Structure–activity or Structure–property Relationship (QSAR/QSPR) ″. SAR and QSAR in Environmental Research, Vol, 20, pp, 241–266.

94- S. Lozano, M.-P. Halm-Lemeille, A. Lepailleur, S. Rault, R. Bureau. (2010) ″Consensus QSAR Related to Global or MOA Models: Application to Acute Toxicity for Fish″. Molecular Informatics, Vol, 29, pp, 803–813.

95- P. P. Roy, S. Kovarich. and P. Gramatica. (2011) ″QSAR Model Reproducibility and Applicability: A Case Study of Rate Constants of Hydroxyl Radical Reaction Models Applied to Polybrominated Diphenyl Ethers and (benzo-)triazoles″. Journal of Computational Chemistry Vol , 32, pp, 2386-2396.

96- N. Chirico. and P. Gramatica, P. Real. (2011) ″External Predictivity of QSAR Models: How To Evaluate It? Comparison of Different Validation Criteria and Proposal of Using the Concordance Correlation Coefficient″. Journal of Chemical Information and Modeling, Vol, 51, pp, 232.

97- A. Golbraikh, M. Shen, Z. Y. Xiao, Y. D. Xiao, K. H. Lee, A. Tropsha, (2003) ″Rational Selection of Training and Test Sets for the Development of Validated QSAR Models″. Journal of Computer-Aided Molecular Design, Vol, 17, pp, 241–253.

98- A. Tropsha, P. Gramatica. and V. K. Gombar. (2003) ″The Importance of Being Earnest: Validation Is the Absolute Essential for Successful Application and Interpretation of QSPR Models″. Qsar & Combinatorial Science, Vol, 22, pp, 69–77.

99- G. Schüürmann, R. –U. Ebert, J. Chen, B. Wang, R. Kühne. (2008) ″External Validation and Prediction Employing the Predictive Squared Correlation Coefficient - Test Set Activity Mean Vs Training Set Activity Mean″. Journal of Chemical Information and Modeling, Vol, 48, pp, 2140-2145.

100- V. Consonni, D. Ballabio. and R. Todeschini. (2009) ″Comments on the Definition of the Q2 Parameter for QSAR Validation″. Journal of Chemical Information and Modeling, Vol, 49, pp, 1669–1678.

101- L. I. –K. Lin. (1989) ″A Concordance Correlation Coefficient to Evaluate Reproducibility″. Biometrics, Vol, 45, pp, 255–268.

102- L. I.-K. Lin. (1992) ″Assay Validation Using the Concordance Correlation Coefficient″. Biometrics, Vol, 48, pp, 599–604.

103- F. Gharagheizi, S. A. Mirkhani, P. Ilani-Kashkouli, et al, (2013), Fluid Phase Equil., Vol, 354, pp, 250.

104- A.R. Katritzky, M. Kuanar, S. Slavov, C.D. Hall, M. Karelson, I. Kahn, D.A. Dobchev, (2010), Chemical Reviews., Vol, 110, pp, 5714.

105- T. Le, V.C. Epa, F.R. Burden, D.A. Winkler, (2012), Chemical Reviews., Vol, 112, pp, 2889.

106- W.J. Lyman, W.F. Reechl, D.H. Rosenblatt, (1990), Handbook of Chemical Property Estimation Methods, American Chemical Society, Washington, DC.

ANNEXES

Tableau I: Nom, classe et définition des descripteurs sélectionnés

N	Descripteur	Classe	Définition
1	**R2e** Autocorrélation R de décalage 2 / pondérée par l'électronégativité de Sanderson	Descripteurs GETAWAY	Les indices HATS, H, R et R maximal basés sur l'autocorrélation spatiale, encodent des informations sur les fragments structuraux et semblent donc particulièrement appropriés pour décrire les différences dans les séries congénères de molécules. A la différence des autocorrélations de Moreau-Broto, les GETWAY sont des descripteurs géométriques encodant des informations sur la position effective de substituants et de fragments dans l'espace moléculaire. De plus, ils sont indépendants de l'alignement de la molécule et, dans une certaine mesure, tiennent également compte des informations sur la taille et la forme moléculaires ainsi que sur les propriétés atomiques spécifiques.
2	**R4m** Autocorrélation R de décalage 4 pondérée par les masses atomiques		
3	**HATS5u** Autocorrélation à effet de levier pondéré de décalage 5 / non pondérée		
4	**H6m** Autocorrélation H de décalage 6 / pondérée par les masses atomiques.		
5	**R1p+** Autocorrélation maximale R de décalage 1 / pondérée par les polarisabilités atomiques.		
6	**H5e** AutocorrélationH de retard 5 / pondérée par l'électronégativité atomique de Sanderson.		

N	Descripteur	Classe	Définition
7	**piPC01** Dénombrement de trajets moléculaires multiples d'ordre 01 [somme des ordres de liaisons conventionnels [SCBO]	Dénombrement de marches et de trajets	Les dénombrements de trajets moléculaires multiples (piPCk) sont des dénombrements de trajets de graphes pondérés par une longueur donnée k, ou chaque trajet est pondéré par le produit des ordres de liaisons conventionnels des arêtes impliquées. Pour les molécules saturées, les dénombrements de trajets moléculaires multiples coïncident avec le(s) dénombrement(s) de chemin moléculaire.
8	**SIC2** Contenu d'information structurelle (symétrie de voisinage d'ordre 2)	Indices d'information	Les indices de symétrie de voisinage sont calculés en partitionnant les sommets de graphe en classes d'équivalence ; l'équivalence topologique de deux sommets réside dans ce que les voisinages correspondants du $k^{ième}$ ordre sont le mêmes. Le voisinage des sommets peut être considéré comme une sphère ouverte comprenant tous les sommets du graphe, de sorte que leur distance par rapport au sommet considéré est inférieure à k. Le logiciel DRAGON calcule les indices de symétrie de voisinage de l'ordre 0 jusqu'à l'ordre 5.
9	**CIC0** Contenu d'information complémentaire (Symétrie de voisinage d'ordre 0)		
10	**BIC2** Contenu d'information de liaison (Symétrie de voisinage d'ordre 2)		

N°	Descripteur	Classe	Définition
11	**Vev1** Somme des coefficients du vecteur propre d'après la matrice distance pondérée de van der walls.	Indices basés sur les valeurs propres	v étant une matrice carrée représentant un graphe moléculaire dépourvu des atomes d'hydrogène, ces descripteurs moléculaires sont obtenus en sommant les coefficients du vecteur propre associé à la plus basse (la plus négative) valeur propre de la matrice v).
12	**MAXDN** Variation électrotopologique négative maximale.	Indices topologiques	la variation électrotopologique négative maximale (MAXDN) est calculée comme valeur négative de DI_i dans la molécule.
13	**GGI1** Indice de charge topologique d'ordre 1	Indices de charge topologique	L'indice de charge topologique GGIk est défini comme la demi-somme de tous les termes de charge CIij (en valeurs absolues correspondant à une paire de sommets séparés par une distance topologique égale à k. le nombre maximal de GGIk dans une molécule est égal à la distance topologique maximale dans le graphe moléculaire correspondant ; cependant, pour obtenir des descripteurs de longueur uniforme pour un ensemble de molécules, le logiciel DRAGON calcule les indices de charge topologique d'ordres 1 à 10 pour toutes les molécules.
14	**GNar** Indice topologique géométrique de Narumi	Descripteurs topologiques	L'ancien indice topologique géométrique de Narumi est obtenu en divisant le nombre d'atomes, autres que l'hydrogène, par la somme des degrés des sommets, qui est remplacée dans la définition actuelle de ce descripteur par la moyenne géométrique des degrés des sommets.
15	**ATS8v** Autocorrélation de Broto-Morceau d'une structure topologique de retard pondérée par les volumes atomiques de Van Der Waals	Indices d'autocorrélation 2 D	Autocorrélation de Broto-Morceau ATS8kw, W étant la propriété atomique utilisé pour pondérer le graphe moléculaire et k le retard est évaluée en considérant séparément toutes les contributions de chaque longueur (topologique) de chemin (le retard) dans le graphe moléculaire, collectées dans la matrice distance topologique. En d'autres termes, l'autocorrélation spatiale totale à logk, ATS8kw est obtenu en additionnant tous les produits Wi*Wj de tous les paires d'atomes i et j, pour lesquels la distance que est égale au décalage

Tableau II : Valeurs des descripteurs sélectionnés

N	piPC01	R2e	R4m	SIC2	GGI1	MAXDN	VEv1	HATS5u	H6m	R1p+	GNar	CIC0	BIC2	ATS8v	H5e
1	0	2,199	0	0	0	0	1	0	0	0	0	1,6	0	0	0
2	0,693	1,861	0	0,27	0	0	1,414	0	0	0,059	1	2,189	0	0	0
3	1,099	1,862	0,006	0,487	0,5	0,25	1,716	0	0	0,061	1,26	2,614	0,507	0	0
4	1,386	1,832	0,026	0,484	0,5	0,181	1,974	0,658	0	0,056	1,414	2,944	0,498	0	0
5	1,609	1,824	0,051	0,532	1,5	0,449	2,202	0,88	0	0,057	1,431	3,213	0,543	0	0,169
6	1,792	1,739	0,074	0,424	3	0,708	2,411	0,914	0	0,051	1,414	3,441	0,432	0	0,33
7	1,792	1,86	0,09	0,363	2	0,481	2,412	1,295	0	0,052	1,442	3,441	0,37	0	0,443
8	1,792	1,838	0,086	0,554	1,5	0,435	2,412	0,762	0,001	0,054	1,513	3,441	0,564	0	0,239
9	1,792	1,77	0,092	0,531	1,5	0,398	2,409	1,226	0,001	0,054	1,513	3,441	0,54	0	0,367
10	1,792	1,83	0,06	0,474	0,5	0,145	2,412	0,712	0,003	0,047	1,587	3,441	0,482	0	0,069
11	1,946	1,81	0,119	0,464	3	0,7	2,604	0,654	0,001	0,048	1,486	3,637	0,471	0	0,428
12	1,946	1,877	0,127	0,452	2,5	0,458	2,607	0,671	0,004	0,048	1,511	3,637	0,458	0	0,357
13	1,946	1,786	0,116	0,372	3,5	0,75	2,607	1,376	0	0,047	1,426	3,637	0,377	0	0,788
14	1,946	1,868	0,08	0,557	1,5	0,431	2,605	0,594	0,004	0,052	1,575	3,637	0,566	0	0,284
15	1,946	1,823	0,109	0,557	1,5	0,319	2,616	1,055	0	0,048	1,575	3,637	0,566	0	0,505
16	1,946	1,843	0,151	0,485	1,5	0,347	2,598	1,241	0,003	0,052	1,575	3,637	0,492	0	0,275
17	1,946	1,842	0,06	0,447	0,5	0,139	2,604	0,642	0,004	0,046	1,641	3,637	0,454	0	0,093
18	2,079	1,863	0,17	0,449	4	0,728	2,788	0,589	0,005	0,044	1,488	3,81	0,454	0	0,395
19	2,079	1,776	0,15	0,243	4,5	0,813	2,789	1,564	0	0,037	1,414	3,81	0,246	0	1,19
20	2,079	1,864	0,101	0,482	3	0,698	2,784	0,464	0,006	0,048	1,542	3,81	0,488	0	0,405
21	2,079	1,891	0,089	0,468	2,5	0,447	2,788	0,525	0,011	0,043	1,565	3,81	0,474	0	0,445
22	2,079	1,87	0,139	0,527	2,5	0,454	2,784	0,778	0,005	0,047	1,565	3,81	0,534	0	0,528
23	2,079	1,818	0,17	0,465	3,5	0,741	2,784	1,269	0,001	0,045	1,488	3,81	0,471	0	1,039
24	2,079	1,793	0,152	0,488	3	0,658	2,779	0,997	0,002	0,044	1,542	3,81	0,494	0	0,929
25	2,079	1,786	0,17	0,471	3,5	0,708	2,783	1,536	0	0,042	1,488	3,81	0,477	0	1,323
26	2,079	1,826	0,19	0,443	3	0,625	2,777	1,479	0,001	0,051	1,542	3,81	0,449	0	0,952
27	2,079	1,864	0,152	0,515	1,5	0,333	2,776	0,949	0,004	0,051	1,622	3,81	0,521	0	0,325
28	2,079	1,838	0,121	0,521	1,5	0,37	2,779	0,791	0,004	0,049	1,622	3,81	0,527	0	0,568

