وزارة التعليم العالى و البحث العلمى

Université BADJI Mokhtar – Annaba
BADJI Mokhtar – Annaba University

جـامـعـة بـاجـي مـخـتـار – عـنـابـة

Faculté des Sciences
Département de Chimie

MEMOIRE

Présenté pour l'obtention du diplôme MAGISTER

Par M^r^. HADDAG Hamza

DES en Chimie

Option : Chimie et environnement

THEME

Modélisation du facteur acentrique de plusieurs familles de composés

Devant le jury :

PRESIDENT :	**M^r^. D. MESSADI**	**Pr**	**UBMA**
EXAMINATEURS :	**M^me^. S. ALI-MOKHNACHE**	**Pr**	**UBMA**
	M^r^. A. DJALLAL	**MC**	**UBMA**
RAPPORTEUR :	**M^me^. Z. HABES**	**MC**	**UBMA**
Invité :	**M^r^. M.L. SOUICI**	**MC (B)**	**UBMA**

Année 2009

ملخص

التنبؤ بخواص المركبات الكيميائية أصبح أمرا ضروريا باستعمال مختلف العلاقات.

في هذا الإطار طورنا نماذج مختلفة للتنبؤ بقيم المعامل الغير مركزي نظرا لأهميته في الديناميكية الحرارية.

استعملنا لاختيار الواصفات الجزيئية ذات المعني, الخوارزمي الجيني و لحساب النماذج التراجع المتعدد الخطية أو الشبكة العصبية الاصطناعية.

الإحصاءات المتحصل عليها لمجموعتي المعايرة و التصديق كمعاملي :التنبؤ و التحديد المتعدد و كذلك جذور الأخطاء المربعة المتوسطة تثبت النوعية الجيدة للنماذج المدروسة مع ملاحظة الجودة الإضافية للنموذج العصبي.

الكلمات الجوهرية:
المعامل الغير مركزي, QSAR/QSPR, التراجع المتعدد الخطي, الشبكة العصبية الاصطناعية, الواصفات الجزيئية.

ABSTRACT

The empirical way to estimate properties will always impose itself and remain unavoidable when dealing with characterization of organic compounds.

In this context we developed different acentric factor's prediction models that had an unquestionable interest in thermodynamics.

The selection of significant molecular descriptors was done by genetic algorithm, and the models calculated using a multilinear regression or an artificial neural network.

The statistical parameters obtained for training end validation sets such as: prediction and multiple determination coefficients, roots of mean quadratic errors; are proofs of the quality and the pertinence of the models. We must remark here the superiority of the neural model.

Key-words:
Acentric factor, QSAR/QSPR, multilinear regression, artificial neural network, molecular descriptors.

RESUME

L'estimation empirique des propriétés s'imposera toujours et demeurera incontournable pour caractériser les produits organiques.

Dans cette optique nous avons développé différents modèles pour la prédiction du facteur acentrique qui revêt un intérêt certain en thermodynamique.

La sélection des descripteurs moléculaires significatifs a été faite par algorithme génétique, et les modèles établis par régression multilinéaire ou réseau de neurones artificiels.

Les paramètres statistiques obtenus pour les ensembles de calibration et de validation (coefficients de détermination multiple et de prédiction, les racines des erreurs quadratiques moyennes) font ressortir la qualité et la pertinence des modèles calculés. Signalons tout de même la supériorité du modèle neural.

Mots-clés:
Facteur acentrique, QSAR/QSPR, régression linéaire multiple, réseau de neurones artificiels, descripteurs moléculaires.

Remerciements

Cette étude a été réalisée au laboratoire de sécurité environnementale et alimentaire de l'université d'Annaba sous la direction de **M^r^ D. Messadi**, à qui j'exprime ma profonde reconnaissance pour l'intérêt qu'il a porté à ce travail et ses conseils éclairés tout au long de ces années.

Aussi, je tiens à remercier **M^me^ Z. Habes** pour son aide et les membres du jury pour avoir accepté de juger ce modeste travail.

Enfin je témoigne ma reconnaissance à toutes celles et ceux qui m'ont accompagné et aidé à l'accomplissement de ce travail.

Dédicace

Je dédie ce travail à :

- Mes parents
- Toute ma famille
- Mes amis
- Et toute l'équipe du 34.

PARTIE GENERALITES

PARTIE EXPERIMENTALE

LISTE DES TABLEAUX

LISTE DES FIGURES

SYMBOLES ET ABREVIATIONS

AM1 :	Austin Model 1.
EQM:	Erreur quadratique moyenne.
EQMC:	Ecart quadratique moyen calculé sur l'ensemble de calibration.
EQMP:	Ecart quadratique moyen de prédiction.
$EQMP_{ext.}$:	Ecart quadratique moyen calculé sur l'ensemble de validation externe
e_i :	Résidu ordinaire.
$e_{i\ std}$:	Résidu standardisé.
F :	Statistique de Fisher.
FIT:	Fonction de Kubinyi.
FIV:	Facteur d'inflation de la variance.
GA:	Algorithme génétique (Genetic Algorithm).
h_i :	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).
n_{ext} :	Dimension de l'ensemble de validation.
PRESS :	Somme des carrés des erreurs de prédiction.
P :	Pression.
P^*_{vp} :	Pression de vapeur réduite.
P_c :	Pression critique.
p :	Nombre de descripteurs dans le modèle.
QSAR :	Quantitative Structure/ Activity Relationships. (Relations Quantitatives Structure/ Activité).
QSPR :	Quantitative Structure/ Property Relationships. (Relations Quantitatives Structure/ Propriété).
Q^2_{LOO}	Coefficient de prédiction.
RLM:	Régression linéaire multiple.
RNA:	Réseaux de neurones artificiels.
R^2 :	Coefficient de détermination.
S :	Erreur standard.
SCE :	Somme des carrés des écarts.
SCT :	Somme des carrés totale.
T :	Température.
T^* :	Température réduite.
T_c:	Température critique.
y_i:	Valeur observée.
$\hat{y}_i$	Valeur estimée.
ω :	Facteur acentrique.
ω_{obs} :	Valeur observée du facteur acentrique.
ω_{cal} :	Valeur calculée du facteur acentrique (ensemble de calibration).
ω_{pred} :	Valeur prédite du facteur acentrique (ensemble de validation).

INTRODUCTION GENERALE

De très nombreux produits chimiques, ne sont pas caractérisés par leurs propriétés physico-chimiques ou leurs activités biologiques utiles pour contrôler leur évolution dans l'environnement ou leurs toxicités potentielles. Les raisons en sont :

- Leur grand nombre, plus de 120.000 actuellement.
- Coûts élevés des procédés expérimentaux de mesure.
- Indisponibilité des appareillages ou des réactifs nécessaires, ou leurs complexités dans le premier cas ou leur dangerosité dans le second.

Cette carence en données expérimentales a contraint les chercheurs à élaborer des modèles fiables afin de prédire rapidement et avec précision les propriétés ou activités manquantes des composés organiques d'intérêt. Les modèles QSP(A)R pour "Quantitative Structure Property (Activity) Relationship " sont devenus un outil indispensable à cet effet.

Le facteur acentrique (auquel on s'est intéressé dans ce travail) est d'une grande importance pour la caractérisation des substances pures, pour appliquer la loi des états correspondants, pour utiliser les équations d'état et pour déterminer les propriétés thermodynamiques. Depuis son introduction par Pitzer en 1957, les valeurs du facteur acentrique sont tabulées et présentes dans la littérature. Néanmoins certains composés organiques n'y figurent pas, ou cette littérature n'est elle-même pas aisément exploitable.

Nous suggérons dans ce travail d'appliquer la méthodologie QSPR pour la modélisation du facteur acentrique d'un mélange de composés organiques (halogénures, esters, éthers, cétones, alcools et phénols). Dans ce but nous avons choisi de recourir à des méthodes hybrides : (AG/RLM) algorithme génétique/ régression linéaire multiple et (AG/RNA) algorithme génétique/ réseau de neurones artificiels. Par algorithme génétique, nous avons procédé à la sélection d'un sous-ensemble significatif de descripteurs moléculaires parmi quelques 3.000 calculés par les logiciels spécialisés disponibles au laboratoire.

Notre mémoire comporte en plus de la bibliographie, de l'introduction et de la conclusion générale ; deux grandes parties :

- La première qui traite de généralités et dans laquelle nous avons, défini les modèles QSAR et QSPR, développé le processus menant au calcul des descripteurs moléculaires. Nous y avons aussi posé la base théorique nécessaire à ce travail tel que : Algorithmes génétiques, régression linéaire multiple, réseaux de neurones artificiels, traitement statistique pour l'évaluation de la qualité de l'ajustement (robustesse des modèles, validation externe …).
- Et une partie expérimentale où nous avons présenté et discuté les modèles obtenus.

PARTIE GENERALITES

I- LES MODELES QSAR/QSPR

Au cours des décennies passées, les Relations Quantitatives Structure- Activité/ Propriété (QSAR/QSPR) sont devenues un puissant outil théorique, alternatif à la mécanique quantique, pour la description et la prédiction des propriétés des systèmes moléculaires complexes dans différents environnements. L'approche QSAR/QSPR 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 [1]. Cette dernière peut être représentée par la composition chimique, la connectivité des atomes, la surface d'énergie potentielle, et la fonction d'onde électronique d'un composé. Différents descripteurs moléculaires physico- chimiques reflétant la structure peuvent être déterminés empiriquement ou en utilisant des méthodes théoriques et computationnelles de différentes complexités. Il est à souligner que la connaissance de la constitution chimique exacte et/ou de la structure moléculaire tridimensionnelle des composés chimiques étudiés est un pré- requis à l'application de l'approche QSAR/QSPR.

Le succès de l'approche QSAR/QSPR dépend de façon critique de la définition précise et de l'utilisation appropriée des descripteurs moléculaires. On distingue, arbitrairement, **les descripteurs moléculaires empiriques** des **descripteurs moléculaires théoriques**.

Les descripteurs empiriques peuvent être divisés en deux classes générales (tableau 1), la première reflète les interactions électroniques intramoléculaires **(descripteurs structurels)** alors que la seconde tient compte des interactions intermoléculaires dans les milieux condensés tels que les liquides et les solutions **(descripteurs de solvatation)**.

Tableau I-1 : Classification d'ensemble des descripteurs moléculaires empiriques

classe	Sous- classe
Descripteurs structurels	- Constantes d'induction - Constantes de résonance - Constantes stérique
Descripteurs de solvatation	- Echelles de polarité - Echelles de polarisabilité - Echelles d'acidité - Echelles de basicité - Echelles mixtes

Les descripteurs structurels les plus répandus ont été définis pour quantifier les propriétés d'induction, l'effet mésomère ou de résonance, et les effets stériques des composés chimiques. Les descripteurs de solvatation reflètent les interactions du soluté avec la masse du solvant environnant **(effets de solvant macroscopiques** ou **non spécifiques)**, et les liaisons spécifiques, souvent des liaisons hydrogène entre le soluté et les molécules individuelles de solvant **(effets de solvant spécifiques** ou **microscopiques)**. Les effets de solvant macroscopiques sont quantifiés en utilisant diverses échelles de polarité et de polarisabilité. Les descripteurs des effets de solvant microscopiques impliquent les échelles générales d'acidité et de basicité. Certaines échelles empiriques d'effets de solvant (échelles mixtes) peuvent impliquer en même temps ces deux effets macroscopique et microscopique. Le coefficient de partage octanol/ eau, log P, est le représentant typique de tels descripteurs.

Les descripteurs moléculaires théoriques peuvent, conventionnellement, être répartis en un certain nombre de classes, selon leur complexité ou leur méthode de calcul. Les descripteurs théoriques les plus simples sont des **descripteurs constitutionnels** qui peuvent être construits à partir de l'information sur la composition chimique du composé considéré. Les nombres, absolus et relatifs, des différents types d'atomes et de liaisons chimiques, la masse molaire, et le nombre de différents cycles dans le composé représentent quelques descripteurs constitutionnels typiques. **Les descripteurs**, **ou indices**, **topologiques** décrivent la connectivité des atomes dans la molécule. On a avancé [1] que les indices topologiques pouvaient encoder des interactions moléculaires subtiles et non pas seulement renseigner sur le degré de ramification des liaisons chimiques ou la distribution de la masse spécifique dans la molécule. **Les descripteurs géométriques** sont obtenus à partir de la structure tri-dimensionnelle des molécules définie par les coordonnées des noyaux atomiques et la grosseur de la molécule représentée, par exemple, par le rayon atomique de Van der Waals. Les molécules de la plupart des composés chimiques possèdent une certaine flexibilité conformationnelle et les surfaces de potentiels moléculaires respectives possèdent de multiples minima locaux. Selon la structure de la molécule, le nombre de ces minima peut être très grand et, par conséquent, il est plutôt difficile de trouver le minimum d'énergie global pour des conditions expérimentales établies.
Evidemment, les descripteurs géométriques peuvent varier de façon significative selon les conformations utilisées dans le calcul de ces descripteurs. Dans une certaine mesure, **les descripteurs théoriques liés à la distribution de charge** peuvent également dépendre de la conformation. Ces descripteurs sont basés sur la structure tri- dimensionnelle et la distribution des charges dans la molécule. Ces dernières peuvent se présenter comme charges atomiques partielles obtenues à partir d'un schéma empirique ou en utilisant des fonctions plus sophistiquées basées sur la fonction d'onde de la molécule calculée par la chimie quantique.

Un certain nombre de **descripteurs quanto- chimiques basés sur les OM** ont été employés dans le développement d'équations QSAR/QSPR. Les plus utilisés sont les énergies des OM frontières, c'est-à-dire, l'énergie calculée de la plus basse orbitale moléculaire inoccupée ($\varepsilon_{LUMO)}$, et l'énergie de la plus haute orbitale moléculaire occupée (ε_{HOMO}), et la

différence entre ces énergies. De même, différents indices de réactivité déduits de la théorie de la superdélocalisabilité de Fukui ou d'autres constructions théoriques ont gagné en popularité parmi les chercheurs.

Tous les descripteurs théoriques ne peuvent être strictement classés selon le schéma présenté dans le tableau 2. Par exemple, les indices topographiques sont déduits de l'information contenant à la fois la topologie et la géométrie des molécules. **Les indices électrotopologiques** sont fondés sur la topologie et la distribution de charge alors que les aires de surfaces partielles chargées sont des descripteurs qui encodent à la fois la distribution de charge et la géométrie des molécules. De tels descripteurs peuvent être classés comme **descripteurs moléculaires mixtes ou combinés**.

Les descripteurs moléculaires peuvent être définis pour tout le système moléculaire étudié ou pour n'importe laquelle de ses parties (fragments). Par exemple, la majorité des descripteurs empiriques structurels sont reliés à des fragments moléculaires appelés substituants. En conséquence, les molécules d'une série congénère de composés chimiques sont divisées formellement en deux ou plusieurs fragments qui correspondent à une unité structurale constante Y (c'est-à-dire le centre de réaction) et à des unités structurales variables Xi (les substituants). Les relations QSAR/QSPR sont ainsi présentées comme suit :

$$P = P_0^{(Y)} + \sum_i \sum_k \alpha_{ik}^{(Y)} D_{ik}^{(X)} \qquad \text{(I-1)}$$

Où $P_0^{(Y)}$ est l'ordonnée à l'origine correspondant au fragment moléculaire constant Y, les $D_{ik}^{(X)}$ sont les descripteurs moléculaires de type k pour les fragments variables X_i, et les $\alpha_{ik}^{(Y)}$ sont les coefficients de développement caractéristiques d'une série donnée de composés $X_i Y$.

Tableau I-2 : Classification générale des descripteurs moléculaires théoriques

classe	Sous- classe
Descripteurs constitutionnels	- Dénombrement des atomes ou des liaisons. - Descripteurs basés sur les masses atomiques.
Descripteurs topologiques	- Indices topologiques (connectivité). - Descripteurs théoriques d'information. - Descripteurs topochimiques.
Descripteurs géométriques	- Descripteurs liés à la distance. - Descripteurs liés à l'aire de la surface. - Descripteurs liés au volume. - Descripteurs du champ stérique moléculaire.
Descripteurs liés à la distribution de charge	- Charges atomiques partielles. - Moments électriques moléculaires - Polarisabilités moléculaires. - Descripteurs du champ électrique moléculaire.
Descripteurs liés aux orbitales moléculaires	- Energie des OM frontières - Ordres de liaison - Indices de réactivité de Fukui.
Descripteurs température dépendants	- Fonctions thermodynamiques. - Descripteurs facteurs de Boltzmann pondérés.
Descripteurs de solvatation	- Energie électrostatique de solvatation. - Energie de dispersion de solvatation. - Enthalpie libre de formation de cavité. - Descripteurs de liaison hydrogène. - Entropie de solvatation. - Descripteurs d'énergie de solvatation linéaire théorique.
Descripteurs mixtes	- Descripteurs topographiques. - Descripteurs électrotopologiques. - Descripteurs de la charge partielle de l'aire de la surface.

La plupart des descripteurs théoriques qui apparaissent dans le tableau 2 peuvent être calculés soit pour la molécule entière soit pour un fragment moléculaire pré- défini.

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

II-1- Introduction

L'application pratique des gammes des descripteurs moléculaires dans le développement de modèles QSAR/QSPR n'est pas une tâche aisée [1]. Tout d'abord, un très grand nombre (>3000) 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) et d'analyse en composantes principales (ACP) est la technique de régression par les moindres carrés partiels (MCP ou PLS) [2-7].

On a également appliqué les méthodes modernes de l'intelligence artificielle au développement de modèles QSAR/QSPR [8-12]. 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.

II-2- Méthodes de régressions linéaire et multilinéaire

II-2-1 Aperçu général

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

a) Un nombre maximal de données expérimentales (de préférence toutes) doivent être caractérisées par des valeurs de descripteurs originaux complémentaires.
b) Les valeurs des descripteurs doivent être obtenues de la même source et, de préférence, mesurées selon le même protocole expérimental ou calculées en utilisant le même logiciel.
c) Le nombre de descripteurs dans les modèles de régression multiples doit être minimisé, sans perte d'information, ce que mettent en évidence les critères statistiques (valeurs des tests t et F…).
d) Dans les modèles RLM, les descripteurs utilisés doivent être statistiquement orthogonaux.
e) Pourvu que les autres critères soient similaires, la nature physique ou chimique du descripteur sélectionné doit être la plus proche de la propriété ou du phénomène étudié.

En réalité, il est difficile de se conformer pratiquement aux 5 principes énoncés. Cependant, la négligence de plusieurs d'entre eux peut conduire à des équations inutiles sans aucun pouvoir prédictif sinon très limité.

II-2-2- Evaluation préliminaire des données

Avant d'entamer le développement effectif des équations de régression QSAR et QSPR, il est hautement recommandé d'examiner la qualité statistique des données de départ, à la fois les données à corréler (variable dépendante) et les descripteurs utilisés dans la corrélation (variables indépendantes).

On distingue habituellement dans un tel pré- traitement des données les analyses univariées des analyses bivariées [13-18].

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

Pour un ensemble de descripteurs différents, il est nécessaire d'effectuer une analyse des données bivariée, c'est-à-dire de calculer le coefficient de corrélation linéaire R entre chacune des paires de l'ensemble des descripteurs. Si R est statistiquement significatif (R>0,95), ces deux descripteurs ne peuvent être utilisés simultanément lors de l'analyse par RLM.

II-2-3- Régression linéaire multiple

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

$$yi= \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} + \varepsilon \qquad \text{(I-2)}$$

ou les y_i, x_{i1}, x_{i2},…, x_{ip} sont des données respectivement relatives aux variables Y, X_1,……., X_p.

Les estimateurs β_j sont calculés en utilisant la méthode des moindres carrés ordinaires. Les variables aléatoires ε_i représentent les termes d'erreur non observables du modèle. On peut estimer ces erreurs par les résidus ordinaires e_i, différence entre les valeurs observées y_i et les valeurs estimées $\hat{y}_i$.

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

a) Les résidus (**e**) ont une espérance mathématique nulle :

$$E(\mathbf{e}) = \mathbf{0} \qquad \text{(I-3)}$$

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

c) Les résidus sont indépendants entre eux :

$$E(e_i, e_j) = 0 \qquad \text{si } i \neq j \qquad \text{(I-4)}$$

leurs covariances sont nulles.

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

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

e) Les résidus suivent une distribution normale (de Laplace- Gauss).

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

Pour minimiser l'influence des erreurs de détermination des valeurs explicatives (ou régresseurs) sur la précision des résultats de la régression 4 à 5 données (variables dépendantes, ou encore observations) doivent, à la limite, être associées à chaque variable explicative. Le nombre de degré de liberté final (n-p-1) doit être [19] tel que :

$$n- p- 1 \geq 10 \qquad \text{(I-5)}$$

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

Pour les modèles à plus de deux descripteurs, de faibles coefficients de corrélation croisés n'assurent pas forcément l'orthogonalité des descripteurs. Une indépendance globale acceptable des descripteurs sera vérifiée lorsque les facteurs d'inflation de la variance (FIV) calculés pour chacun d'eux obéissent [19] à la condition FIV< 5.

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

- Le coefficient de détermination multiple :

$$R^2 = 1 - \frac{SCE}{SCT} = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} \qquad \text{(I-6)}$$

Où $\bar{y}$ est la valeur moyenne des valeurs observées pour l'ensemble de calibrage.

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

$$\sigma_N = \mathrm{EQMC} = \sqrt{\frac{\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2}{n}} \qquad \text{(I-7)}$$

Il est intéressant de considérer, également, la racine de l'écart quadratique moyen de prédiction (EQMP), et celle calculée sur l'ensemble de validation externe (EQMPext) :

$$\mathrm{EQMP} = \sqrt{\frac{1}{n}\sum_{1}^{n}\left(y_i - \hat{y}_{(i)}\right)^2} = \sqrt{\frac{\mathrm{PRESS}}{n}} \qquad \text{(I-8)}$$

$$\mathrm{EQMP}_{ext} = \sqrt{\frac{\sum_{i=1}^{n_{ext}}\left(y_i - \hat{y}_i\right)^2}{n_{ext}}} \qquad \text{(I-9)}$$

La validation croisée par « leave – one - out » (LOO) [20] consiste à recalculer le modèle sur (n-1) observations, et à utiliser le modèle ainsi obtenu pour calculer la grandeur d'intérêt du composé écarté, notée $\hat{y}_{(i)}$. On répète le procédé pour chacune des grandeurs d'intérêt. La somme des carrés des erreurs de prédiction, désignée par le symbole PRESS (eq. (8)), est une mesure de la dispersion des estimations. On l'utilise pour définir le coefficient de prédiction [20] :

$$Q^2_{\mathrm{LOO}} = \frac{\mathrm{SCT} - \mathrm{PRESS}}{\mathrm{SCT}} \qquad \text{(I-10)}$$

Contrairement à R^2 qui augmente avec le nombre de paramètres du modèle, le facteur Q^2_{LOO} affiche une courbe avec maximum (ou avec palier) obtenu pour un certain nombre de descripteurs, puis décroît de façon monotone. Ce fait confère une grande importance au coefficient Q^2_{LOO}. Une valeur de $Q^2_{\mathrm{LOO}} > 0{,}5$ est considérée comme satisfaisante, et une valeur supérieure à 0,9 est excellente [21].

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

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

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

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

Pour une grande valeur de Q^2_{LOO}, une valeur élevée de Q^2_{ext} permet de présager d'une bonne capacité prédictive du modèle.

II-3- Sélection d'un sous-ensemble de variables par algorithme génétique (GA- VSS)

Des logiciels informatiques spécialisés permettent le calcul de plus de 3000 descripteurs moléculaires. L'objectif étant de construire, sur un nombre minimal de descripteurs significatifs, des modèles conduisant à de faibles erreurs, il est important de réduire le pool de descripteurs disponibles, de façon à ne retenir que ceux encodant l'information la plus riche. Parmi les stratégies de sélection des variables explicatives on peut citer : les méthodes de pas à pas, ainsi que les algorithmes évolutionnaires et génétiques [22, 23] ; la comparaison se fait souvent à l'avantage des algorithmes génétiques [24].

La modélisation de processus génétiques a initié le développement des algorithmes génétiques, qui peuvent être exploités dans une grande variété de problèmes d'optimisation [24].

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 [25], en maximisant Q^2_{LOO}. Le nombre

de descripteurs est fixé par la valeur optimale de la fonction FIT de Kubinyi [36], calculée selon :

$$FIT = \frac{R^2}{1 - R^2} \frac{n - p - 1}{(n + p)^2} \qquad \text{(I-12)}$$

Ce critère permet de comparer des modèles construits sur n données avec des nombres de variables p différents ; R^2 est le coefficient de détermination.

II-4- Modèles QSAR / QSPR non linéaires- Réseaux de neurones artificiels

Les modèles QSAR /QSPR intrinsèquement non linéaires peuvent être développés en utilisant les réseaux de neurones artificiels [27-29].

Un réseau de neurones artificiels (RNA) est un programme informatique conçu pour apprendre à partir de données qui lui sont présentées, en s'inspirant du schéma d'apprentissage effectué par le cerveau humain où le neurone est l'unité fonctionnelle de base du système nerveux. Les RNA représentent un moyen puissant pour le développement des relations non linéaires entre variables, ce qui en fait d'excellents outils de prédiction dans différents domaines.

Les réseaux multicouches, qui utilisent l'apprentissage supervisé, sont les plus puissants réseaux de neurones ; ils comportent en plus de l'entrée et de la couche de sortie, de une à plusieurs couches cachées (figure 1). Chaque neurone est uniquement relié à tous les

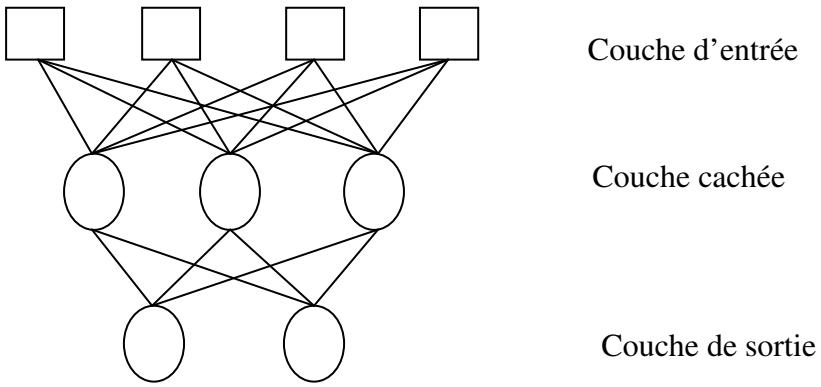


Figure I-1 : *Réseau de neurones avec 4 entrées, une couche cachée avec 3 neurones et une couche de sortie comprenant 2 neurones.*

neurones de la couche suivante, chaque connexion étant caractérisée par un poids. La façon dont chaque neurone transforme son entrée dépend du poids et du biais qui lui sont associés, ces 2 paramètres étant modifiables. L'apprentissage d'un RNA peut être réalisé à l'aide de l'algorithme de rétropropagation. Dans ce but, les valeurs d'entrée sont présentées après une

transformation éventuelle, au réseau qui les propage jusqu'à la couche de sortie et donne ainsi la réponse au réseau. Dans une deuxième étape les bonnes sorties correspondantes sont représentées aux neurones de la couche de sortie qui calculent l'écart, modifient leur poids et leurs biais, et rétropropagent l'erreur jusqu'aux entrées pour permettre aux neurones cachés de modifier leurs poids et leurs biais de la même façon. Ce processus itératif est stoppé en utilisant comme critère « l'arrêt précorce », c'est-à-dire dès que l'indice de performance (erreur quadratique moyenne : EQM) calculé sur les données de test cessent de s'améliorer.