Tableau II : (suite)

N	piPC01	R2e	R4m	SIC2	GGI1	MAXDN	VEv1	HATS5u	H6m	R1p+	GNar	CIC0	BIC2	ATS8v	H5e
29	2,197	1,884	0,214	0,329	5,5	0,766	2,96	0,547	0,009	0,039	1,47	3,964	0,332	0	0,469
30	2,197	1,887	0,087	0,476	3	0,699	2,954	0,375	0,01	0,046	1,587	3,964	0,481	0	0,522
31	2,197	1,852	0,219	0,386	4	0,77	2,956	1,177	0,004	0,042	1,489	3,964	0,39	0	1,17
32	2,197	1,798	0,209	0,371	4,5	0,804	2,954	1,564	0,001	0,043	1,47	3,964	0,375	0	1,66
33	2,197	1,899	0,072	0,514	1,5	0,429	2,954	0,459	0,008	0,048	1,661	3,964	0,519	0	0,523
34	2,197	1,867	0,063	0,401	0,5	0,133	2,951	0,536	0,007	0,041	1,714	3,964	0,405	0,693	0,429
35	2,303	1,917	0,107	0,359	5,5	0,743	3,122	0,337	0,026	0,035	1,516	4,104	0,362	0	0,651
36	2,303	1,883	0,066	0,381	0,5	0,131	3,11	0,502	0,008	0,04	1,741	4,104	0,385	1,099	0,606
37	2,398	1,888	0,069	0,364	0,5	0,13	3,261	0,471	0,01	0,038	1,763	4,231	0,367	1,386	0,752
38	2,485	1,9	0,072	0,348	0,5	0,129	3,406	0,448	0,012	0,037	1,782	4,348	0,351	1,609	0,902
39	2,565	1,904	0,075	0,335	0,5	0,128	3,544	0,425	0,022	0,035	1,798	4,456	0,337	1,792	1,016
40	2,639	1,913	0,079	0,322	0,5	0,128	3,678	0,407	0,036	0,034	1,811	4,557	0,324	1,946	1,137
41	2,773	1,924	0,084	0,301	0,5	0,127	3,932	0,375	0,066	0,031	1,834	4,739	0,303	2,197	1,323
42	2,89	1,932	0,09	0,284	0,5	0,127	4,17	0,348	0,096	0,029	1,852	4,901	0,285	2,398	1,479
43	2,197	1,881	0,166	0,493	4	0,727	2,955	0,617	0,009	0,046	1,537	3,964	0,499	0	0,677
44	2,197	1,908	0,265	0,411	3	0,583	2,943	1,441	0,004	0,049	1,587	3,964	0,415	0	0,765
45	2,708	1,916	0,082	0,311	0,5	0,128	3,807	0,389	0,051	0,032	1,823	4,651	0,313	2,079	1,228
46	2,996	1,935	0,093	0,27	0,5	0,126	4,395	0,325	0,125	0,026	1,866	5,047	0,271	2,565	1,607
47	1,609	1,577	0,026	0,5	0,5	0	1,964	1,195	0	0,067	1,414	2,667	0,5	0	0,144
48	1,609	1,712	0,03	0,5	0,5	0	1,964	1,062	0	0,068	1,414	2,667	0,5	0	0,457
49	1,792	1,56	0,05	0,816	0,5	0,347	2,209	0,644	0	0,063	1,516	2,989	0,816	0	0
50	1,792	1,584	0,049	0,769	1,5	0,412	2,2	1,011	0	0,062	1,431	2,989	0,769	0	0,395
51	1,792	1,609	0,047	0,769	0,5	0,337	2,196	0,918	0,002	0,063	1,516	2,989	0,769	0	0,042
52	1,792	1,729	0,052	0,769	0,5	0,337	2,196	0,795	0,001	0,06	1,516	2,989	0,769	0	0,397
53	1,792	1,586	0,052	0,479	1,5	0,287	2,191	1,36	0	0,063	1,431	2,989	0,479	0	0,446
54	1,946	1,603	0,062	0,777	0,5	0,323	2,418	0,721	0,001	0,062	1,587	3,252	0,777	0	0,11
55	2,079	1,641	0,064	0,737	0,5	0,312	2,61	0,701	0,003	0,061	1,641	3,474	0,737	0	0,162
56	1,386	1,454	0,005	0,763	0,5	0,25	1,715	0	0	0,076	1,26	2,252	0,763	0	0

Tableau II : (suite)

N	piPC01	R2e	R4m	SIC2	GGI1	MAXDN	VEv1	HATS5u	H6m	R1p+	GNar	CIC0	BIC2	ATS8v	H5e
57	1,609	1,348	0,014	0,546	1,5	0,5	1,969	0	0	0,068	1,316	2,667	0,546	0	0
58	1,792	1,533	0,049	0,667	1,5	0,685	2,206	0,9	0	0,072	1,431	2,989	0,667	0	0,007
59	1,946	1,685	0,054	0,793	0,5	0,267	2,408	0,661	0,003	0,056	1,587	3,252	0,793	0	0,151
60	1,946	1,714	0,064	0,793	0,5	0,267	2,408	0,635	0	0,06	1,587	3,252	0,793	0	0,405
61	1,946	1,694	0,051	0,58	0,5	0,326	2,403	0,637	0,004	0,052	1,587	3,252	0,58	0	0,024
62	1,946	1,768	0,059	0,58	0,5	0,326	2,403	0,613	0,003	0,056	1,587	3,252	0,58	0	0,34
63	1,946	1,679	0,079	0,677	1,5	0,331	2,401	0,858	0,002	0,061	1,513	3,252	0,677	0	0,433
64	1,946	1,646	0,087	0,703	1,5	0,309	2,4	1,429	0,001	0,051	1,513	3,252	0,703	0	0,826
65	1,946	1,778	0,1	0,703	1,5	0,309	2,4	1,368	0,001	0,058	1,513	3,252	0,703	0	0,299
66	1,946	1,681	0,084	0,677	1,5	0,616	2,407	0,739	0,002	0,058	1,513	3,252	0,677	0	0,02
67	1,946	1,71	0,094	0,677	1,5	0,616	2,407	0,689	0,002	0,065	1,513	3,252	0,677	0	0,196
68	1,946	1,625	0,083	0,294	2	0,241	2,4	1,74	0	0,051	1,442	3,252	0,294	0	0,745
69	1,946	1,639	0,089	0,677	2	0,676	2,411	1,385	0	0,068	1,442	3,252	0,677	0	0,327
70	1,946	1,486	0,071	0,53	3	0,944	2,417	0,862	0	0,069	1,414	3,252	0,53	0	0,149
71	2,079	1,597	0,113	0,566	3,5	0,944	2,607	1,423	0	0,052	1,426	3,474	0,566	0	0,6
72	2,303	1,702	0,067	0,636	0,5	0,302	2,956	0,592	0,007	0,056	1,714	3,837	0,636	0,693	0,423
73	2,485	1,749	0,073	0,557	0,5	0,298	3,266	0,512	0,01	0,051	1,763	4,126	0,557	1,386	0,738
74	2,565	1,769	0,076	0,526	0,5	0,297	3,41	0,481	0,013	0,049	1,782	4,252	0,526	1,609	0,884
75	2,639	1,785	0,079	0,498	0,5	0,296	3,549	0,455	0,021	0,045	1,798	4,367	0,498	1,792	1,003
76	2,833	1,824	0,088	0,433	0,5	0,294	3,935	0,395	0,067	0,04	1,834	4,667	0,433	2,197	1,316
77	2,944	1,844	0,093	0,4	0,5	0,294	4,173	0,364	0,099	0,036	1,852	4,837	0,4	2,398	1,476
78	1,609	1,847	0,005	0,256	0	0	2	0	0	0,057	2	2,667	0,256	0	0
79	1,792	1,989	0,01	0,235	0	0	2,236	0	0	0,049	2	2,989	0,235	0	0
80	1,946	1,964	0,039	0,564	1	0,287	2,431	0,558	0	0,053	1,906	3,252	0,564	0	0,06
81	1,946	2,081	0,045	0,22	0	0	2,449	0,554	0	0,046	2	3,252	0,22	0	0
82	2,079	2,12	0,091	0,515	1	0,297	2,629	0,776	0	0,048	1,919	3,474	0,515	0	0,275
83	2,079	2,092	0,099	0,515	1	0,236	2,614	0,798	0,001	0,055	1,919	3,474	0,515	0	0,246
84	2,197	2,127	0,118	0,514	2	0,314	2,797	0,865	0,001	0,041	1,861	3,667	0,514	0	0,622

Tableau II : (suite)