Dans la plupart des applications des RNA à la chimie l'utilisation d'une seule couche cachée semble suffire [30]. Nous avons donc utilisé dans ce travail un réseau standard à 3 couches comprenant l'entrée, la sortie et une couche cachée. L'algorithme de Levenberg-Marquardt conçu pour faciliter certains problèmes de convergence est l'un des plus utilisés pour l'apprentissage des réseaux, d'autant plus qu'il s'adapte très bien avec le choix de l'erreur quadratique moyenne comme indice de performance. Nous avons donc utilisé l'algorithme Levenberg- Marquardt de rétropopagation (fonction TRAINLM de la boîte à outils du logiciel MATLAB 7.0 [31]) pour l'apprentissage du réseau. Les fonctions de transfert sigmoïde (tangente hyperbolique) et linéaire ont été adoptées comme fonctions d'activation, respectivement pour les couches cachée et de sortie.

III- CALCUL DES DESCRIPTEURS MOLECULAIRES

La représentation numérique de la structure chimique (descripteurs moléculaires) est une étape importante de l'investigation QSPR. Les performances du modèle élaboré et la précision des résultats sont étroitement liées au mode de détermination de ces descripteurs.

Nous avons utilisé le logiciel de modélisation moléculaire Hyperchem 6.03 [32] pour représenter les molécules puis, à l'aide de la méthode semi-empirique AM1, obtenir les géométries finales. Tous les calculs ont été menés dans le cadre du formalisme RHF (pour restricted Hartree-Fock ou formalisme de Hartree-Fock avec contrainte de spin) sans interaction de configuration. Les structures moléculaires ont été optimisées à l'aide de l'algorithme Polak- Ribiere avec pour critère une racine du carré moyen du gradient égale à 0,01 kcal/mol. Les géométries ainsi optimisées ont été transférées dans le logiciel informatique DRAGON [33] pour le calcul de plus de 1600 descripteurs (si l'on tient compte de ceux calculés à l'aide du logiciel Hyperchem) appartenant à 20 classes différentes.

En utilisant les options correspondantes du logiciel DRAGON, nous avons d'abord éliminé les descripteurs à valeurs constantes (écarts types inférieurs à 0,001) qui n'apportent aucune information, ensuite ceux qui sont hautement corrélés ($R \geq 0,95$) et qui véhiculent une information redondante. Pour chaque paire de descripteurs corrélés, est éliminé automatiquement celui qui présente les plus hautes corrélations croisées avec les autres descripteurs.

PARTIE EXPERIMENTALE

I- Définition du facteur acentrique :

Le facteur acentrique [34,35] est l'une des constantes des corps purs les plus courantes. Comme proposé à l'origine par Pitzer, le facteur acentrique ω représente l'acentricité ou la non sphéricité d'une molécule. De ce fait, le facteur acentrique est très utilisé pour la détermination des propriétés thermodynamiques des substances comme le facteur de compressibilité [36-41], les équilibres de phases des substances [42,43], les pressions de vapeur [44,45] , l'enthalpie de vaporisation [35,46] et les coefficients de l'équation du viriel [47,48].

En général, les valeurs du facteur acentrique sont déterminées à partir des données expérimentales de la pression de vapeur et des paramètres des points critiques.

On utilise [49] l'équation:

$$\omega = -\log P^*_{vp} - 1 \qquad \text{(I-1)}$$

dans laquelle les différents paramètres ont la signification suivante :

P^*_{vp} : pression de vapeur réduite $P^*=P/P_c$ pour $T^*= 0,7$.

P_c : pression critique.

P_{vp} : pression de vapeur.

T^* : température réduite $T^*=T/T_c$; pour le calcul du facteur acentrique elle doit être égale à $0,7$.

T : température en kelvin.

T_c: température critique.

Pour l'utilisation de cette équation, on doit étudier les substances dans l'état critique; ce qui n'est pas toujours aisé. Les méthodes d'estimation du facteur acentrique proposées, comme la méthode de contribution de groupes, donnent des valeurs exactes mais sont, ou très compliquées et utilisent des valeurs elles-mêmes estimées ou applicables à des domaines réduits. La méthode de contribution de groupes, par exemple, n'est applicable qu'aux hydrocarbures saturés.

Toutes les contraintes rencontrées dans le calcul, ou l'estimation du facteur acentrique militent pour la recherche d'un modèle QSPR fiable pouvant être exploité dans la pratique.

II- Données expérimentales :

Les valeurs du facteur acentrique exploitées dans le présent travail ont été prélevées dans l'annexe A de l'ouvrage, '' ***The properties of gases and liquids*** '' [49].On a divisé les 119 composés ainsi prélevés en trois groupes :

- Les alcools et les phénols au nombre de 18 forment le 1er groupe.
- Les halogénures d'alcanes, d'alcènes et de benzène forment le 2ème groupe qui compte 64 composés.
- Le reste qui est un mélange de 37 composés, comprenant des cétones, des éthers et des esters, forme le 3ème groupe.

Un modèle QSPR a été développé pour chacun de ces 3 groupes pris séparément, puis un modèle général a été recherché pour l'ensemble des composés. Dans le premier cas des corrélations linéaires entre grandeurs dépendantes et variables explicatives ont été imposées, alors que dans le second des corrélations linéaires et non linéaires ont été testées.

La liste des composés et leurs numérations idoines ont été reportées en annexe.

III- Modélisation du premier groupe :

III-1- Calcul du modèle :

Vu le nombre réduit des données pour ce groupe nous n'avons pas cherché à le décomposer en un ensemble de calibration (pour le calcul du modèle) et un ensemble de test (pour la validation statistique externe).

Parmi les modèles obtenus par algorithme génétique, notre choix s'est porté sur le premier caractérisé par les valeurs de Q^2= 88,15 % et R^2= 91,53 %. Les descripteurs entrant dans le modèle, leurs classes et deux brèves définitions sont donnés dans le tableau III-1.

Tableau III-1 : Descripteurs moléculaires du modèle choisi:cas des alcools et des phénols.

N°	Descripteur	Classe	Définition
1	MATS6p	Indice d'autocorrélation 2D	Indice d'autocorrélation de Moran
2	HATS3e	Descripteurs GETAWAY	Autocorrélation influence- pondérée de rang 3

Dans le tableau III-2 on trouve les valeurs du coefficient de corrélation pour les deux descripteurs et le facteur acentrique ω.

Tableau III-2 : Matrice de corrélation

	ω	MATS6p
MATS6p	-0,83 0	
HATS3e	0,844 0	-0,53 0,024

III-2- Equation et analyse de régression :

L'équation du modèle calculé est la suivante :

$$\omega = 0{,}48786\ (\pm 0{,}01808) - 0{,}09988\ (\pm 0{,}01666)\ \mathrm{MATS6p} + 0{,}09392\ (\pm 0{,}01481)\ \mathrm{HATS3e}$$

$$n = 18\ ;\quad R^2 = 91{,}53\ \%\ ;\quad Q^2 = 88{,}15\ \%\ ;\quad F = 81{,}04$$

$$\mathrm{EQMP} = 0{,}0226\ ;\ \mathrm{EQMC} = 0{,}0191 \qquad \text{(III-1)}$$

Les paramètres statistiques reproduits ci-dessus montrent une bonne explication de la variabilité de ω, par les descripteurs choisis, de l'ordre de 91,53% pour le coefficient de détermination, et une robustesse du modèle due à la grande valeur de Q^2 (proche de 90%).Les valeurs des écarts quadratiques moyens de prédiction (EQMP) et de calcul (EQMC) sont proches et faibles. La grande valeur du paramètre de Fisher montre un modèle hautement significatif.

III-3- Analyse des résidus et diagnostics d'influence :

Le calcul effectué par le logiciel MobyDigs entre autres des résidus ordinaires, e_i , et standardisés, que nous noterons $e_{i\ std}$, donne les résultats qui apparaissent respectivement dans les colonnes 3 et 4 du tableau III-3.

Tableau III-3 : Valeurs des ω observés (ω_{obs}) et calculés (ω_{cal}), des résidus ordinaires (e_i) et standardisés ($e_{i\ std}$) ainsi que des leviers (h_i).

Composé	ω_{obs}	ω_{cal}	e_i	$e_{i\ std}$	h_i
1	0,454	0,4458	-0,0082	-0,5812	0,229
2	0,505	0,5133	0,0083	0,5647	0,211
3	0,433	0,4434	0,0104	0,7436	0,235
4	0,438	0,4443	0,0063	0,4692	0,26
5	0,502	0,494	-0,008	-0,5807	0,241
6	0,587	0,5594	-0,0276	-1,779	0,181
7	0,56	0,5668	0,0068	0,4195	0,16
8	0,528	0,5563	0,0283	1,7434	0,155
9	0,56	0,5792	0,0192	1,144	0,137
10	0,579	0,5602	-0,0188	-0,9779	0,056
11	0,593	0,5649	-0,0281	-1,6144	0,116
12	0,592	0,5896	-0,0024	-0,1326	0,087
13	0,577	0,601	0,024	1,3155	0,087
14	0,612	0,593	-0,019	-1,0401	0,086
15	0,623	0,6331	0,0101	0,6287	0,164
16	0,665	0,6276	-0,0374	-2,2487	0,143
17	0,644	0,6628	0,0188	1,6775	0,34
18	0,556	0,5734	0,0174	0,9973	0,114

Tous les résidus ordinaires (colonne3) sont compris dans l'intervalle ±3S, S étant l'erreur standard, c'est-à-dire ± 3× 0,0209= ± 0,0627.

Les valeurs des résidus standardisés (colonne 4) sont toutes comprises dans les bondes ±3.

Les valeurs de h_i , reproduites dans la colonne 5, étant toutes inférieures à la valeur critique $h^* = \frac{3\times(p+1)}{n} = \frac{3\times(2+1)}{18} = 0,5$, ne mettent pas en évidence de point levier.

La figure III-1 qui représente la variation des résidus standardisés $e_{i\ std}$, en fonction des leviers h_i des composés ne laisse pas apparaître de points aberrants.

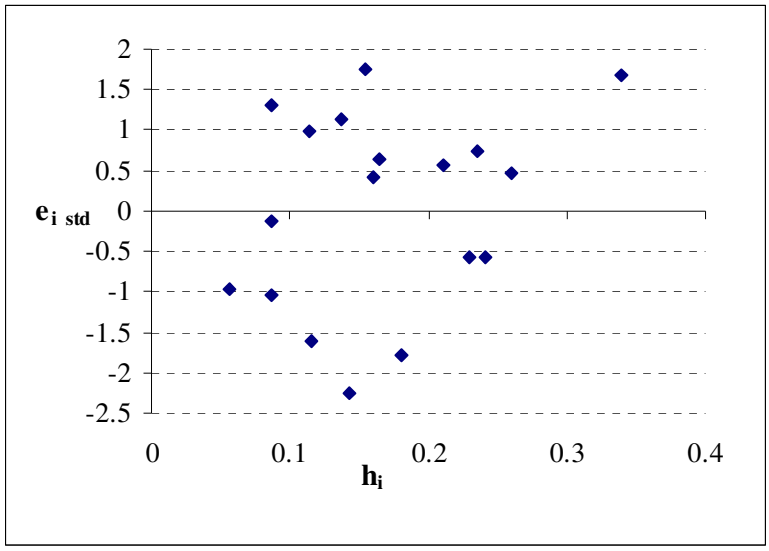


Figure III-1 : *Diagramme de Williams pour les composés du 1er groupe (alcools et phénols).*

III-4- Qualité de l'ajustement :

La qualité de l'ajustement à été vérifiée en représentant les valeurs observées ou expérimentales ω_{obs} (colonne1 tableau III-3) en fonction des valeurs calculées ω_{cal} (colonne 2 tableau III-3) par notre modèle. La figue III-2 (page suivante) montre une faible dispersion autour de la droite d'ajustement (qui peut être assimilée à la première bissectrice) définie par l'équation (III-2).

$$\omega_{obs} = -0.0000000 + 1\ \omega_{cal}$$
$$S = 0.0202736 \qquad R^2 = 91.5\ \% \qquad R^2_{ajust} = 91.0\ \% \qquad \text{(III-2)}$$

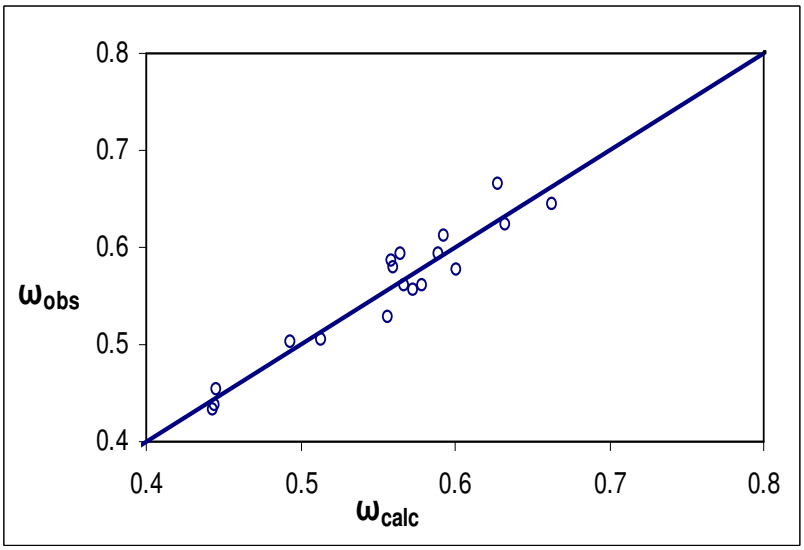


Figure III-2 : *Droite d'ajustement des* ω *observés en fonction des* ω *calculés: cas des alcools et des phénols.*

Afin de nous assurer que notre modèle n'est pas dû au hasard on a procédé au test de randomisation de y. La figure III-3 fait apparaître clairement que les valeurs de R^2, pour les modèles dont on a modifié le vecteur ‘‘facteur acentrique’’, sont très inférieures à celle du modèle réel (cercle plein), de plus on retrouve des valeurs négatives de Q^2 pour les modèles randomisés (cercles vides).Ceci nous confirme que notre modèle relie réellement le facteur acentrique aux deux variables explicatives choisies et qu'il n'est pas fortuit.