N	piPC01	R2e	R4m	SIC2	GGI1	MAXDN	VEv1	HATS5u	H6m	R1p+	GNar	CIC0	BIC2	ATS8v	H5e
85	2,197	2,205	0,155	0,2	0	0	2,828	0,933	0	0,044	2	3,667	0,2	0	0,404
86	1,386	1,706	0	0,29	0	0	1,732	0	0	0,072	2	2,252	0,29	0	0
87	2,079	1,898	0,071	0,507	2,5	0,556	2,62	0,812	0	0,053	1,811	3,474	0,507	0	0,17
88	2,079	1,917	0,079	0,577	1,5	0,319	2,616	1,166	0	0,051	1,842	3,474	0,577	0	0,171
89	2,079	2,022	0,081	0,577	1,5	0,384	2,601	0,964	0,002	0,051	1,842	3,474	0,577	0	0,362
90	2,197	2,068	0,138	0,55	1,5	0,33	2,802	1,142	0	0,045	1,861	3,667	0,55	0	0,484
91	2,197	2,055	0,134	0,568	2	0,321	2,802	0,877	0	0,048	1,834	3,667	0,478	0	0,311
92	2,197	2,119	0,131	0,475	1	0,247	2,797	0,913	0,001	0,049	1,929	3,667	0,475	0	0,446
93	2,197	1,909	0,12	0,582	3	0,597	2,795	1,386	0	0,048	1,769	3,667	0,582	0	0,462
94	2,197	1,942	0,114	0,582	3,5	0,584	2,792	0,789	0,001	0,046	1,769	3,667	0,582	0	0,189
95	2,197	2,102	0,11	0,504	1	0,222	2,789	0,622	0,003	0,051	1,929	3,667	0,504	0	0,22
96	2,197	1,924	0,132	0,568	1,5	0,391	2,795	0,927	0,001	0,048	1,861	3,667	0,568	0	0,727
97	2,303	2,141	0,139	0,474	1	0,233	2,959	0,72	0,002	0,048	1,937	3,837	0,474	0	0,476
98	2,303	2,018	0,179	0,509	3	0,344	2,967	0,845	0,001	0,046	1,817	3,837	0,509	0	0,346
99	2,398	2,155	0,127	0,446	1	0,228	3,116	0,586	0,005	0,048	1,943	3,989	0,446	0	0,59
100	2,398	2,107	0,16	0,529	2	0,414	3,122	0,602	0,007	0,045	1,888	3,989	0,529	0	0,631
101	2,398	2,108	0,195	0,521	1,5	0,347	3,119	0,947	0,011	0,045	1,888	3,989	0,521	0	0,687
102	2,398	2,063	0,208	0,502	3	0,68	3,127	0,949	0,005	0,042	1,813	3,989	0,502	0	0,776
103	2,485	2,114	0,085	0,422	1	0,216	3,268	0,404	0,006	0,052	1,948	4,126	0,422	1,099	0,52
104	2,565	2,126	0,086	0,401	1	0,216	3,414	0,375	0,008	0,05	1,953	4,252	0,401	1,609	0,717
105	2,773	2,119	0,095	0,352	1	0,218	3,816	0,324	0,021	0,048	1,962	4,574	0,352	2,079	1,039
106	2,833	2,196	0,127	0,338	1	0,228	3,94	0,383	0,031	0,046	1,964	4,667	0,338	2,197	1,23
107	2,944	2,132	0,106	0,316	1	0,219	4,179	0,3	0,063	0,046	1,968	4,837	0,316	2,398	1,334
108	2,996	2,123	0,108	0,306	1	0,219	4,294	0,286	0,08	0,046	1,97	4,915	0,306	2,485	1,386
109	3,045	2,125	0,11	0,298	1	0,22	4,405	0,284	0,094	0,045	1,971	4,989	0,298	2,565	1,481
110	1,792	0,735	0,017	0,579	0,5	0,361	1,985	0,331	0	0,062	1,414	2,351	0,556	0	0
111	1,792	1,356	0,024	0,797	0,5	0,236	1,967	1,124	0	0,087	1,414	2,351	0,765	0	0,145
112	1,946	1,258	0,046	0,667	0,5	0,583	2,217	0,348	0	0,059	1,516	2,739	0,649	0	0,01

Tableau II : (suite)

N	piPC01	R2e	R4m	SIC2	GGI1	MAXDN	VEv1	HATS5u	H6m	R1p+	GNar	CIC0	BIC2	ATS8v	H5e
113	1,946	1,308	0,049	0,594	1,5	0,523	2,194	1,453	0	0,086	1,431	2,739	0,578	0	0,135
114	1,946	1,356	0,045	0,818	0,5	0,25	2,202	0,748	0	0,083	1,516	2,739	0,795	0	0,007
115	1,946	1,519	0,051	0,818	0,5	0,455	2,202	0,635	0,001	0,083	1,516	2,739	0,795	0	0,103
116	1,609	0,472	0,003	0,491	0,5	0,25	1,716	0	0	0,072	1,26	1,822	0,46	0	0
117	2,485	1,605	0,071	0,727	0,5	0,299	3,115	0,548	0,008	0,054	1,741	3,989	0,594	1,099	0,59
118	2,079	1,728	0,037	0,769	1	0,175	2,423	0,83	0	0,057	1,906	3,046	0,753	0	0,059
119	2,079	1,875	0,04	0,625	0	0,181	2,448	0,515	0	0,063	2	3,046	0,612	0	0
120	2,197	1,861	0,092	0,724	1	0,222	2,605	0,943	0,001	0,058	1,919	3,298	0,712	0	0,247
121	1,946	1,749	0,007	0,667	0	0,181	2,234	0	0	0,07	2	2,739	0,649	0	0
122	1,386	0,133	0	0,5	0	0	1,414	0	0	0,057	1	1	0,431	0	0
123	1,609	1,34	0,003	0,758	0,5	0,347	1,714	0	0	0,099	1,26	1,822	0,709	0	0
124	1,792	1,551	0,023	0,413	0,5	0,181	1,959	1,956	0	0,071	1,414	2,351	0,396	0	0,533
125*	1,946	0,569	0,015	0,917	0,5	0,597	1,987	0,288	0	0,097	1,414	2	0,828	0	0
126*	1,386	1,805	0,015	0,376	1,5	0,5	1,972	0	0	0,059	1,316	2,944	0,386	0	0
127*	1,609	1,801	0,052	0,496	0,5	0,156	2,204	0,806	0,001	0,054	1,516	3,213	0,507	0	0
128*	1,946	1,697	0,122	0,452	3	0,667	2,601	1,291	0	0,045	1,486	3,637	0,458	0	0,877
129*	1,946	1,802	0,126	0,502	2	0,468	2,603	1,282	0,001	0,049	1,511	3,637	0,51	0	0,813
130*	2,079	1,854	0,184	0,37	2,5	0,491	2,785	1,246	0,002	0,048	1,51	3,81	0,375	0	1,076
131*	2,079	1,84	0,13	0,527	2	0,463	2,782	1,007	0,004	0,045	1,565	3,81	0,534	0	0,828
132*	2,079	1,834	0,186	0,517	2	0,454	2,779	1,221	0,005	0,046	1,565	3,81	0,523	0	0,576
133*	2,079	1,882	0,074	0,537	1,5	0,429	2,785	0,511	0,007	0,05	1,622	3,81	0,544	0	0,389
134*	2,079	1,826	0,155	0,501	2	0,37	2,779	0,791	0,004	0,049	1,565	3,81	0,507	0	1,021
135*	2,079	1,838	0,121	0,521	1,5	0,417	2,78	1,273	0,003	0,044	1,622	3,81	0,527	0	0,568
136*	2,079	1,862	0,062	0,423	0,5	0,135	2,782	0,582	0,006	0,043	1,682	3,81	0,428	0	0,282
137*	2,197	1,902	0,102	0,479	4	0,718	2,958	0,397	0,015	0,043	1,537	3,964	0,484	0	0,576
138*	2,197	1,838	0,22	0,368	4	0,75	2,954	1,561	0,002	0,038	1,489	3,964	0,372	0	1,53
139*	2,303	1,851	0,213	0,408	4,5	0,803	3,11	1,224	0,006	0,043	1,516	4,104	0,412	0	1,561
140*	2,303	1,884	0,183	0,512	4	0,685	3,111	0,82	0,011	0,046	1,578	4,104	0,517	0	1,274

Tableau II : (suite et fin)

N	piPC01	R2e	R4m	SIC2	GGI1	MAXDN	VEv1	HATS5u	H6m	R1p+	GNar	CIC0	BIC2	ATS8v	H5e
141*	1,609	1,516	0,025	0,266	3	0,75	2,204	0	0	0,051	1,32	3,213	0,272	0	0
142*	2,197	1,869	0,162	0,493	3,5	0,74	2,951	0,986	0,005	0,046	1,537	3,964	0,499	0	0,998
143*	2,833	1,926	0,087	0,292	0,5	0,127	4,052	0,36	0,082	0,03	1,843	4,823	0,294	2,303	1,4
144*	2,944	1,935	0,092	0,277	0,5	0,127	4,284	0,336	0,111	0,028	1,859	4,976	0,278	2,485	1,541
145*	1,099	0,75	0	0,355	0	0	1,414	0	0	0,059	1	1,667	0,355	0	0
146*	1,609	1,52	0,024	0,796	0,5	0,417	1,978	0,613	0	0,067	1,414	2,667	0,796	0	0
147*	2,197	1,677	0,066	0,684	0,5	0,306	2,788	0,641	0,005	0,06	1,682	3,667	0,684	0	0,278
148*	2,197	1,737	0,063	0,736	0,5	0,232	2,784	0,643	0,005	0,048	1,682	3,667	0,736	0	0,377
149*	2,398	1,729	0,07	0,594	0,5	0,286	3,114	0,514	0,012	0,069	1,741	3,867	0,719	1,099	0,682
150*	2,708	1,8	0,082	0,474	0,5	0,295	3,682	0,433	0,035	0,044	1,811	4,474	0,474	1,946	1,119
151*	2,773	1,812	0,085	0,452	0,5	0,295	3,811	0,412	0,051	0,041	1,823	4,574	0,452	2,079	1,216
152*	2,079	2,177	0,084	0,209	0	0	2,646	0,902	0	0,046	2	3,474	0,209	0	0,378
153*	2,197	1,988	0,127	0,478	2,5	0,571	2,804	0,943	0	0,047	1,861	3,667	0,568	0	0,305
154*	2,197	2,077	0,146	0,55	1,5	0,33	2,802	1,172	0	0,048	1,861	3,667	0,55	0	0,448
155*	2,197	1,972	0,121	0,568	2	0,321	2,802	0,963	0,001	0,047	1,861	3,667	0,568	0	0,019
156*	2,197	2,127	0,118	0,514	2	0,314	2,797	0,865	0,001	0,041	1,861	3,667	0,514	0	0,62
157*	2,197	1,967	0,14	0,521	2,5	0,514	2,791	1,174	0,001	0,055	1,834	3,667	0,521	0	0,406
158*	2,303	2,092	0,169	0,537	1,5	0,398	2,965	0,947	0,003	0,044	1,876	3,837	0,537	0	0,664
159*	2,639	2,127	0,09	0,383	1	0,217	3,553	0,351	0,011	0,051	1,956	4,367	0,383	1,792	0,818
160*	2,708	2,13	0,092	0,366	1	0,217	3,687	0,341	0,017	0,048	1,959	4,474	0,366	1,946	0,964
161*	2,89	2,127	0,104	0,327	1	0,218	4,062	0,303	0,047	0,048	1,966	4,754	0,327	2,303	1,229
162*	2,079	1,348	0,064	0,625	0,5	0,441	2,426	0,732	0,001	0,057	1,587	3,046	0,612	0	0,062
163*	1,946	1,277	0,044	0,735	1,5	0,648	2,207	0,903	0	0,082	1,431	2,739	0,714	0	0,094
164*	1,792	1,443	0,023	0,797	0,5	0,653	1,978	0,508	0	0,107	1,414	2,351	0,765	0	0
165*	1,946	1,535	0,052	0,818	0,5	0,583	2,21	0,419	0	0,098	1,516	2,739	0,795	0	0,074

PUBLICATION

ISSN: 0975-8585

Research Journal of Pharmaceutical, Biological and Chemical Sciences

QSPR Study of the Boiling Point of Diverse Hydrocarbons: Hybrid (GA/ MLR) Approach.

Nour-eddine Kertiou, Amel Bouakkadia, and Djelloul Messadi*.

Environmental and Food Safety Laboratory, Badji Mokhtar University, Annaba 23000, Algeria.

ABSTRACT

A quantitative structure- property relationship (QSPR) was performed for the prediction of the boiling points of hydrocarbons which consists of alkanes, alkenes, dienes, alkynes, cycloalkanes, and cycloalkenes. The entire set of 165 compounds was divided into a training set of 125 hydrocarbons and a test set of 40 compounds. A five descriptor model, with squared correlation coefficient (R^2) of 99.80% and standard error of estimation (s) of 4.67, was developed by applying multiple linear regression analysis using the ordinary least square regression method and genetic algorithm- variable subset selection. The reliability of the proposed model was further illustrated using various evaluation technics: leave- one- out cross- validation, bootstrap, randomization tests, and validation through the test set.

Keywords: Hydrocarbons-Boiling points- QSPR-Molecular Descriptors- Multiple Linear Regression.

Corresponding author

ISSN: 0975-8585

INTRODUCTION

Boiling point Bp is one of the most important physical property, used to describe the volatility of a compound (its presence in the atmospheric environment), defined as the temperature at which the vapor pressure of a pure saturated liquid is 760 mmHg [1]. Also, to estimate other properties such as critical temperatures, vapor pressure and flash points [2-4].

For many hydrocarbons, the values of boiling point are not available in the literature. Their experimental measurement is expensive, consumes a long time and it requires pure compounds. Moreover, the compounds of high molecular weight decompose before reaching their boiling points and require measures under reduced pressure and subsequent correction for atmospheric pressure. Therefore, the direct measurement of the boiling point of the organic compound is laborious [5].

The aim of the present work is to develop a robust QSPR[6] model that could predict the boiling point values for a diverse set of hydrocarbons (which consists of alkanes, alkenes, dienes, alkynes, cycloalkanes, and cycloalkenes) using the general molecular descriptors computed with the help of DRAGON software [7].

In this study, we present a new QSPR model for the prediction of the boiling point of a set of 165 hydrocarbons. Our goal is to develop an accurate, simple, fast, and less expensive method for calculation of boiling point values. The predictive power of resulting model is demonstrated by testing it on test data that were not used during model generation

METHODS

Experimental Data

The experimental Bp values (K) of 165 selected, structurally heterogeneous, hydrocarbons were taken from the literature [8]. The boiling point values span between 111.6 and 628.12K (Table 1). The detailed structures of all studied compounds are available as Supporting Information.

Descriptor Generation

The chemical structure of each compound was sketched on a PC using the HYPERCHEM program [9] and preoptimized using MM+ molecular mechanics method (Polack- Ribiere algorithm). The final geometries of the minimum energy conformation were obtained by the semi-empirical PM3 method at a restricted Hartree-Fock level with no configuration interaction, applying a gradient norm limit of 0.01 kcal. $Å^{-1}.mol^{-1}$ as a stopping criterion. Then the geometries were used as input for the generation of 1664 descriptors from 20 different classes such as Constitutional, Topological, Geometrical, Charge, GETAWAY (Geometry, Topology and Atoms-Weighted Assembly), WHIM (Weighted Holistic Invariant Molecular descriptors), 3D-MoRSE (3D-Molecular Representation of Structure based on Electron diffraction), and Molecular Walk Counts using Dragon software (version 5.4) [7].

Constant values and descriptors found to be correlated pairwise were excluded in a pre-reduction step (when there was more than 98% pairwise correlation, one variable was deleted), and the genetic algorithm was applied for variables selection to a final set of 1230 descriptors.

Selection of the training and test sets It is important to rationally define a training set from which the model is built and external test set on which to evaluate its prediction power. The object of this selection should be to generate two sets with similar molecular diversity, in order to be reciprocally representative and to cover all the main structural and physicochemical characteristics of the global data set.

Several procedures can be adopted for the selection of the training and test sets, the later should contain between 15 and 40% of the compounds in the full data set.

The Duplex algorithm [10] was applied in this study to separate data into two independent subsets: a training set of – compounds to build the model and a test set of the remained- compounds to evaluate its prediction ability.

ISSN: 0975-8585

The algorithm begin with a list of the n (=165) observations where the k regressors are standardized to unit length; that is,

$$z_{ij} = \frac{x_{ij} - \bar{x}_j}{S_{jj}^{1/2}}, i = 1, 2, \dots, n ; j = 1, 2, \dots, k \quad (1)$$

Where $S_{jj} = \sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2$ is the corrected sum of squares of the jth regressor. The standardized regressors are then orthonormalized. This can be done by factoring the $Z'Z$ matrix as:

$$Z'Z = T'T \quad (2)$$

Where T' is unique k×k upper triangular matrix. The elements of T can be found using the square root or cholesky method [11]. Then make the transformation

$$W = Z T^{-1} \quad (3)$$

Resulting in a new set of variables (the w's) that are orthogonal and have unit variance. Then the Euclidian distance between all possible pairs of points is calculated. The two points which are farthest apart are assigned to the estimation set. The two points in the remaining list which are farthest are assigned to the prediction set. At the third step the point which is farthest from the two points in the estimation set is added to the estimation set. At the fourth step the point which is farthest from the two points in the prediction set is included in the prediction set. The alternation between the estimation and the prediction set continues until all points in the list have been assigned to one of the two sets. Of course, once a point is assigned to a set, it is deleted from further consideration.

This algorithm was applied in the present study to separate data into two independent subsets: a training set of 125 compounds to build the model and a test set of the remained 40 compounds to evaluate its prediction ability.

Model Development and Validation

Multiple linear regression analysis (MLR) and variable selection were performed by the software MobyDigs [12] using the Ordinary Least Square regression (OLS) method and Genetic Algorithm-Variable Subset Selection (GA-VSS) [13].

The outcome of the application of the genetic algorithms is a population of 100 regression models, ordered according to their decreasing internal predictive performance, verified by Q^2. The models with lower Q^2 are those with fewer descriptors. First of all, models with 1-2 variables were developed by the all – subset – method procedure in order to explore all the low dimension combinations. The number of descriptors was subsequently increased one by one, and new models were formed. The best models are selected at each rank, and the final model must be chosen from among them. This has to be sufficiently correlated and, at the same time, protect against any over parameterization, which would lead to a loss of predictive power for molecules outside training set. From a statistical view point the ratio of the number of samples (n) to the number of descriptors (m) should not be too low. Usually, it is recommended that $n/m \geq 5$ [14]. The GA was stopped when increasing the model size did not increase the Q^2 value to any significant degree. Particular attention was paid to the collinearity of the selected molecular descriptors: by applying the QUIK rule (Q Under Influence of K) [15] a necessary condition for the model validity. Acceptable model is only that with a global correlation of [x + y] block (K_{xy}) greater than the global correlation of the x block (K_{xx}) variable, x being the molecular descriptors and y the response variable.

The collinearity in the original set of molecular descriptors results in many similar models that more or less yield the same predictive power (in MOBYDIGS software 100 models of different dimensionality). Therefore, when there were models of similar performance, those with higher ΔK (K_{xy}- K_{xx}) were selected and further verified.

In this work, the "breaking point" rule was used to manage this problem. This method consists of analysing the improvement in the correlation with the number of variables in the model. By plotting the R^2 values as functions of the number of descriptors, asymptotic behavior was observed, and the improvement in the correlation became less significant after a certain rank (ΔR^2<0.02-0.03). At this point (the "breaking point"), the model is considered to be optimal, representing the best compromise between correlation and parameterization.

The models were justified by the R^2, the adjusted R^2, the external Q^2_{ext}, the F ratio values, the standard error of estimation s and the significance level value p. the R^2 and adjusted R^2 were calculated using the following formula:

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y}_i)^2} \quad (4)$$

$$R^2_{adj} = 1 - \left[\frac{N-1}{N-M-1}(1 - R^2)\right] \quad (5)$$

Where N is the number of members of the training set and M is the number of descriptors involved in the correlation. The adjusted R^2 is a better measure of the proportion of variance in the data explained by the correlation than R^2, because R^2 is somewhat sensitive to changes in N and M. The adjusted R^2 corrects for the artificiality introduced when M approaches N through the use of a penalty function which scales the result. A variance inflation factor (VIF) was calculated to test if multicollnearities existed among the descriptors, which is defined as:

$$VIF = \frac{1}{1 - R_j^2} \quad (6)$$

Where R_j^2 is the squared correlation coefficient between the jth coefficient regressed against all the other descriptors in the model. Models would not be accepted if they contain descriptors with VIFs above a value of five.

Randomization tests were also carried out to prove the possible existence of chance correlation. To do this, the dependent variable was randomly scrambled and used in the experiment. Models were then investigated with all members in the descriptor pool to find the most predictive models. The resulting models obtained on the training set with the randomized IR values should have significantly lower R^2 values than the proposed one because the relationship between the structure and property is broken. This is a proof of the proposed model's validity as it can be reasonably excluded that the originally proposed mode was obtained by chance correlation.

Validation of the models was further performed by using the external test set composed of data not used to develop the prediction model. The Q^2_{ext} is determined with Eq. (7):

$$Q^2_{ext} = 1 - [(\sum_{i=1}^{n_{ext}} (y_i - \hat{y}_{(i)})^2 / n_{ext}) / (\sum_{i=1}^{n_{tr}} (y_i - \bar{y}_{tr})^2 / n_{tr})] \quad (7)$$

Here n_{ext} and n_{tr} are the number of objects in the external set (or left out by bootstrap) and the number of training set objects, respectively.

According to Golbraikh et al, [16,17] a QSPR model can provide an acceptable prediction if it verifies the following conditions:

$$Q^2_{EXT} > 0.5 \quad (8\text{-a})$$

$$r^2 > 0.6 \quad (8\text{-b})$$

$$(r^2 - r_0^2)/r^2 < 0.1 \text{ or } (r^2 - r_0'^2)/r^2 < 0.1 \quad (8\text{-c})$$

$$0.85 < k < 1.15 \text{ or } 0.85 < k' < 1.15 \quad (8\text{-d})$$

r^2 is the correlation coefficient between the calculated and experimental values in the test set; r^2_0 (calculated versus observed versus) and r'^2_0 (observed versus calculated values) are the coefficients of determination ; k, k' are slopes of the regression lines through the origin of calculated versus observed and observed versus respectively.