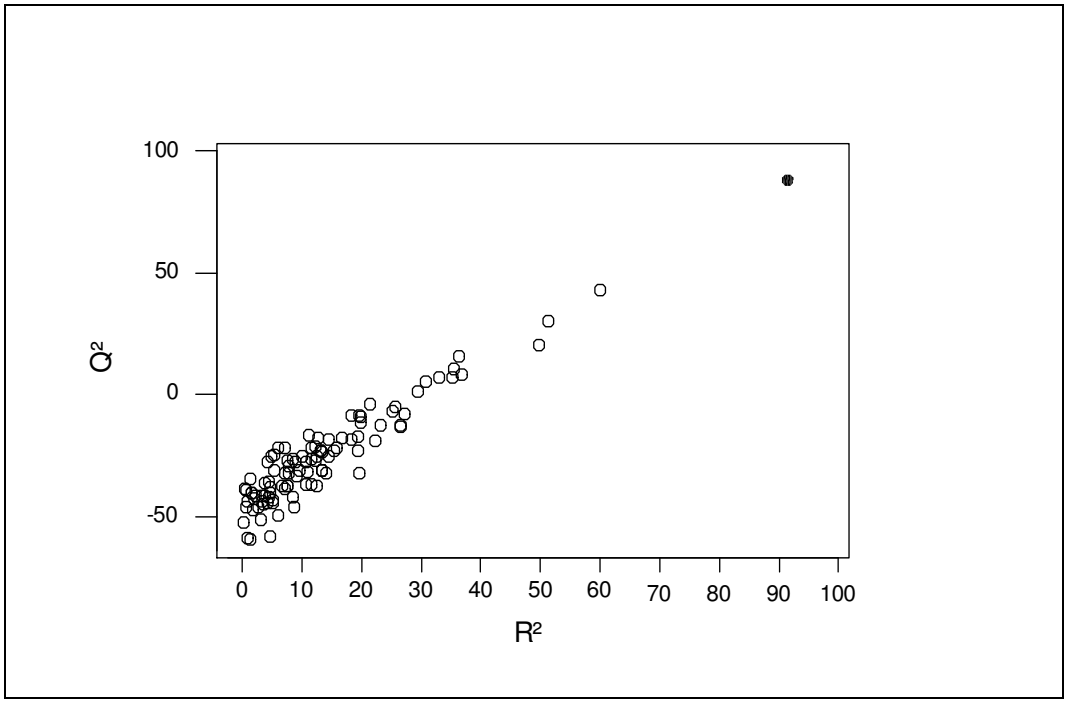


Figure III-3 : *Test de randomisation: cas des alcools et des phénols.*

IV- Modélisation du deuxième groupe :

Afin de procéder à une modélisation complète, nous devons valider le ou les modèle (s) obtenu (s).On a donc scindé notre ensemble de 64 observations en deux sous-groupes, un de calibration (à 51 composés) et un autre (à 13 composés) pour la validation statistique externe. Ce choix a été fait aléatoirement en prenant garde à ne pas répéter les mêmes valeurs du facteur acentrique dans le même sous-ensemble.

Le fractionnement ainsi fait, on a élaboré notre modèle (choix de sa dimension et des descripteurs) sur l'ensemble de calibration. Puis on s'est assuré de sa validité sur le second ensemble.

IV-1- Calcul du modèle :

Nous avons fixé la dimension du modèle à 4 descripteurs pour les halogénures ; bien que la valeur maximale de la fonction FIT de Kubinyi soit de 5 (Figure IV-1).On a opté pour cette taille pour les raison suivantes :

- La plus grande progression du FIT s'opère lorsque le modèle passe de 3 à 4 descripteurs.
- Les modèles à 5 descripteurs sont caractérisés par une mauvaise validation externe (valeurs de Q^2_{ext} nulles ou très petites), ils ne sont donc pas généralisables.

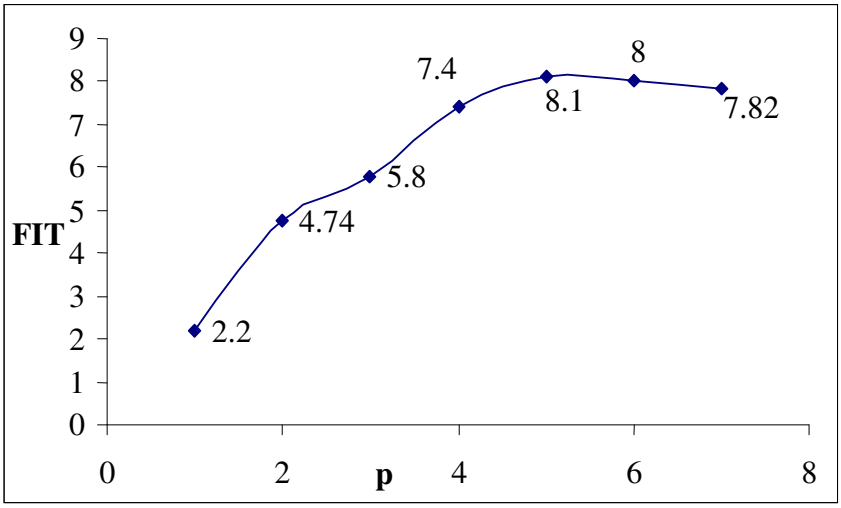


Figure IV-1 : *Variation du FIT en fonction du nombre de descripteurs.*

Après le choix de la dimension du modèle, on a procédé à la recherche des descripteurs qui expliquent au mieux la variation de la propriété à modéliser ou variable à expliquer. On a donc maximisé le coefficient de prédiction interne Q^2_{loo} .

Parmi les nombreux modèles obtenus, un s'est démarqué par les plus grandes valeurs de Q^2, R^2 et Q^2_{ext} ; les descripteurs dont il est composé sont dans le tableau IV-1. Ce tableau comporte aussi une brève définition des descripteurs ainsi que les blocs ou classes auxquels ils appartiennent.

Tableau IV-1 : Descripteurs du modèle pour le 2ème groupe constitué d'halogénures.

N°	Descripteur	Classe	Définition
1	Log P	Descripteur empirique	Coefficient de partage octanol/eau
2	SNar	Descripteur topologique	Indice simple topologique de Narumi
3	PCR	Comptes de parcours et de chemin	Ratio de parcours multiple
4	Mor23m	Descripteur MoRSE-3D	MoRSE-3D signal-32 pondéré par les masses atomiques

Le tableau IV-2 reproduit la matrice de corrélation.

Tableau IV-2 : Corrélations entre les paires de variables.

	ω	LogP	SNar	PCR
LogP	0,019			
	0,896			
SNar	0,799	0,438		
	0	0,001		
PCR	0,426	0,157	0,732	
	0,002	0,27	0	
Mor23m	-0,18	-0,283	-0,369	-0,088
	0,206	0,044	0,008	0,539

IV-2- Equation et analyse de régression :

Le modèle obtenu a pour équation :

$$\omega = 0{,}597(\pm 0{,}04344) - 0{,}0439\ (\pm 0{,}004209)\ \text{Log P} + 0{,}0674\ (\pm 0{,}003324)\ \text{SNar} - 0{,}392\ (\pm 0{,}04352)\ \text{PCR} + 0{,}0384\ (\pm 0{,}009614)\ \text{Mor23m}$$

$$n=51\ ;\ R^2= 92{,}05\ \%\ ;\ Q^2= 91{,}08\ \%\ ;\ F= 133{,}1901$$
$$EQMP= 0{,}0236\ ;\ EQMC= 0{,}0223. \qquad \text{(IV-1)}$$

Le facteur acentrique pour les 51 composés utilisés pour la calibration (élaboration du modèle) est bien corrélé avec les quatre descripteurs d'où la grande valeur du coefficient de détermination R^2. Notre modèle a de très bonnes capacités prédictives confirmées par la valeur de Q^2 qui est supérieure à 90%. La statistique de Fisher montre que notre modèle est très significatif. Les écarts quadratiques (EQMP/C) sont faibles et proches.

IV-3- Analyse des résidus et diagnostics d'influence :

Les valeurs des résidus, ordinaires et standardisés, ainsi que les valeurs de h_i sont présentées dans le tableau IV-3, dont les colonnes 1 et 2 reproduisent les valeurs expérimentales et calculées du facteur acentrique des composés considérés.

Tableau IV-3 : Valeurs des ω observés (ω_{obs}) et calculés (ω_{cal}), des résidus ordinaires (e_i) et standardisés ($e_{i\ std}$) ainsi que des leviers (h_i).

Composé	ω_{obs}	ω_{cal}	e_i	$e_{i\ std}$	h_i
1	0,177	0,1914	0,0144	0,6536	0,042
2	0,26	0,2224	-0,0376	-1,7553	0,059
3	0,198	0,1922	-0,0058	-0,2643	0,046
4	0,187	0,1885	0,0015	0,0768	0,102
5	0,171	0,1855	0,0145	0,7622	0,131
6	0,221	0,2289	0,0079	0,359	0,039
7	0,204	0,1916	-0,0124	-0,6138	0,095
8	0,153	0,169	0,016	0,7623	0,07
9	0,184	0,1832	-0,0008	-0,0359	0,074
10	0,189	0,1909	0,0019	0,1086	0,184
11	0,199	0,209	0,01	0,4575	0,046
12	0,193	0,1915	-0,0015	-0,1174	0,327
13	0,251	0,2261	-0,0249	-1,1204	0,036
14	0,279	0,284	0,005	0,2289	0,045
15	0,215	0,2214	0,0064	0,2999	0,065
16	0,263	0,2797	0,0167	0,7704	0,051
17	0,191	0,1998	0,0088	0,398	0,042
18	0,245	0,2549	0,0099	0,492	0,095
19	0,256	0,2751	0,0191	0,8952	0,062
20	0,24	0,2349	-0,0051	-0,2341	0,044
21	0,278	0,2354	-0,0426	-1,9145	0,036
22	0,248	0,2414	-0,0066	-0,346	0,126
23	0,325	0,3243	-0,0007	-0,0384	0,136
24	0,308	0,2912	-0,0168	-0,7706	0,047
25	0,235	0,2254	-0,0096	-0,4273	0,029
26	0,31	0,2623	-0,0477	-2,1525	0,037
27	0,374	0,371	-0,003	-0,2065	0,269
28	0,19	0,2232	0,0332	1,4746	0,028
29	0,3	0,2345	-0,0655	-2,9225	0,031
30	0,218	0,2537	0,0357	1,5975	0,033
31	0,14	0,1843	0,0443	2,081	0,063
32	0,223	0,2489	0,0259	1,1592	0,034

Tableau IV-3: suite et fin

Composé	ω_{obs}	ω_{cal}	e_i	$e_{i\ std}$	h_i
33	0,252	0,2413	-0,0107	-0,4766	0,027
34	0,22	0,2217	0,0017	0,0777	0,06
35	0,213	0,2119	-0,0011	-0,0522	0,05
36	0,238	0,226	-0,012	-0,5319	0,025
37	0,13	0,1779	0,0479	2,2485	0,063
38	0,346	0,3509	0,0049	0,2474	0,102
39	0,373	0,38	0,007	0,3526	0,108
40	0,299	0,3048	0,0058	0,2922	0,106
41	0,249	0,2485	-0,0005	-0,026	0,174
42	0,355	0,3726	0,0176	0,9413	0,14
43	0,251	0,2308	-0,0202	-1,1945	0,197
44	0,622	0,6204	-0,0016	-0,6761	0,782
45	0,272	0,2541	-0,0179	-1,0505	0,192
46	0,4	0,3961	-0,0039	-0,2049	0,137
47	0,217	0,2293	0,0123	0,5987	0,085
48	0,21	0,2295	0,0195	0,8998	0,051
49	0,232	0,2004	-0,0316	-1,5538	0,091
50	0,246	0,2758	0,0298	1,3654	0,046
51	0,229	0,1914	-0,0376	-1,7202	0,046

Tous les résidus ordinaires (e_i , colonne 3) sont inférieurs, en valeur absolue, à 3 fois l'erreur standard, soit 3S= 0,0704.