Here

$$r = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})}{\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2 \sum(\hat{y}_i - \bar{\hat{y}})^2}} \quad \text{(9-a)}$$

$$r_0^2 = 1 - \frac{\sum_{i=1}^{n}(\hat{y}_i - \hat{y}_i^{r0})^2}{\sum_{i=1}^{n}(\hat{y}_i - \bar{\hat{y}})^2} \quad \text{(9-b)}$$

$$r'^2_0 = 1 - \frac{\sum_{i=1}^{n}(y_i - y_i^{r0})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \quad \text{(9-c)}$$

$$k = \frac{\sum y_i \hat{y}_i}{\sum y_i^2} \quad \text{(9-d)}$$

$$k' = \frac{\sum y_i \hat{y}_i}{\sum \hat{y}_i^2} \quad \text{(9-e)}$$

Where $\hat{y}^{r0}$ and y^{r0} are defined as $\hat{y}^{r0} = ky$ and $y^{r0} = k'\hat{y}$, respectively.

The reason to use r_0^2 and require k values that are chose to 1 is that when actual versus predicted boiling points are compared, an exact fit is required, not just a correlation.

The robustness of the models and their predictivity were evaluated by both Q^2_{LOO} and bootstrap. In this last procedure K n-dimensional groups are generated by a randomly repeated selection of n- objects from the original data set.

The model obtained on the first selected objects is used to predict the values for the excluded sample, and then Q^2 is calculated for each model. The bootstrapping was repeated 8000 times.

The proposed model was also checked for reliability and robustness by permutation testing: new models are recalculated for randomly recorded response (Y- scrambling) by using the same original independent variable matrix. After repeating this test several times (100 times in this work) it is expected to obtain new models that have significantly lower R^2 and Q^2 than the original model. If this condition is not verified the original model is not acceptable, as it was due to a chance correlation or a structural redundancy in the training set.

Obtaining a robust model does not give real information about its prediction power. This is evaluated by predicting the compounds included in the test set.

ISSN: 0975-8585

Analysis

The applicability domain (AD) [18,19] is a theoretical region in the space defined by the descriptors of the model and the modeled response, for which a given QSPR should make reliable predictions. In this work, the structural AD was verified by the leverage (h_{ii}) approach [20].

The warning leverage h^* is, generally, fixed at $3(m + 1)/n$, where n is the total number of samples in the training set and m is the number of descriptors involved in the correlation.

The presence of both the response outliers (Y outliers) and the structurally influential compounds (X outliers) was verified by the Williams plot [21], the plot of standardized residuals versus leverage values.

Table 1: Experimental and calculated Bp for the studied compounds

		Descriptors					Boiling Point		
N	Sample	MAXDN	VEv1	HATS5u	H6m	R1p+	Expt.	Calc.	Residual
1	Methane	0	1	0	0	0	111.6	102.67	8.93
2	Ethylene	0	1.414	0	0	0.059	169.4	188.41	-19.01
3	Ethane	0	1.414	0	0	0.059	184.5	188.41	-3.91
4	Acetylene	0	1.414	0	0	0.057	188.4	187.62	0.78
5	Propylene	0.25	1.715	0	0	0.076	225.5	232.59	-7.09
6	Propane	0.25	1.716	0	0	0.061	231	227.46	3.54
7	Propadiene	0.25	1.716	0	0	0.072	238.7	231.37	7.33
8	Cyclopropane	0	1.732	0	0	0.072	240.3	243.14	-2.84
9	Propyne	0.347	1.714	0	0	0.099	249.9	236.76	13.14
10	Isobutane	0.5	1.972	0	0	0.059	261.4	258.11	3.29
11	isobutylene	0.5	1.969	0	0	0.068	266.2	260.74	5.46
12*	But-1-ene	0.417	1.978	0.613	0	0.067	266.9	270.04	-3.14
13*	Buta-1,3-diene	0.361	1.985	0.331	0	0.062	268.7	268.98	-0.28
14	Butane	0.181	1.974	0.658	0	0.056	272.6	274.72	-2.12
15	E-But-2-ene	0	1.964	1.195	0	0.067	274	288.01	-14.01
16	Z-But-2-ene	0	1.964	1.062	0	0.068	276.9	287.36	-10.46
17*	vinylacetylene	0.597	1.987	0.288	0	0.097	278.1	272.25	5.85
18	But-1-yne	0.653	1.978	0.508	0	0.107	281.2	274.2	7
19	2,2-Dimethylpropane	0.75	2.204	0	0	0.051	282.6	282.79	-0.19
20	3,3-Dimethylhexane	0.658	2.779	0.997	0.002	0.044	284.00	286.17	-2.17
21*	Cyclobutane	0	2	0	0	0.057	285.7	280.21	5.49
22*	3-Methyl but-1-ene	0.685	2.206	0.9	0	0.072	293.3	300.14	-6.84
23	Penta-1,4-diene	0.583	2.217	0.348	0	0.059	299.1	296.69	2.41
24	But-2-yne	0.181	1.959	1.956	0	0.071	300.1	288.43	11.67
25	Isopentane	0.449	2.202	0.88	0	0.057	301	303.06	-2.06
26	Pent-1-ene	0.347	2.209	0.644	0	0.063	303.1	308.02	-4.92
27	2-Methyl but-1-ene	0.412	2.2	1.011	0	0.062	304.3	306.93	-2.63
28	2-Methyl buta-1,3-diene	0.648	2.207	0.903	0	0.082	307.2	305.29	1.91
29	Pentane	0.156	2.204	0.806	0.001	0.054	309.2	312.31	-3.11
30*	E-pent-2-ene	0.337	2.196	0.918	0.002	0.063	309.5	308.16	1.34
31	Z-pent-2-ene	0.337	2.196	0.795	0.001	0.06	310.1	306.19	3.91
32	2-Methyl but-2-ene	0.287	2.191	1.36	0	0.063	311.7	313.41	-1.71
33	Pent-1-yne	0.583	2.21	0.419	0	0.098	313.3	309.6	3.7
34	3-Methyl buta-1,2-diene	0.523	2.194	1.453	0	0.086	314	313.86	0.14

ISSN: 0975-8585

N	Sample	Descriptors					Boiling Point		
		MAXDN	VEv1	HATS5u	H6m	R1p+	Expt. Bp	Calc. Bp	Residual
35*	3,3-Dimethyl but-1-ene	0.944	2.417	0.862	0	0.069	314.4	322.72	-8.32
36	E-penta-1,3-diene	0.25	2.202	0.748	0	0.083	315.2	318.25	-3.05
37	cyclopentene	0.181	2.234	0	0	0.07	317.4	315.1	2.3
38	Penta-1,2-diene	0.455	2.202	0.635	0.001	0.083	318	309.33	8.67
39	Cyclopentane	0	2.236	0	0	0.049	322.4	314.88	7.52
40*	2,2-Dimethylbutane	0.708	2.411	0.914	0	0.051	322.9	324.57	-1.67
41*	2,3-Dimethyl but-1-ene	0.676	2.411	1.385	0	0.068	328.8	335.57	-6.77
42	Z-4-Methyl pent-2-ene	0.616	2.407	0.739	0.002	0.058	329.6	327.97	1.63
43	2,3-Dimethylbutane	0.481	2.412	1.295	0	0.052	331.1	336.84	-5.74
44*	E-4-Methyl pent-2-ene	0.616	2.407	0.689	0.002	0.065	331.7	329.79	1.91
45	Hexa-1,5-diene	0.441	2.426	0.732	0.001	0.057	332.6	337.28	-4.68
46	2-Methylpentane	0.435	2.412	0.762	0.001	0.054	333.4	334.51	-1.11
47*	3- Methylpentane	0.398	2.409	1.226	0.001	0.054	336.4	339.14	-2.74
48	Hex-1-ene	0.323	2.418	0.721	0.001	0.062	336.6	341.85	-5.25
49	Z-Hex-3-ene	0.326	2.403	0.637	0.004	0.052	339.6	334.66	4.94
50*	E-Hex-3-ene	0.326	2.403	0.613	0.003	0.056	340.3	335.91	4.39
51	2-Methyl pent-2-ene	0.331	2.401	0.858	0.002	0.061	340.5	339.59	0.91
52*	Z-3-Methyl pent-2-ene	0.309	2.4	1.429	0.001	0.051	340.9	341.69	-0.79
53	E-Hex-2-ene	0.267	2.408	0.635	0	0.06	341	341.32	-0.32
54	Hexane	0.145	2.412	0.712	0.003	0.047	341.9	341.92	-0.02
55*	Z-Hex-2-ene	0.267	2.408	0.661	0.003	0.056	342	339.44	2.56
56	E-3-Methyl pent-2-ene	0.309	2.4	1.368	0.001	0.058	343.6	343.74	-0.14
57	Methylcyclopentane	0.287	2.431	0.558	0	0.053	344.9	341.16	3.74
58	2,3-Dimethyl but-2-ene	0.241	2.4	1.74	0	0.051	346.4	347.21	-0.81
59	1-Methylcyclopentene	0.175	2.423	0.83	0	0.057	348.95	347.7	1.25
60	2,3,3-Trimethyl but-1-ene	0.944	2.607	1.423	0	0.052	351	351.62	-0.62
61	2,2-Dimethylpentane	0.7	2.604	0.654	0.001	0.048	352.3	352.08	0.22
62	2,4-Dimethylpentane	0.458	2.607	0.671	0.004	0.048	353.6	360.76	-7.16
63*	Cyclohexane	0	2.449	0.554	0	0.046	353.9	352.03	1.87
64	2,2,3-Trimethylbutane	0.75	2.607	1.376	0	0.047	354	356.67	-2.67
65	Cyclohexene	0.181	2.448	0.515	0	0.063	356.1	350.84	5.26
66	3,3-Dimethylpentane	0.667	2.601	1.291	0	0.045	359.2	357.39	1.81
67*	1,1Dimethylcyclopentane	0.556	2.62	0.812	0	0.053	361	363.05	-2.05
68	2,3-Dimethylpentane	0.468	2.603	1.282	0.001	0.049	362.9	366.12	-3.22
69	2-Methylhexane	0.431	2.605	0.594	0.004	0.052	363.2	362.34	0.86
70	E-1,2-dimethylcyclopentane	0.319	2.616	1.055	0	0.048	365	371.67	-6.67
71*	3-Methylhexane	0.384	2.601	0.964	0.002	0.051	365	366.52	-1.52
72*	3-Ethylpentane	0.347	2.598	1.241	0.003	0.052	366.6	369.85	-3.25
73	Hept-1-ene	0.312	2.61	0.701	0.003	0.061	366.8	371.76	-4.96
74	Heptane	0.139	2.604	0.642	0.004	0.046	371.6	371.31	0.29
75*	2,2,4-Trimethylpentane	0.728	2.788	0.589	0.005	0.044	372.4	377.25	-4.85
76	Z-1,2-Dimethylcyclopentane	0.319	2.616	1.166	0	0.051	372.7	373.64	-0.94
77	Methylcyclohexane	0.297	2.629	0.776	0	0.048	374.1	372.19	1.91
78	Ethylcyclopentane	0.236	2.614	0.798	0.001	0.055	376.6	374.43	2.17
79	1,1,3-Trimethylcyclopentane	0.584	2.792	0.789	0.001	0.046	378	386.53	-8.53
80	1-Ethylcyclopentene	0.222	2.605	0.943	0.001	0.058	379.45	375.78	3.67
81*	2,2,3,3-Tetramethylbutane	0.813	2.789	1.564	0	0.037	379.6	381.06	-1.46