IV-4- Validation externe :

Pour vérifier les capacités prédictives de notre modèle on a eu recours à sa validation sur un ensemble prévu à cet effet et choisi dès le départ. Cet ensemble de validation, qui n'a pas servi à l'élaboration du modèle, est constitué des composés numérotés de 52 à 64 (tableau IV-4).

Tableau IV-4 : Quelques caractéristiques des éléments de l'ensemble de validation externe pour les halogénures.

Composé	ω_{obs}	ω_{pred}	$e_{i\ std}$	h_i
52	0,396	0,4087	0,5808	0,127
53	0,344	0,3492	0,2318	0,102
54	0,251	0,2302	-0,9028	0,032
55	0,256	0,2562	0,0109	0,078
56	0,24	0,2367	-0,1421	0,028
57	0,232	0,2084	-1,0194	0,031
58	0,218	0,225	0,3088	0,06
59	0,355	0,3508	-0,1893	0,101
60	0,249	0,2748	1,1982	0,156
61	0,244	0,2643	0,9428	0,162
62	0,281	0,2961	0,6585	0,047
63	0,271	0,2268	-1,968	0,083
64	0,157	0,1647	0,3538	0,137

La valeur de $Q^2_{ext} = 91,28$ % nous renseigne sur la validité du modèle et sa capacité à prédire des valeurs qui n'ont pas servi à le générer. L'écart quadratique moyen de prédiction externe $EQMP_{ext} = 0,019$ est faible, gage aussi d'une bonne aptitude à la prédiction.

IV-5- Diagramme de Williams :

On a représenté, sur la même figure IV-2; pour les deux ensembles (calibration et validation); les valeurs de $e_{i\ std}$ (tableau IV-3 colonne4 et tableau IV-4 colonne 3) et celle de h_i (tableau IV-3 colonne5 et tableau IV-4 colonne 4).

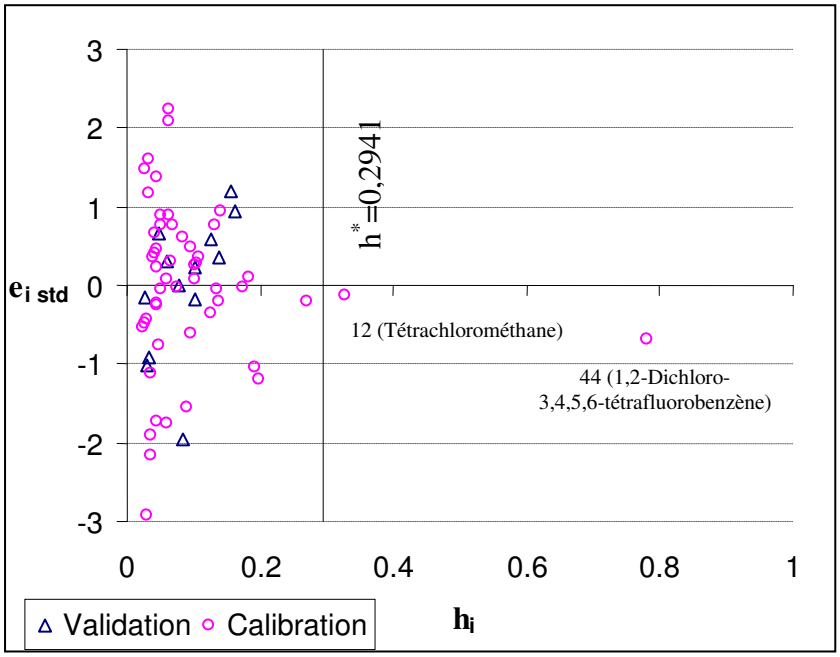


Figure IV-2 : *Diagramme de Williams pour les composés du 2ème groupe (les halogénures).*

Comme le montre la figure IV-2, les valeurs de $e_{i\ std}$ sont toutes comprises entre les bornes ±3. On retrouve deux composés influents 12 (Tétrachlorométhane) et 44 (1,2-Dichloro-3,4,5,6-tétrafluorobenzène)qui apparaissent en dehors de la limite fixée par la valeur critique $h^* = 0{,}2941$ symbolisée par la droite parallèle à l'axe des $e_{i\ std}$.

IV-6- Qualité de l'ajustement :

La figure IV-3 représente les deux droites d'ajustement pour l'ensemble de calibration (ω_{obs} en fonction de ω_{cal} définie par l'équation (IV-2)) et celle pour l'ensemble de validation (ω_{obs} en fonction de ω_{pred} définie par l'équation (IV-3)).

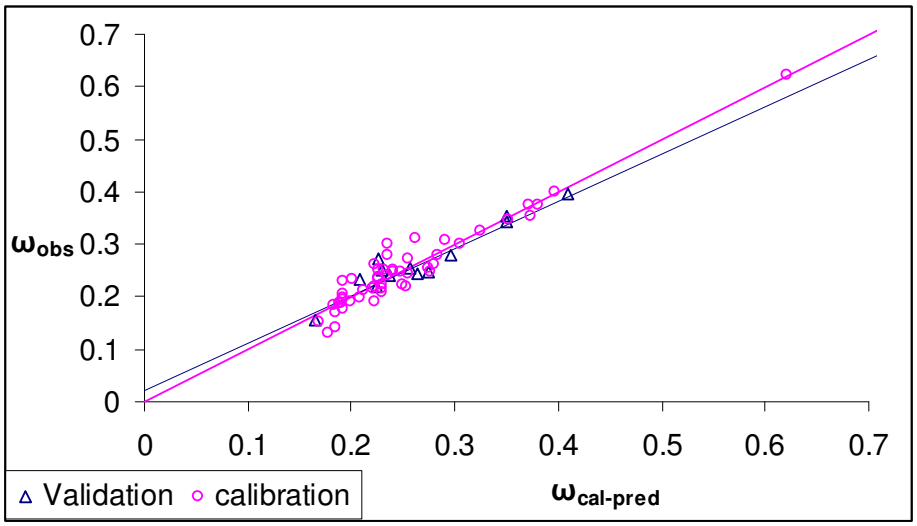


Figure IV-3 : *Droites d'ajustement des deux ensembles: cas des halogénures.*

$$\omega_{obs} = 0{,}0000018 + 0{,}99999\ \omega_{cal} \qquad \text{(IV-2)}$$
$$S = 0{,}0227437 \quad R^2 = 92{,}1\ \% \quad R^2_{ajust} = 91{,}9\ \%$$

$$\omega_{obs} = 0{,}0268489 + 0{,}900646\ \omega_{pred} \qquad \text{(IV-3)}$$
$$S = 0{,}0191567 \quad R^2 = 91{,}6\ \% \quad R^2_{ajust} = 90{,}9\ \%$$

On remarque, relativement, une faible dispersion autour des droites d'ajustement (calibration ou validation) ce qui traduit la faiblesse des erreurs lors du calcul (Calibration) et de la prédiction (Validation).

On déduit donc, de la figure précédente, qu'on a un bon ajustement confirmé par les valeurs du coefficient de régression R^2 supérieures à 90%, pour les deux ensembles. Cela prouve la performance du modèle établi.

Pour juger au mieux de la valeur du modèle on a réuni dans le tableau IV-5 les paramètres statistiques du modèle utilisé pour la prédiction du facteur acentrique pour cette classe.

Tableau IV-5 : Paramètres statistiques.

EQMP	0,0236	51 observations	R^2	92,05 %
EQMC	0,0223		Q^2	91,08 %
$EQMP_{ext}$	0,019	13 observations	Q^2_{ext}	91,28 %

Les valeurs de Q^2, R^2 et Q^2_{ext} sont assez élevées et très similaires. Les écarts quadratiques moyens de calcul, de prédiction et de prédiction externe ont des valeurs petites et voisines. Ces deux remarques faites sur le tableau précédent résultent de la stabilité de notre modèle. La validité du modèle a été éprouvée par le test de randomisation de y (Figure IV-4).

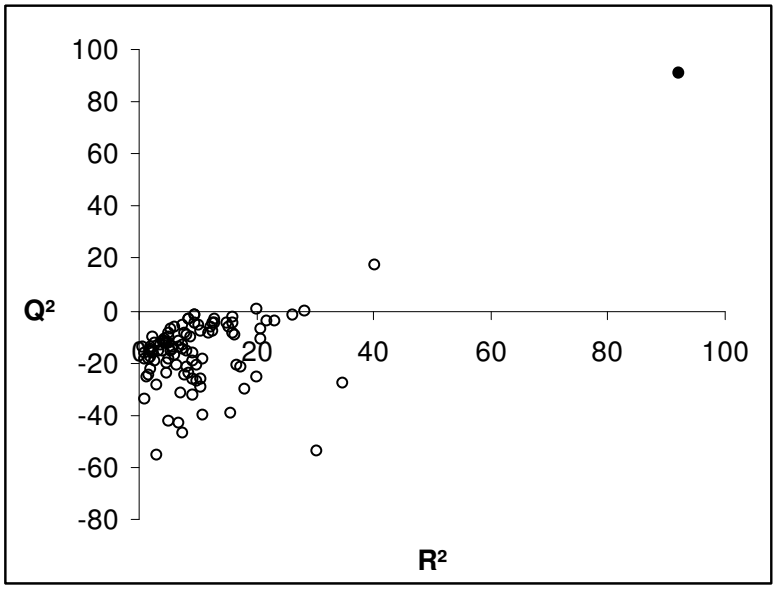


Figure IV-4 : *Test de randomisation: cas des halogénures.*

Les cercles vides regroupés dans la région des valeurs négatives de Q^2 ont des valeurs petites de R^2 ; ce sont les représentations des modèles randomisés. Seul le cercle noirci a des valeurs élevées et proches pour ces deux statistiques, il représente notre modèle qui, par conséquent, n'est pas dû au hasard.

V- Modélisation du troisième groupe :

Pour ce groupe l'ensemble de validation compte 7 composés et celui de calibration 30. Le choix des deux sous-ensembles a été fait de façon aléatoire.

V-1- Calcul du modèle :

Nous avons observé (Figure V-1) que la fonction FIT atteint un presque palier à 17,08 pour un nombre de descripteurs p=3. L'augmentation de cette fonction n'est pas très significative quand le modèle passe de 3 à 4 régresseurs. Nous avons donc construit des modèles à 3 descripteurs en procédant à l'optimisation de la fonction Q^2 par algorithme génétique.

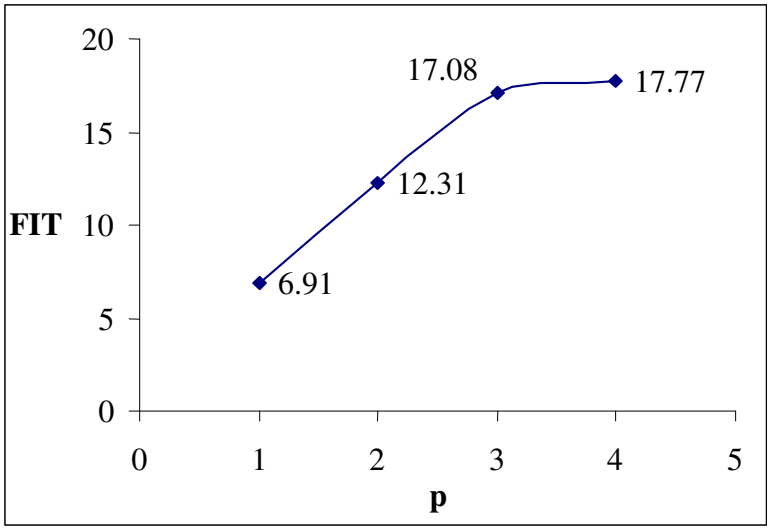


Figure V-1 : *Variation du FIT en fonction du nombre de descripteurs.*

Le modèle que nous avons choisi de présenter pour ce groupe est composé des descripteurs regroupés dans le tableau V-1.

Tableau V-1 : Descripteurs du modèle pour le 3ème groupe.

N°	Descripteur	Classe	Définition
1	GMTIV	Descripteur topologique	MTI de Gutman par le degré de valence des sommets
2	BELm2	Valeur propre de Burden	2ème plus petite valeur propre de la matrice de Burden pondérée par la masse atomique
3	Mor20v	Descripteur MoRSE-3D	MoRSE-3D signal-20 pondéré par les volumes atomiques de Van der Waals

Les corrélations linéaires entre les variables du modèle sont reproduites dans le tableau V-2

Tableau V-2 : Matrice de corrélation.

	ω	GMTIV	BELm2
GMTIV	0,912		
	0		
BELm2	0,553	0,297	
	0,002	0,111	
Mor20v	-0,573	-0,357	-0,377
	0,001	0,053	0,04

V-2- Equation et analyse de régression :

Le modèle obtenu par RLM est résumé par l'équation de régression suivante :

$$\omega = 0,06709\ (\pm 0,03632) + 0,00033109\ (\pm 0,00002010)\ GMTIV + 0,13201\ (\pm 0,02483)\ BELm2 - 0,14206\ (\pm 0,03290)\ Mor20v$$

$$n=30\ ;\ R^2= 95,29\ \%\ ;\ Q^2= 93,49\ \%\ ;\ F= 175,2773$$
$$EQMP= 0,019\ ;\ EQMC= 0,016 \qquad (V\text{-}1)$$

Le modèle calculé sur les 30 observations que compte le groupe de calibration corrèle très bien le facteur acentrique à ces trois variables puisque la valeur de R^2 est grande. La valeur de Q^2 est grande et montre les capacités prédictives du modèle. La valeur élevée de la statistique de Fisher indique un modèle très significatif. Les écarts quadratiques (EQMP/C) sont faibles et proches.