ISSN: 0975-8585

N	Sample	Descriptors					Boiling Point		
		MAXDN	VEv1	HATS5	H6m	R1p+	Expt. Bp	Calc.	Residual
82	2,2-Dimethylhexane	0.698	2.784	0.464	0.006	0.048	380	377.81	2.19
83	2,5-Dimethylhexane	0.447	2.788	0.525	0.011	0.043	382.3	385.32	-3.02
84*	2,4-Dimethylhexane	0.454	2.784	0.778	0.005	0.047	382.6	389.37	-6.77
85*	2,2,3-Trimethylpentane	0.741	2.784	1.269	0.001	0.045	383	382.97	0.03
86	3,3-Dimethylhexane	0.658	2.779	0.997	0.002	0.044	385.1	382.56	2.54
87*	2,3,4-Trimethylpentane	0.491	2.785	1.246	0.002	0.048	386.6	392.98	-6.38
88	1,1,2-Trimethylcyclopentane	0.597	2.795	1.386	0	0.048	386.9	392.29	-5.39
89	2,3,3-Trimethylpentane	0.708	2.783	1.536	0	0.042	387.9	385.67	2.23
90	2,3-Dimethylhexane	0.463	2.782	1.007	0.004	0.045	388.8	390.19	-1.39
91*	3-Ethyl-2-methylpentane	0.454	2.779	1.221	0.005	0.046	388.8	391.96	-3.16
92*	2-Methylheptane	0.429	2.785	0.511	0.007	0.05	390.8	388.58	2.22
93	3,4-Dimethylhexane	0.417	2.78	1.273	0.003	0.044	390.9	393.71	-2.81
94*	4-Methylheptane	0.37	2.779	0.791	0.004	0.049	390.9	392.72	-1.82
95	3-Ethyl-3-methylpentane	0.625	2.777	1.479	0.001	0.051	391.4	390.17	1.23
96*	Cycloheptane	0	2.646	0.902	0	0.046	391.6	386.23	5.37
97	3-Ethylhexane	0.333	2.776	0.949	0.004	0.051	391.7	395.63	-3.93
98*	3-Methylheptane	0.37	2.779	0.791	0.004	0.049	392.1	392.72	-0.62
99	E-1,4-Dimethylcyclohexane	0.314	2.797	0.865	0.001	0.041	392.5	396.22	-3.72
100*	1,1-Dimethylcyclohexane	0.571	2.804	0.943	0	0.047	392.7	390.6	2.1
101*	Z-1,3-Dimethylcyclohexane	0.321	2.802	0.877	0	0.048	393.3	399.33	-6.03
102*	Oct-1-ene	0.306	2.788	0.641	0.005	0.06	394.4	398.62	-4.22
103	1-Ethyl-1-methylcyclopentane	0.514	2.791	1.174	0.001	0.055	394.7	395.29	-0.59
104	2,2,4,4-Tetramethylpentane	0.766	2.96	0.547	0.009	0.039	395.4	400	-4.6
105	E-1,2-Dimethylcyclohexane	0.33	2.802	1.142	0	0.045	396.6	400.36	-3.76
106	2,2,5-Trimethylhexane	0.718	2.958	0.397	0.015	0.043	397.2	400.14	-2.94
107*	Z-1,4-Dimethylcyclohexane	0.314	2.797	0.865	0.001	0.041	397.5	396.22	1.28
108	E-1,3-Dimethylcyclohexane	0.321	2.802	0.963	0.001	0.047	397.6	399.64	-2.04
109	E-Oct-2-ene	0.232	2.784	0.643	0.005	0.048	398.1	396.78	1.32
110*	Octane	0.135	2.782	0.582	0.006	0.043	398.8	397.41	1.39
111	Isopropylcyclopentane	0.391	2.795	0.927	0.001	0.048	399.6	395.99	3.61
112*	2,2,4-Trimethylhexane	0.727	2.955	0.617	0.009	0.046	399.7	403.65	-3.95
113	Z-1,2-Dimethylcyclohexane	0.33	2.802	1.172	0	0.048	402.9	401.47	1.43
114	Propylcyclopentane	0.222	2.789	0.622	0.003	0.051	404.1	399.27	4.83
115*	Ethylcyclohexane	0.247	2.797	0.913	0.001	0.049	404.9	401.85	3.05
116	2,2-Dimethylheptane	0.699	2.954	0.375	0.01	0.046	405.8	402.25	3.55
117	2,2,3,4-Tetramethylpentane	0.77	2.956	1.177	0.004	0.042	406.2	406.74	-0.54
118*	2,2,3-Trimethylhexane	0.74	2.951	0.986	0.005	0.046	406.8	406.59	0.21
119	2,2,5,5-Tetramethylhexane	0.743	3.122	0.337	0.026	0.035	410.6	419.05	-8.45
120	2,2,3,3-Tetramethylpentane	0.804	2.954	1.564	0.001	0.043	413.4	409.49	3.91
121	1,E-3,5-Trimethylcyclohexane	0.344	2.967	0.845	0.001	0.046	413.7	423.53	-9.83
122	2,3,3,4-Tetramethylpentane	0.75	2.954	1.561	0.002	0.038	414.7	409.5	5.2
123	2-Methyloctane	0.429	2.954	0.459	0.008	0.048	416.4	414.1	2.3
124	3,3-Diethylpentane	0.583	2.943	1.441	0.004	0.049	419.3	416.23	3.07
125	Non-1-ene	0.302	2.956	0.592	0.007	0.056	420	423.22	-3.22
126	Cyclooctane	0	2.828	0.933	0	0.044	422	414.57	7.43
127	Nonane	0.133	2.951	0.536	0.007	0.041	424	422.88	1.12

ISSN: 0975-8585

N	Sample	Descriptors					Boiling Point		
		MAXDN	VEv1	HATS5	H6m	R1p+	Expt. Bp	Calc. Bp	Residual
128	Isopropylcyclohexane	0.398	2.965	0.947	0.003	0.044	427.7	420.91	6.79
129	3,3,5-Trimethylheptane	0.685	3.111	0.82	0.011	0.046	428.8	431.09	-2.29
130	Propylcyclohexane	0.233	2.959	0.72	0.002	0.048	429.9	425.77	4.13
131	2,2,3,3-Tetramethylhexane	0.803	3.11	1.224	0.006	0.043	433.5	430.13	3.37
132	Deca-1,3-diene	0.286	3.114	0.514	0.012	0.069	442	451.41	-9.41
133	Dec-1-ene	0.299	3.115	0.548	0.008	0.054	443.7	447.17	-3.47
134	Isobutylcyclohexane	0.414	3.122	0.602	0.007	0.045	444.5	441.62	2.88
135*	tert-Butylcyclohexane	0.68	3.127	0.949	0.005	0.042	444.7	434.94	9.76
136	Decane	0.131	3.11	0.502	0.008	0.04	447.3	447.38	-0.08
137	sec-Butylcyclohexane	0.347	3.119	0.947	0.011	0.045	452.5	445.57	6.93
138	Butylcyclohexane	0.228	3.116	0.586	0.005	0.048	454.1	448.92	5.18
139	Undec-1-ene	0.298	3.266	0.512	0.01	0.051	465.8	469.26	-3.46
140	Undecane	0.13	3.261	0.471	0.01	0.038	469.1	469.86	-0.76
141	Hexylcyclopentane	0.216	3.268	0.404	0.006	0.052	476.3	473	3.3
142	Dodec-1-ene	0.297	3.41	0.481	0.013	0.049	486.5	490.39	-3.89
143	Dodecane	0.129	3.406	0.448	0.012	0.037	489.5	491.8	-2.3
144	Heptylcyclopentane	0.216	3.414	0.375	0.008	0.05	497.3	494.66	2.64
145	Tridec-1-ene	0.296	3.549	0.455	0.021	0.045	505.9	508.68	-2.78
146	Tridecane	0.128	3.544	0.425	0.022	0.035	508.6	510.38	-1.78
147	Octylcyclopentane	0.217	3.553	0.351	0.011	0.051	516.9	516.02	0.88
148	Tetradec-1-ene	0.295	3.682	0.433	0.035	0.044	524.3	525.85	-1.55
149	Tetradecane	0.128	3.678	0.407	0.036	0.034	526.7	527.5	-0.8
150	Nonylcyclopentane	0.217	3.687	0.341	0.017	0.048	535.3	534.65	0.65
151	Pentadec-1-ene	0.295	3.811	0.412	0.051	0.041	541.5	541.2	0.3
152	Pentadecane	0.128	3.807	0.389	0.051	0.032	543.8	543.68	0.12
153	Decylcyclopentane	0.218	3.816	0.324	0.021	0.048	552.5	553.9	-1.4
154	Hexadec-1-ene	0.294	3.935	0.395	0.067	0.04	558	556.76	1.24
155	Hexadecane	0.127	3.932	0.375	0.066	0.031	560	559.43	0.57
156*	Decylcyclohexane	0.228	3.94	0.383	0.031	0.046	570.8	570.32	0.48
157	Heptadecane	0.127	4.052	0.36	0.082	0.03	575.2	574.11	1.09
158*	Dodecylcyclopentane	0.218	4.062	0.303	0.047	0.048	584.1	586.16	-2.06
159	Octadec-1-ene	0.294	4.173	0.364	0.099	0.036	588	585.1	2.9
160	Octadecane	0.127	4.17	0.348	0.096	0.029	589.5	588.97	0.53
161	Tridecylcyclopentane	0.219	4.179	0.3	0.063	0.046	598.6	600.32	-1.72
162*	1-Cyclopentyltetradecane	0.219	4.294	0.286	0.08	0.046	599	614.06	-15.06
163	Nonadecane	0.127	4.284	0.336	0.111	0.028	603.1	602.96	0.14
164	Eicosane	0.126	4.395	0.325	0.125	0.026	617	616.42	0.58
165	1-Cyclopentylpentadecane	0.22	4.405	0.284	0.094	0.045	625	628.12	-3.12

RESULTS AND DISCUSSION

Results of the MLR Model

A multiple linear regression (MLR) was employed to describe the relation between critical properties and their molecular descriptors. The best model and the number of descriptors (p) in the final QSPR model was determined on the basis of the correlation coefficient R^2. At first, the optimal p is tested using p= 2 to 8. An increase of the R^2 value less than 0.02 was chosen as a threshold. Figure. 1 shows the application of the

breaking point criterion [22] in the present case suggest a best five-parameters equation was obtained, which is as the following:

$$Bp = -56.17 + 157.85\ \mathbf{VEv1} - 34.74\ \mathbf{MAXDN} + 7.62\ \mathbf{HATS5u} - 231.67\ \mathbf{H6m} + 360.75\ \mathbf{R1p+} \quad (10)$$

$R^2 = 99.77\%$ $R^2_{adj} = 99.80\%$ $Q^2_{LOO} = 99.73\%$ $Q^2_{EXT} = 99.57\%$ $Q^2_{BOOT} = 99.61\%$ $s = 4.79$

$F = 10423.55$ $K_{xx} = 41$ $K_{xy} = 50.27$

Here, VEV1 is the eigenvector coefficient sum from van der Waals weighted distance matrix; MAXDN is the maximal electrotopological negative variation [23,24] ; HATS5u is the leverage-weighted autocorrelation of lag 5 / unweighted [25,26] ; H6m is the H autocorrelation of lag 6 / weighted by atomic masses ; R1p+ is the R maximal autocorrelation of lag 1 / weighted by atomic polarizabilities [25,26]

More information about these descriptors can be found in [27] and the references therein.