V-3- Analyse des résidus et diagnostics d'influence :

Les valeurs de ω calculées et expérimentales, les résidus ordinaires et standardisés ainsi que las valeurs des leviers pour chaque composé sont rassemblés dans les colonnes de 1 à 5 du tableau V-3.

Les valeurs de la colonne 3 (résidus ordinaires e_i) sont toutes inférieures en valeur absolue à trois fois l'erreur standard du modèle, S soit $3\times 0,0172= 0,0516$.

Tableau V-3 : Valeurs des ω observés (ω_{obs}) et calculés (ω_{cal}), des résidus ordinaires (e_i) et standardisés ($e_{i\ std}$) ainsi que des leviers (h_i).

Composé	ω_{obs}	ω_{cal}	e_i	$e_{i\ std}$	h_i
1	0,538	0,5119	-0,0261	-2,2407	0,229
2	0,417	0,4617	0,0447	3,3691	0,159
3	0,461	0,4494	-0,0116	-0,8692	0,158
4	0,431	0,4348	0,0038	0,2777	0,135
5	0,396	0,3879	-0,0081	-0,6832	0,217
6	0,314	0,3415	0,0275	2,0357	0,15
7	0,326	0,2898	-0,0362	-2,4301	0,092
8	0,369	0,3608	-0,0082	-0,6223	0,161
9	0,316	0,3122	-0,0038	-0,2389	0,051
10	0,333	0,3154	-0,0176	-1,2799	0,138
11	0,269	0,2743	0,0053	0,3444	0,075
12	0,268	0,2826	0,0146	1,0119	0,112
13	0,281	0,2814	0,0004	0,0302	0,152
14	0,266	0,2559	-0,0101	-0,6911	0,101
15	0,244	0,2434	-0,0006	-0,0446	0,118
16	0,2	0,2184	0,0184	1,4216	0,172
17	0,483	0,4632	-0,0198	-1,5759	0,188
18	0,378	0,3963	0,0183	1,3769	0,158
19	0,385	0,3878	0,0028	0,2144	0,169
20	0,35	0,3593	0,0093	0,6066	0,073
21	0,344	0,3495	0,0055	0,3977	0,134
22	0,331	0,3262	-0,0048	-0,3341	0,119
23	0,32	0,3014	-0,0186	-1,3172	0,124
24	0,347	0,364	0,017	1,0663	0,051
25	0,455	0,4496	-0,0054	-0,38	0,12
26	0,392	0,3999	0,0079	0,5703	0,135
27	0,285	0,2775	-0,0075	-0,7129	0,28
28	0,391	0,394	0,003	0,1995	0,088
29	0,38	0,379	-0,001	-0,0634	0,072
30	0,362	0,3629	0,0009	0,0607	0,066

V-4- Validation externe :

L'ensemble écarté avant le calcul du modèle nous a servi à valider ce dernier afin d'évaluer sa capacité à prédire d'éventuelles nouvelles valeurs. Les composés de validation sont numérotés de 31 à 37 dans le tableau V-4. Les valeurs des erreurs standards de prédiction $e_{i\ std}$, les valeurs des leviers ainsi que les valeurs prédites ω_{pred} et observées ω_{obs} du facteur acentrique y sont aussi rapportées.

Tableau V-4 : Quelques caractéristiques des éléments de l'ensemble de validation externe.

Composé	ω_{obs}	ω_{pred}	$e_{i\ std}$	h_i
31	0,331	0,3246	-0,423	0,216
32	0,35	0,3268	-1,403	0,074
33	0,362	0,3401	-1,3331	0,086
34	0,391	0,3915	0,0305	0,129
35	0,346	0,3461	0,0055	0,108
36	0,502	0,4748	-1,7634	0,194
37	0,271	0,2722	0,0733	0,078

Les valeurs introduites pour la validation du modèle ont été prédites sans trop d'erreur (valeur de l'écart quadratique moyen de prédiction externe faible $EQMP_{ext} = 0,016$). Cette constatation est confirmée par la grande valeurs de Q^2_{ext}=95,26 % qui est une preuve de l'aptitude du modèle à la prédiction.

V-5- Diagramme de Williams :

La figure V-2 est une représentation des valeurs des résidus standardisés $e_{i\ std}$ des composés pour les deux sous-groupes (calibration et validation) en fonction de leurs leviers respectifs.

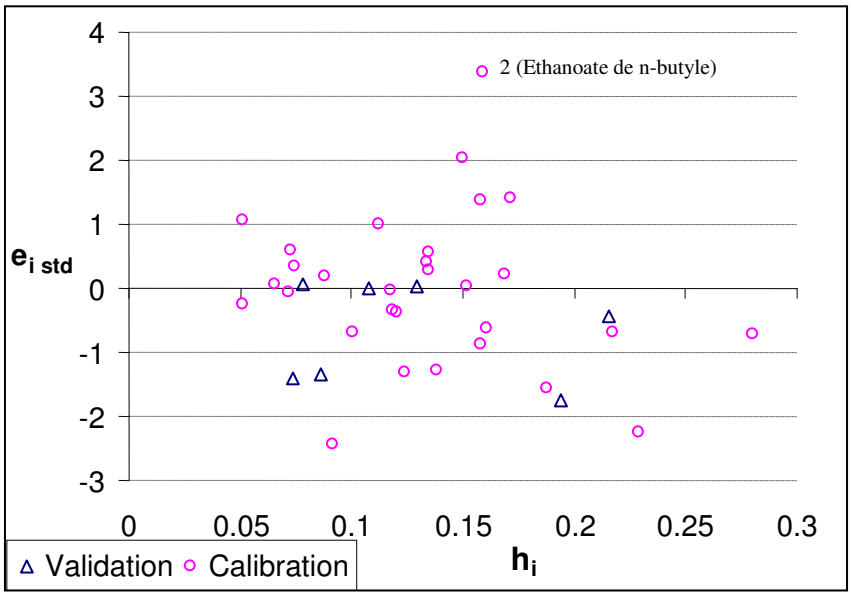


Figure V-2: *Diagramme de Williams pour le 3ème groupe.*

Les valeurs des leviers sont toutes inférieures à la valeur critique h^*=0,4 donc on n'a pas de point influent, mais comme le montre le diagramme on détecte un point aberrant (composé 2: Ethanoate de n-butyle) caractérisé par la valeur de son résidu standardisé, supérieure à 3.

V-6- Qualité de l'ajustement :

La qualité de l'ajustement peut être appréciée dans la figure V-3 par les deux droites d'ajustements dont les équations sont :

- Pour l'ensemble de calibration

$$\omega_{obs} = -0{,}0000323 + 1{,}00009\ \omega_{cal}$$
$$S = 0{,}0165858 \qquad R^2 = 95{,}3\ \% \qquad R^2_{ajust} = 95{,}1\ \% \qquad \text{(V-2)}$$

- Pour l'ensemble de validation

$$\omega_{obs} = -0{,}0216247 + 1{,}09219\ \omega_{pred}$$
$$S = 0{,}0122221 \qquad R^2 = 97.5\ \% \qquad R^2_{ajust} = 97.0\ \% \qquad \text{(V-3)}$$

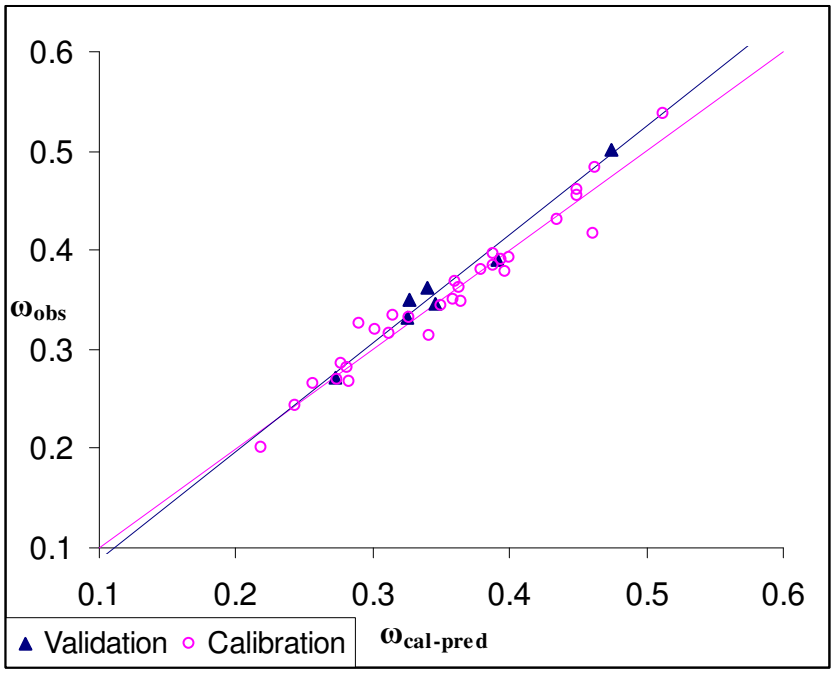


Figure V-3: *Droites d'ajustement.*

La faible dispersion des points autour des droites montre que les valeurs prédites (pour l'ensemble de validation) et calculées (pour l'ensemble de calibration) sont en adéquation avec les valeurs expérimentales. Les valeurs de R^2 montrent la bonne qualité d'ajustement qu'offre notre modèle. Pour avoir une vue globale des autres qualités de ce dernier on a rassemblé toutes les statistiques s'y rapportant dans le tableau suivant:

Tableau V-5: Statistiques du modèle

EQMC	0,016	30 observations	R^2	95,29 %
EQMP	0,019		Q^2	93,49 %
$EQMP_{ext}$	0,016	7 observations	Q^2_{ext}	95,26 %

Les valeurs de Q^2, R^2 et Q^2_{ext} sont comme pour les autres modèles grandes et proches. Les écarts quadratiques sont aussi proches mais faibles, preuves que le modèle prédit ou calcule sans trop d'erreurs les valeurs du facteur acentrique.

Par le test de randomisation (figure V-4), nous nous sommes assuré qu'une relation structure-propriété réelle a été établie par notre modèle.

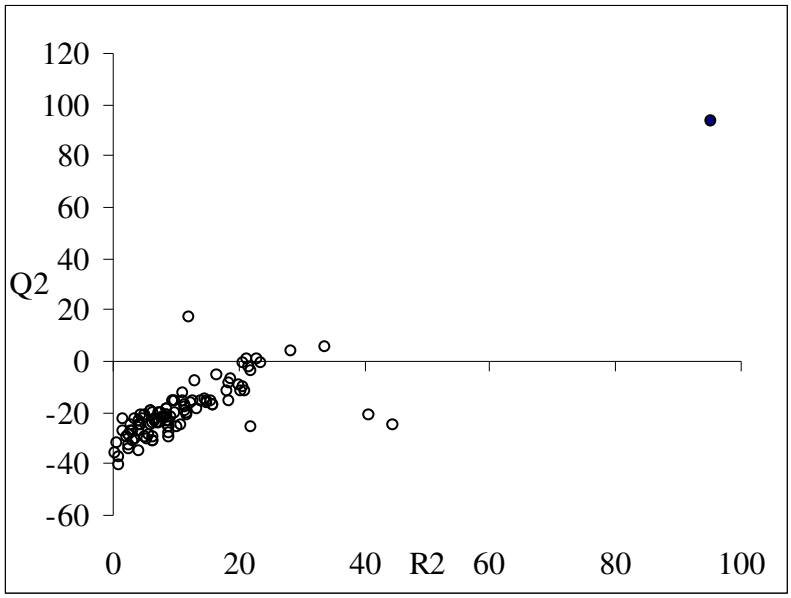


Figure V-4: *Test de randomisation.*

Les 100 modèles pour lesquels nous avons randomisé les valeurs du facteur acentrique ont des valeurs ou faibles ou négatives de Q^2, et des valeurs du coefficient de corrélation multiple (R^2) petites. Seul le modèle (cercle plein) avec le vecteur réel offre des valeurs élevées pour les deux statistiques représentées dans la figure V-4.

VI- Modèles linéaire et non linéaire pour la prédiction du facteur acentrique :

La totalité des 119 composés a servi pour la modélisation linéaire (RLM) et non linéaire (Réseau de Neurones Artificiels).

VI-1- Calcul des modèles :

Le calcul du modèle a été fait sur 96 composés qui constituent l'ensemble de calibration et sa validation a été faite sur les 23 restants. Cet ensemble est constitué de l'union des 2 ensembles de validation du deuxième (13) et du troisième (7) groupe, en plus de 3 alcools choisis au hasard dans le premier groupe.

La dimension du modèle a été fixée à 5 descripteurs en suivant le même raisonnement que pour les deux derniers groupes étudiés. La figure VI-1 montre la variation de la fonction de Kubinyi.

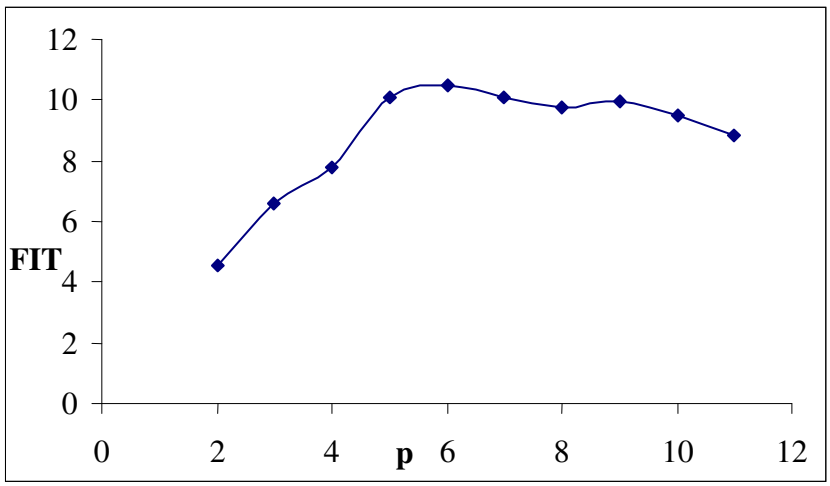


Figure VI-1 : V*ariation du FIT en fonction de la dimension du modèle.*

En procédant à l'optimisation par algorithme génétique du coefficient de prédiction Q^2 on a obtenu les variables explicatives utilisées (tableau VI-1) pour les deux méthodes (tableau VI-1). La matrice de corrélation est dans le tableau VI-2.

Tableau VI-1 : Choix d'un sous-ensemble de descripteurs significatif par algorithme génétique.

N°	Descripteur	Classe	Définition
1	Hy	Propriétés moléculaires	Facteur d'hydrophilie
2	HIC	Descripteurs GETAWAY	Renseignements sur la teneur moyenne sur l'ampleur du levier
3	RDF050m	Descripteurs RDF	Fonction de distribution radiale 5.0/pondérée par la masse atomique
4	ESpm09d	Indice de contiguïté	Moment spectral 09 de la matrice de contiguïté pondérée par le moment dipolaire.
5	J	Descripteurs topologiques	Indice de connectivité de la distance de Balaban

Tableau VI-2 : Matrice de corrélation

	Omega	Hy	HIC	RDF050m	ESpm09d
Hy	0,36				
	0				
HIC	0,628	-0,394			
	0	0			
RDF050m	0,359	-0,075	0,205		
	0	0,467	0,045		
ESpm09d	-0,26	-0,248	-0,121	0,036	
	0,011	0,015	0,242	0,725	
J	-0,041	-0,057	0,078	0,064	0,818
	0,689	0,581	0,452	0,534	0

VI-2- Analyse de régression et architecture du réseau :

L'équation de régression pour le modèle RLM est la suivante :

ω = -0,11032 (±0,02549) + 0,325 (±0,01475) Hy + 0,18418 (±0,007153) HIC + 0,016059 (±0,002102) RDF050m - 0,07532 (±0,01125) J+ 0,017720 (±0,01125) ESpm09d

n=97 ; R^2= 92,24% ; Q^2= 90,68% ; F= 213,88
EQMP= 0.040 ; EQMC= 0,037. (VI-1)

Pour la méthode neuronale on a opté pour huit (8) neurones dans la couche cachée et 500 itérations ce que justifie la figure VI-2. La structure du réseau est reproduite dans le tableau VI-3.

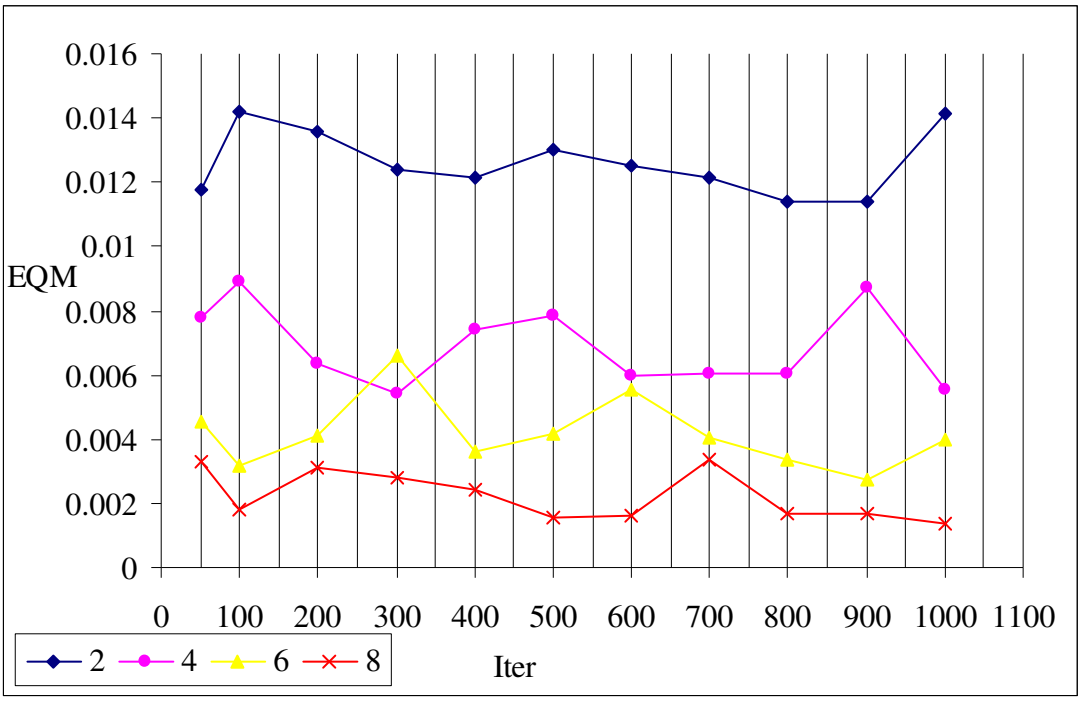


Figure VI-2 : *Variation de l'erreur quadratique moyenne en fonction du nombre d'itérations pour chaque choix du nombre de neurones.*

Tableau VI-3 : Architecture du réseau

Entrées	05 (les descripteurs)
Sortie	01 (le facteur acentrique)
Couche cachée	Une couche cachée
Nombre de neurones dans la couche cachée	08
Nombre d'itérations	500
Algorithme d'apprentissage	Rétro propagation du gradient de l'erreur
Fonction d'apprentissage	Tangente hyperbolique (couche cachée) Linéaire (couche de sortie)

n=97 ; R^2= 96,91% ; Q^2= 96,62 % ; F= 501,62
EQMP= 0,024 ; EQMC= 0,023.

VI-3- Résultats des deux méthodes :

Tableau VI-4 : Valeurs calculées et observées du facteur acentrique, leviers des observations et résidus pour les deux méthodes.

Composé			RLM			RNA		
	ω_{obs}	h_i	ω_{cal}	e_i	$e_{i\ std}$	ω_{cal}	e_i	$e_{i\ std}$
1	0,177	0,054	0,212	0,035	0,9985	0,1831	0,0061	0,2593
2	0,26	0,05	0,2156	-0,0444	-1,258	0,2241	-0,0359	-1,5427
3	0,198	0,054	0,213	0,015	0,4282	0,1837	-0,0143	-0,6090
4	0,187	0,115	0,147	-0,04	-1,258	0,1964	0,0094	0,4134
5	0,171	0,054	0,213	0,042	1,198	0,1851	0,0141	0,6004
6	0,221	0,05	0,2165	-0,0045	-0,1265	0,2248	0,0038	0,1612
7	0,204	0,05	0,2076	0,0036	0,101	0,1871	-0,0169	-0,7188
8	0,153	0,107	0,1549	0,0019	0,059	0,1981	0,0451	2,0173
9	0,184	0,054	0,2137	0,0297	0,8458	0,1852	0,0012	0,0510
10	0,189	0,045	0,1949	0,0059	0,1658	0,1944	0,0054	0,2285
11	0,199	0,058	0,1789	-0,0201	-0,5779	0,2032	0,0042	0,1789
.12	0,193	0,045	0,1859	-0,0071	-0,1982	0,198	0,005	0,2115
13	0,251	0,027	0,2801	0,0291	0,7954	0,2605	0,0095	0,3984
14	0,279	0,061	0,2664	-0,0126	-0,3621	0,2577	-0,0213	-0,9128
15	0,215	0,047	0,2156	0,0006	0,0157	0,2138	-0,0012	-0,0508
16	0,263	0,063	0,2616	-0,0014	-0,0414	0,2567	-0,0063	-0,2691
17	0,191	0,045	0,2154	0,0244	0,6858	0,2142	0,0232	0,9866
18	0,245	0,061	0,2691	0,0241	0,6963	0,257	0,012	0,5126
19	0,256	0,063	0,2629	0,0069	0,2006	0,2578	0,0018	0,0769
20	0,24	0,018	0,244	0,004	0,107	0,2281	-0,0119	-0,4971
21	0,278	0,025	0,2496	-0,0284	-0,7739	0,2527	0,0017	0,0714
22	0,248	0,062	0,2636	0,0156	0,45	0,622	0	0,0000
23	0,325	0,142	0,3347	0,0097	0,3211	0,2909	0,0189	0,7960
24	0,308	0,055	0,314	0,006	0,1719	0,41	0,01	0,4314
25	0,235	0,038	0,256	0,021	0,5833	0,2388	0,0218	0,9150
26	0,31	0,014	0,3017	-0,0083	-0,2232	0,2175	0,0075	0,3171
27	0,374	0,234	0,34	-0,034	-1,3312	0,1872	-0,0448	-1,9375
28	0,19	0,025	0,2934	0,1034	2,8152	0,2586	0,0126	0,5383
29	0,3	0,025	0,2887	-0,0113	-0,307	0,2099	-0,0191	-0,8126
30	0,218	0,046	0,2782	0,0602	1,6933	0,4767	0,0227	0,9621
31	0,14	0,033	0,173	0,033	0,9122	0,4579	-0,0471	-2,0382
32	0,223	0,035	0,1851	-0,0379	-1,0493	0,4599	0,0269	1,1480
33	0,252	0,035	0,1826	-0,0694	-1,923	0,5234	0,0214	0,9412
34	0,22	0,031	0,1829	-0,0371	-1,0189	0,5848	-0,0022	-0,0966
35	0,213	0,036	0,1649	-0,0481	-1,3344	0,5655	0,0055	0,2397

Tableau VI-4 : Suite

Composé			RLM			RNA		
	ω_{obs}	h_i	ω_{cal}	e_i	$e_{i\ std}$	ω_{cal}	e_i	$e_{i\ std}$
36	0,238	0,026	0,2523	0,0143	0,3904	0,5197	-0,0083	-0,3613
37	0,13	0,041	0,1843	0,0543	1,5168	0,5603	0,0003	0,0130
38	0,346	0,03	0,3324	-0,0136	-0,3722	0,5948	0,0018	0,0772
39	0,373	0,03	0,3447	-0,0283	-0,7759	0,5981	0,0061	0,2606
40	0,299	0,03	0,2932	-0,0058	-0,158	0,5988	-0,0132	-0,5702
41	0,249	0,032	0,2561	0,0071	0,1946	0,6232	0,0002	0,0086
42	0,355	0,028	0,3487	-0,0063	-0,1713	0,6513	-0,0137	-0,5925
43	0,251	0,031	0,2668	0,0158	0,4346	0,649	0,005	0,2223
44	0,622	0,871	0,6228	0,0008	0,459	0,5573	0,0013	0,0648
45	0,272	0,03	0,2926	0,0206	0,566	0,4658	-0,0722	-3,2071
46	0,4	0,08	0,4283	0,0283	0,842	0,4649	0,0479	2,0710
47	0,217	0,021	0,2447	0,0277	0,7487	0,4501	-0,0109	-0,4597
48	0,21	0,043	0,2068	-0,0032	-0,0895	0,4089	-0,0221	-0,9364
49	0,232	0,048	0,1546	-0,0774	-2,1881	0,4059	0,0099	0,4144
50	0,246	0,061	0,2676	0,0216	0,6229	0,3647	0,0507	2,1735
51	0,229	0,049	0,209	-0,02	-0,5651	0,3039	-0,0221	-0,9301
52	0,454	0,039	0,4799	0,0259	0,723	0,3649	-0,0041	-0,1737
53	0,505	0,045	0,476	-0,029	-0,8139	0,3155	-0,0005	-0,0212
54	0,433	0,048	0,4769	0,0439	1,2409	0,3153	-0,0177	-0,7495
55	0,502	0,108	0,5555	0,0535	1,6645	0,314	0,045	1,9370
56	0,587	0,113	0,5995	0,0125	0,3926	0,2045	-0,0635	-2,7955
57	0,56	0,1	0,5838	0,0238	0,7324	0,2706	-0,0104	-0,4383
58	0,528	0,097	0,6113	0,0833	2,5452	0,2869	0,0209	0,8814
59	0,56	0,087	0,5716	0,0116	0,35	0,2469	0,0029	0,1222
60	0,593	0,072	0,5567	-0,0363	-1,0655	0,2331	0,0331	1,4195
61	0,592	0,063	0,5596	-0,0324	-0,9375	0,4452	-0,0378	-1,6241
62	0,612	0,081	0,5834	-0,0286	-0,8499	0,3763	-0,0017	-0,0715
63	0,623	0,085	0,5532	-0,0698	-2,0933	0,3648	-0,0202	-0,8530
64	0,665	0,083	0,5809	-0,0841	-2,5133	0,3665	0,0165	0,7088
65	0,644	0,135	0,5822	-0,0618	-2,0169	0,3266	-0,0174	-0,7317
66	0,556	0,312	0,6559	0,0999	4,5942	0,3067	-0,0243	-1,0253
67	0,538	0,034	0,4296	-0,1084	-2,9943	0,3015	-0,0185	-0,7787
68	0,417	0,042	0,431	0,014	0,391	0,3524	0,0054	0,2312
69	0,461	0,037	0,4205	-0,0405	-1,1247	0,446	-0,009	-0,3790
70	0,431	0,039	0,3934	-0,0376	-1,0477	0,4007	0,0087	0,3673
71	0,396	0,023	0,3866	-0,0094	-0,2561	0,2952	0,0102	0,4270
72	0,314	0,021	0,35	0,036	0,9746	0,4125	0,0215	0,9084
73	0,326	0,026	0,3115	-0,0145	-0,3966	0,4181	0,0381	1,6216

Tableau VI-4 : Suite et fin.