The results for the randomized models can be compared with the real starting one only by representing in a plot the statistical coefficients R^2 and Q^2. This is depicted in figure. 2. The statistics for the modified Bp vectors are clearly lower than the real QSPR model. This ensures that a real structure-property relationship has been found out.

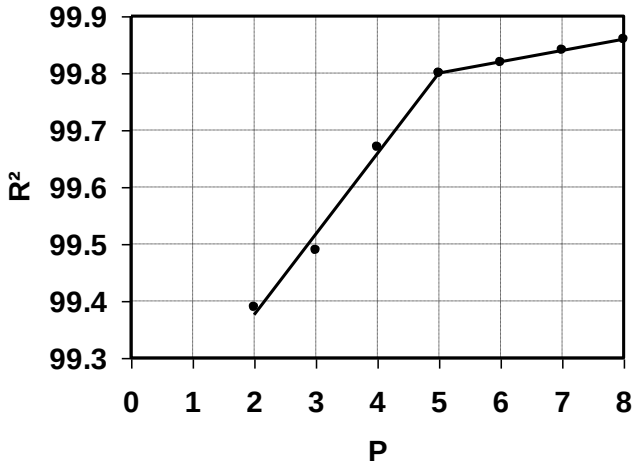


Figure 1: Breaking point rule for determination of the optimum number of the descriptors

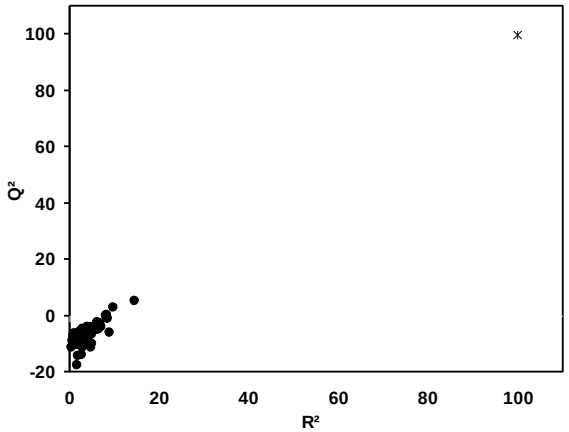


Figure 2: Randomization test associated to previous QSPR model. Black circles represent the randomly ordered, and star corresponds to the real boiling points.

ISSN: 0975-8585

Some important statistical parameters (as given in table 2) were used to evaluate the involved descriptors. The t-value of a descriptor measures the statistical significance of the regression coefficients. The high absolute t-values shown in table 2 express that the regression coefficients of the descriptors involved in the MLR model are significantly larger than the standard deviation. The t-probability of a descriptor can describe the statistical significance when combined together within an overall collective QSPR model (i. e., descriptor's interactions). Descriptors with t-probability values below 0.05 (95% confidence) are usually considered statistically significant in a particular model, which means that their influence on the response variable is not merely by chance [28]. The smaller t-probability suggests the more significant descriptor. The t-probability values of the five descriptors are very small, indicating that all of them are highly significant descriptors. The VIF values suggest that these descriptors are weakly correlated with each others. Thus, the model can be regarded as an optimal regression equation.

For the training and test set are showed in table 1 and figure. 2. Regression lines were used for comparing the values obtained by this model with experimental values. As can be seen from figure. 3, the calculated slope and intercept (a=0.998 ; b=0.88) did not differ greatly from the "ideal" values of 1 and 0, respectively, and most of the predicted Bp values agreed, for all the training and testing sets. Thus, model has been developed that calculate the Bp values for hydrocarbons with accuracy comparable to experiment.

The distribution of errors for the entire data set is given in figure. 4. Residuals are distributed normally around zero (the mean value) as can be clearly seen from the histogram in the right side of the plot,

Table 2: Characteristics of the selected descriptors in the best MLR model

Descriptor	Descriptor type	X	Dx	t- value	t- probability	VIF
Constant		-56.17	3.96	-14.20	0	
VEv1	Eigenvalue-based indices	157.85	1.13	140.32	0	3.03
MAXDN	Topological descriptors	-34.74	2.21	-15.72	0	1.24
HATS5u	GETAWAY descriptors	7.62	1.13	6.75	0	1.30
H6m	GETAWAY descriptors	-231.67	30.08	-7.70	0	2.86
R1p+	GETAWAY descriptors	360.75	35.64	10.12	0	1.50

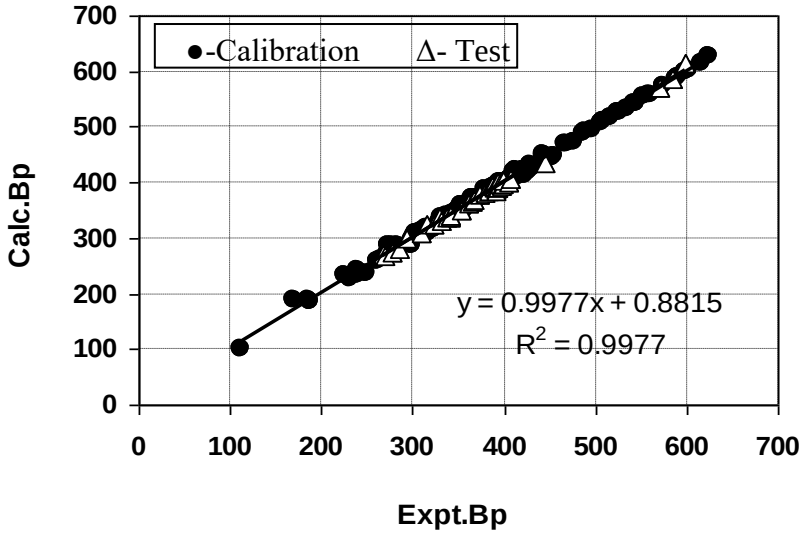


Figure 3: Plot of predicted *vs*. experimental Bp for the entire data set.

ISSN: 0975-8585

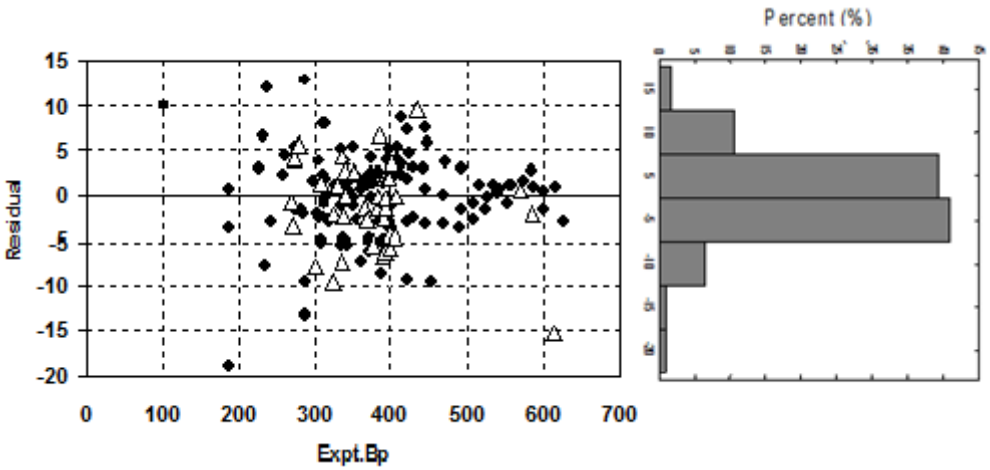


Figure 4: Plot of residual *vs*. experimental Bp for the entire data set.

Descriptor Contribution Analysis and Interpretation

Based on a previously described procedure [29, 30], the relative contribution of the five descriptors to the model were determined and they decrease in the following order: VEv1(67.01%) > MAXDN(11.36%) > HATS5u (07.50%) > H6m (07.29%) > R1p+ (06.84%) . It should be noted that the difference in the descriptor contribution between the three last descriptors used in the model is not significant, but the first one had a very high contribution indicating that these descriptor is indispensable in generating the predictive model (Figure.5).

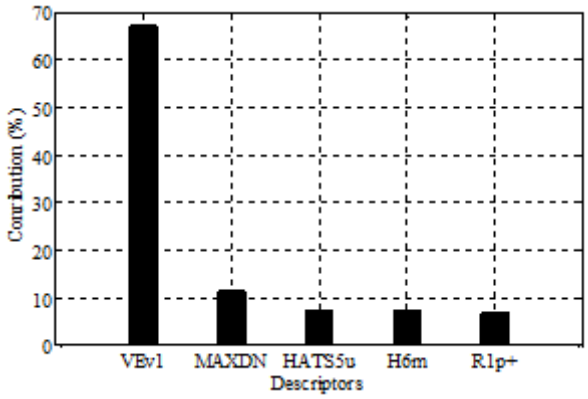


Figure 5: Relative contributions of the selected descriptors to the MLR model.

The first important descriptor is VEv1, which has a relatively very high positive correlation with the experimental Bp values (R= 99.95%). The positive coefficient of VEv1 indicates that the hydrocarbons with larger values for this descriptor would have higher Bp values.

The second important descriptor is MAXDN, a topological descriptor, which has a smaller negative correlation coefficient with the experimental Bp values (R= -10%). The electrotopological state indices are atomic indices calculated from a H-depleted molecular graph as:

$$S_i = I_i + \Delta I_i = I_i + \sum_{j=1}^{A} \frac{I_i - I_j}{(d_{ij}+1)^k} \qquad (11)$$

where I_i is the intrinsic state of the *i*th atom and ΔI_i is the field effect on the *i*th atom calculated as perturbation of the intrinsic state of *i*th atom by all other atoms in the molecule, the MAXDN is calculated as the maximum negative value of ΔI_i in the molecule; d_{ij} is the topological distance between the *i*th and the *j*th

atoms; A is the number of non-hydrogen atoms in the molecule. The exponent *k* is a parameter to modify the influence of distant or nearby atoms for particular studies. In DRAGON it is taken as *k* = 2.

The last three descriptors are HATS5u, H6m and R1p+, there are a GETAWAY descriptors and correlates with the experimental Bp values of -5.40 (p=0.5), 74 and -54.4% respectively. The GETAWAY descriptors [25,26] have been proposed as chemical structure descriptors derived from a new representation of molecular structure, the molecular influence matrix. These descriptors, as based on spatial autocorrelation, encode information on molecular space. Moreover, they are independent of molecule alignment and, to some extent, account also for information on molecular size and shape as well as for specific atomic properties.

HATS5u, H6m and R1p+ are calculated by Eq. (12), (13) and (14) respectively.

$$HATSkw = \sum_{i=1}^{A}\sum_{j>1}^{A}(w_i.h_i)(w_j.h_j).\delta(k;d_{ij}) \qquad \text{for k=1,2,3,...D} \qquad (12)$$

$$RTw+ = \max_{ij}\left(\frac{\sqrt{h_{ii}-h_{jj}}}{r_{ij}}.w_i.w_j.\delta(k;d_{ij})\right) \qquad i \neq j \quad \text{and} \quad k=1,2,3,\ldots,D \qquad (13)$$

$$Hkw = \sum_{i=1}^{A}\sum_{j>1}^{A}h_j.w_i.w_j.\delta(k;d_{ij};h_{ij}) \qquad \text{for k=1,2,3,...D} \qquad (14)$$

where *A* is the number of atoms, w is an atomic weighting scheme, d_{ij} is the topological distance, $\delta(k, d_{ij})$ is a Dirac- delta function ($\delta=1$ *if* $d_{ij}=k$, zero otherwise), r_{ij} is the interatomic distance. D is the molecule topological diameter that is the maximum topological distance in the molecule. The coefficient of R1p+ is positive, meaning that the hydrocarbons with larger values for this descriptor have larger Bp values.

The following statistical parameters obtained for the external tests set verify the well-accepted conditions (8-a to 8-d), which reinforces the predictive capabilities of the present model.

$$Q^2_{EXT} = 0.9971 > 0.5 \qquad r^2 = 0.996 > 0.6$$
$$(r^2 - r_0^2)/r^2 = (0.996 - 0.9997)/0.996 = -0.004 < 0.1$$
$$\text{or} \quad (r^2 - r_0'^2)/r^2 = (0.996 - 0.9997)/0.996 = -0.004 < 0.1$$

$$0.85 < k = 0.9965 < 1.15 \quad \text{or} \quad 0.85 < k' = 1.003 < 1.15$$

Applicability Domain of the MLR Model

Before a QSPR model is put into use for screening compounds, its applicability domain must be defined and predictions for only those compounds that fall in this domain can be considered as reliable.

The AD of the MLR model was analyzed in the Williams plot (shown in figure.6). There are three X outliers (Compounds 1, 64 and 65) with leverage higher than the warning limit of 0.14 is a structurally influential compound, and one Y outlier with residual higher than ± 3 (Compound 66) in the training set. Deleting these observations could alter slightly R^2 between the experimental Bp values and the selected descriptors to 99.75% (Q^2 = 99.71%) and decrease the standard error to 4.58, while utilization of a higer energy conformation geometry for this observation alter negatively the calculated model.

Validation

In order to estimate the predictive power of MLR, in this case we used two validation procedures. Firstly, using the leave-one-out procedure; a Q^2_{LOO} =99.70% and the bootstrap procedure a Q^2_{BOOT} =99.67%, reveal the high predictive ability of the model. Secondly the external validation procedure; by using a set of 40 compounds which have not been explored for training set. The external predictive power is confirmed by a

high Q^2_{ext} value (Q^2_{ext}=99.70%) that reveals model applicability also to predict the boiling points of unknown series compounds. The plot of predicted versus experimental values for data set is shown in figure. 3(Δ).

Remains to be noted that there is a single Y (Compound 162) outlier with residual higher than ± 3

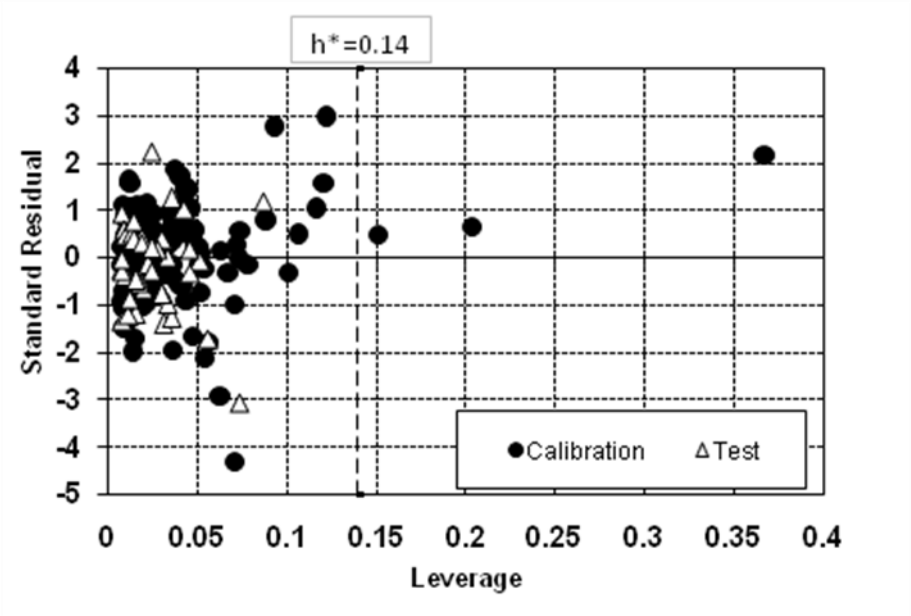


Figure 6: Williams plot of the MLR model for the entire data set.

CONCLUSION

In this paper, the QSPR method was applied to the prediction of the boiling points of organic compounds. A five-parameter linear model was developed by hybrid GA/ MLR approach with R^2 of 99.80 and s of 4.67 for the training set. The selected descriptors express many factors influencing boiling points, to name: molecular size and shape, specific atomic properties. Several validation techniques, including leave-one-out cross-validation and bootstrap, randomization tests, and validation through the test set, illustrated the reliability of the proposed model. All of the descriptors can be directly calculated from the molecular structure of the compound, thus the proposed model is predictive and could be used to estimate the boiling points of hydrocarbons. In this case, the applicability domain will serve as a valuable tool to filter out "dissimilar" chemical structures.

REFERENCES

[1] F. Gharagheizi, S. A. Mirkhani, P. Ilani-Kashkouli, et al, Fluid Phase Equil., 2013, (354), 250.
[2] A.R. Katritzky, M. Kuanar, S. Slavov, C.D. Hall, M. Karelson, I. Kahn, D.A. Dobchev, Chemical Reviews., 2010, (110), 5714.
[3] T. Le, V.C. Epa, F.R. Burden, D.A. Winkler, Chemical Reviews., 2012, (112), 2889.
[4] W.J. Lyman, W.F. Reechl, D.H. Rosenblatt, Handbook of Chemical Property Estimation Methods, American Chemical Society, Washington, DC, 1990.
[5] D. Yi-min, Z. Zhi-ping, C. Zhong, Z. Yue-fei, Z. Ju-lan, and L. Xun, J. Mol. Graphics Modell., 2013, (44), 113.
[6] Katritzky A. R., Fara D. C., Energy Fuel., 2005, (19), 922.
[7] Todeschini R., Consonni V., Mauri A., Pavan M., 2005. DRAGON Software – version 5.4-TALETEsrl
[8] R.C. Reid, J.M. Prausnitz, B.E. Poling, The Properties of Gases & liquids,Fourth Edition, Mc Graw-Hill Book Company, New York, 1987.
[9] Hyperchem™. Release 6.02 for windows. 2000. Molecular Modeling system
[10] Snee R D., *Technometrics*, 1977, (19), 415.
[11] See Graybill, 1976/ Graybill, F. A. Theory and Application of the Linear Model, Duxbury, North Scituate, Mass. pp. 231-236.

[12] Todeschni R., Ballabio D., Consonni V., Mauri A., Pavan M., 2009. MOBYDIGS – version 1.1 – Copyright TALETE srl (2004).
[13] Leardi R., Boggia R., Tarrile M., 1992. *J. Chemom*, 1992, (6), 267.
[14] Xu J., Zhang H., Wang Lei., Liang G., Wang Luoxin., Shen X., Xu W., *Spectrochimica Acta Part A*, 2010, (76), 239.
[15] Todeschini R., Maiocchi A., Consonni V., The K Correlation Index: Theory Development and its Application in Chemometrics. *Chemom,* Int. Lab. Syst, 1999, (46), 13.
[16] Golbraikh A, Tropsha A. J Comput Aided Mol Des 2002; 16: 357-369.
[17] Golbraikh A, Tropsha A. J Mol Graph Model 2002; 20: 269-276.
[18] Tropsha A., Gramatica P., Gombar V K., *QSAR Comb. Sc*i, 2003, (22), 69.
[19] Shen M., Béguin C., Golbraikh A., Stables J P., Kohn H., Tropsha A., *J. Med. Chem*, 2004, (47), 2356.
[20] Weisberg S., 2005. Applied Linear Regression, 3rd edn. (John Wiley and sons, Inc., New Jersey,)
[21] SCAN- Software for Chemometric Analysis- 1995. version 1.1- for Windows, Minitab USA.
[22] Katritzky, A. R., Dobchev, D. A., Tulp, I., Karelson, M., Carlson, D. A. Med. Chem. Lett. 2006, (16), 2306.
[23] A.T. Balaban, D. Ciubotariu, M. Medeleanu, J.Chem.Inf.Comput.Sci. 1991, (31), 517.
[24] P.Gramatica, Corradi M., Consonni V., Chemosphere 2000, (41), 763.
[25] Consonni V., Todeschini R., Pavan M., *J. Chem. Inf. Comput. Sci*, 2002, (42), 682.
[26] Consonni V., Todeschini R., Pavan M., Gramatica P., *J. Chem. Inf. Comput. Sci*, 2002, (42), 693.
[27] Todeschini R., Consonni V. , 2009. Molecular Descriptors for Chemoinformatics Volumes I & II. (WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim, 2009).
[28] Ramsey F. L., Schafer D. W., 2002. The Statistical Sleuth: A Course in Methods of Data Analysis, 2nd edn. (Wadsworth group, USA).
[29] Zheng F., Bayram E., Sumithran S P., Ayers J T., Zhen C G., Schmitt J D., Dwoskim L P., Crooks P A., Bioorg. Med. Chem, 2006, (14), 3017.
[30] Guha R., Jurs P C., *J. Chem. Inf. Model*, 2005, (45), 800.