Composé			RLM			RNA		
	ω_{obs}	h_i	ω_{cal}	e_i	$e_{i\ std}$	ω_{cal}	e_i	$e_{i\ std}$
74	0,369	0,048	0,3638	-0,0052	-0,1457	0,3649	-0,0041	-0,1737
75	0,316	0,053	0,3402	0,0242	0,69	0,3155	-0,0005	-0,0212
76	0,333	0,041	0,3305	-0,0025	-0,0689	0,3153	-0,0177	-0,7495
77	0,269	0,039	0,3365	0,0675	1,8802	0,314	0,045	1,9370
78	0,268	0,041	0,2386	-0,0294	-0,8229	0,2045	-0,0635	-2,7955
79	0,281	0,036	0,2937	0,0127	0,3519	0,2706	-0,0104	-0,4383
80	0,266	0,031	0,3037	0,0377	1,0372	0,2869	0,0209	0,8814
81	0,244	0,038	0,2677	0,0237	0,6588	0,2469	0,0029	0,1222
82	0,2	0,05	0,2435	0,0435	1,2335	0,2331	0,0331	1,4195
83	0,483	0,047	0,412	-0,071	-2,0027	0,4452	-0,0378	-1,6241
84	0,378	0,035	0,3674	-0,0106	-0,2927	0,3763	-0,0017	-0,0715
85	0,385	0,034	0,364	-0,021	-0,5811	0,3648	-0,0202	-0,8530
86	0,35	0,069	0,3464	-0,0036	-0,1044	0,3665	0,0165	0,7088
87	0,344	0,028	0,3326	-0,0114	-0,3121	0,3266	-0,0174	-0,7317
88	0,331	0,029	0,3257	-0,0053	-0,1445	0,3067	-0,0243	-1,0253
89	0,32	0,029	0,3065	-0,0135	-0,3691	0,3015	-0,0185	-0,7787
90	0,347	0,067	0,3711	0,0241	0,7004	0,3524	0,0054	0,2312
91	0,455	0,035	0,4175	-0,0375	-1,0377	0,446	-0,009	-0,3790
92	0,392	0,04	0,3783	-0,0137	-0,3805	0,4007	0,0087	0,3673
93	0,285	0,023	0,3071	0,0221	0,5993	0,2952	0,0102	0,4270
94	0,391	0,034	0,3864	-0,0046	-0,1276	0,4125	0,0215	0,9084
95	0,38	0,029	0,3962	0,0162	0,4435	0,4181	0,0381	1,6216
96	0,362	0,03	0,365	0,003	0,0814	0,368	0,006	0,2519

. Les résidus ordinaires sont tous inférieurs à 3S pour les deux méthodes (0,114 : cas de la RLM et 0,072 : cas du RNA).

VI-4- Validation externe :

Les valeurs prédites du facteur acentriques par RLM et RNA ainsi que les résidus pour l'ensemble de validation (composés numérotés de 97 à 119) sont dans le tableau suivant :

Tableau VI-5 : Statistiques pour le groupe de validation par RNA et RLM.

Composé			RLM		RNA	
	ω_{obs}	h_i	ω_{pred}	$e_{i\ std}$	ω_{pred}	$e_{i\ std}$
97	0,3960	0,029	0,3526	-1,1558	0,3554	-1,7379
98	0,3440	0,03	0,3333	-0,2854	0,335	-0,3856
99	0,2510	0,027	0,277	0,6908	0,2581	0,3033
100	0,2560	0,02	0,2653	0,247	0,2516	-0,1866
101	0,2400	0,014	0,2901	1,3227	0,2761	1,5218
102	0,2320	0,02	0,2595	0,7279	0,234	0,0848
103	0,2180	0,042	0,1813	-0,985	0,2067	-0,4902
104	0,3550	0,031	0,3342	-0,5537	0,3379	-0,7335
105	0,2490	0,03	0,2515	0,0657	0,2325	-0,7070
106	0,2440	0,031	0,2576	0,3619	0,2389	-0,2187
107	0,2810	0,034	0,2797	-0,0342	0,2641	-0,7271
108	0,2710	0,057	0,1975	-1,9863	0,2079	-2,7813
109	0,1570	0,075	0,1244	-0,8896	0,173	0,7189
110	0,5790	0,076	0,5615	-0,4778	0,5689	-0,4543
111	0,4380	0,048	0,4939	1,5022	0,4633	1,1046
112	0,5770	0,064	0,5707	-0,1699	0,59	0,5773
113	0,3310	0,047	0,3563	0,679	0,3511	0,8766
114	0,3500	0,025	0,3458	-0,1111	0,3496	-0,0170
115	0,3620	0,028	0,3426	-0,5152	0,3477	-0,6115
116	0,3910	0,032	0,3775	-0,3605	0,3971	0,2619
117	0,3460	0,032	0,3429	-0,0837	0,3472	0,0515
118	0,5020	0,075	0,4341	-1,8516	0,4356	-2,9837
119	0,2710	0,042	0,2989	0,7475	0,2717	0,03037

VI-5- Diagrammes de Williams :

Les valeurs de h_i (colonne 2 des tableaux VI-5 et VI-6) doivent être comparées à la valeur critique $h^* = \frac{3\times(p+1)}{n} = \frac{3\times(5+1)}{96} = 0{,}1875$. La figure VI-3 montre que dans les deux cas (régression multilinéaire et non linéaire) on a les mêmes points influents 27 (1,1,1,2,2, 3,3,4,4,4-decafluorobutane) 44 (1,2-Dichloro-3,4,5,6-tétrafluorobenzène) et 66 (Méthanol). On doit signaler, tout de même, que ce dernier a une valeur très inhabituelle du résidu standardisé quand il est modélisé par RLM et que cette aberration n'est plus observée par RNA. Les composés numérotés 67 (Formiate de pentyl) pour la RLM et 118 (1-butoxybutane) pour le RNA ont des valeurs très proches de la valeur limite (-3) et l'observation 45 (1,2-Dichlorobenzène) cas du RNA est aberrante.

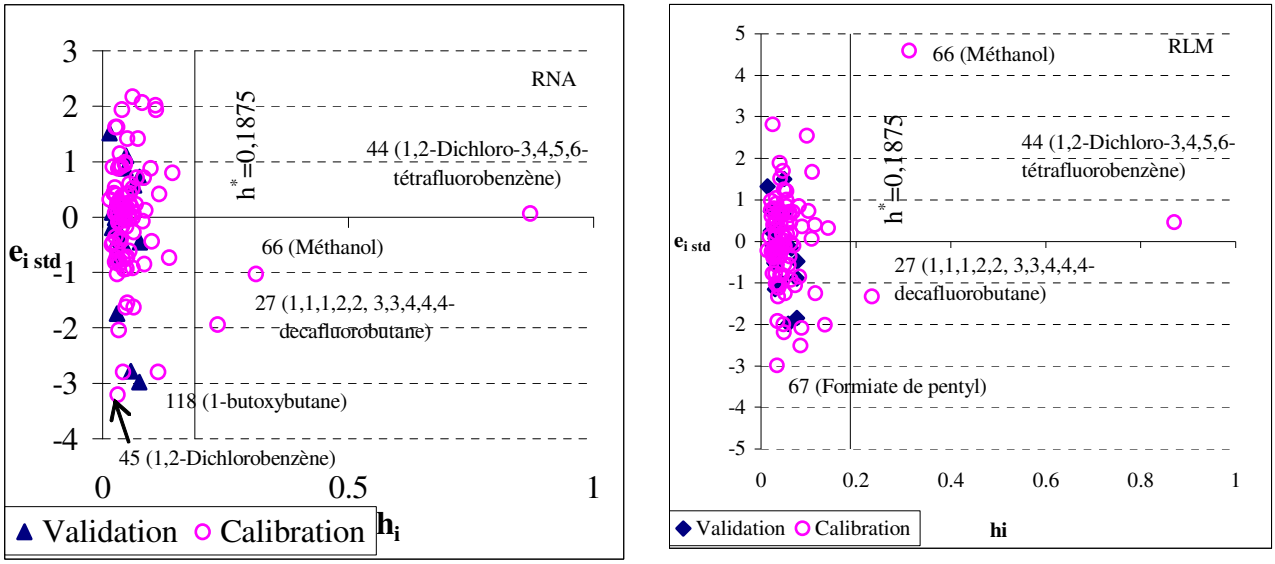


Figure VI-3 : *Diagrammes de Williams par RNA et RLM.*

VI-6- Qualité de l'ajustement :

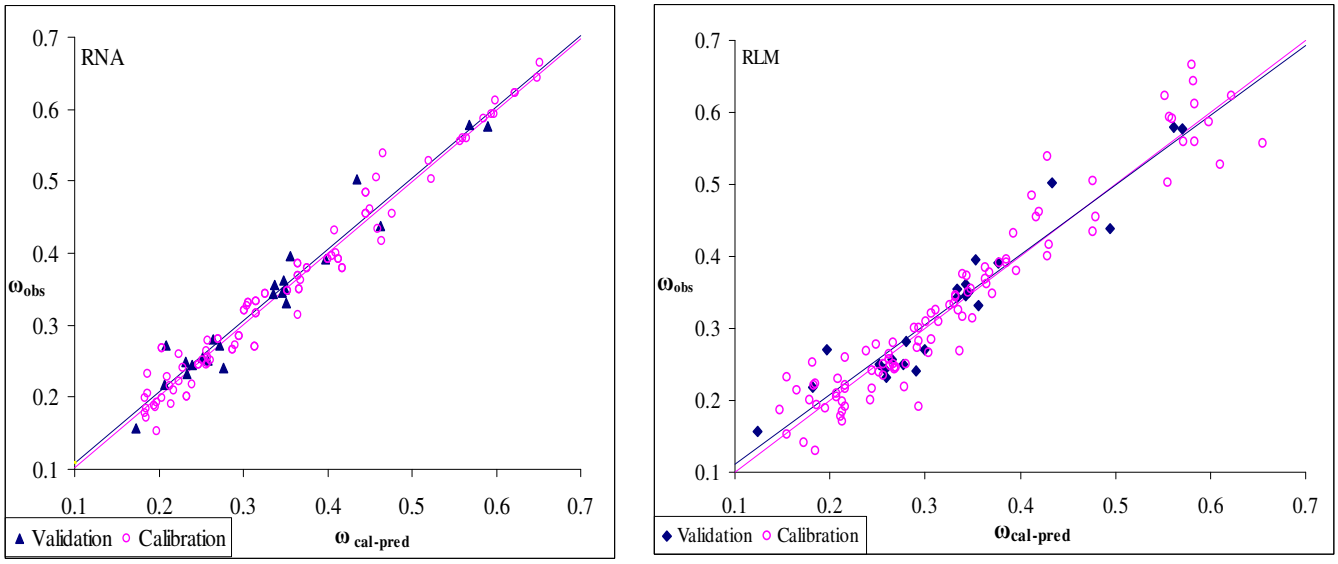


Figure VI-4 : *Qualité de l'ajustement.*

Validation

$\omega_{obs} = 0{,}0107659 + 0{,}986653\ \omega_{pred}$ $S = 0{,}0251866$ $R^2 = 95{,}1\ \%$ $R^2_{ajust} = 94{,}8\ \%$ (VI-2)

$\omega_{obs} = 0{,}0158230 + 0{,}966642\ \omega_{pred}$ $S = 0{,}0335966$ $R^2 = 91{,}2\ \%$ $R^2_{ajust} = 90{,}8\ \%$ (VI-3)

Calibration

$\omega_{obs} = 0{,}0015676 + 0{,}995794\ \omega_{cal}$ $S = 0{,}0235343$ $R^2 = 96{,}7\ \%$ $R^2_{ajust} = 96{,}6\ \%$ (VI-4)

$\omega_{obs} = 0{,}0000038 + 0{,}99999\ \omega_{cal}$ $S = 0{,}0372936$ $R^2 = 92{,}2\ \%$ $R^2_{ajust} = 92{,}2\ \%$(VI-5)

Dans les deux cas (validation externe et calibration) on a une nette amélioration de l'ajustement ou de la prédiction dans le cas du RNA par rapport à la RLM. Car on remarque (figureVI-4) une diminution de la dispersion autour des droites définies par les équations VI-2, 3, 4 et 5, de plus les valeurs de R^2 augmentent sensiblement et pour la calibration (de 92,2 à 96,7) et pour la validation (de 91,2 à 95,1).

Nous avons procédé au test randomisation des modèles obtenus, les résultats apparaissent dans la figure VI-5.

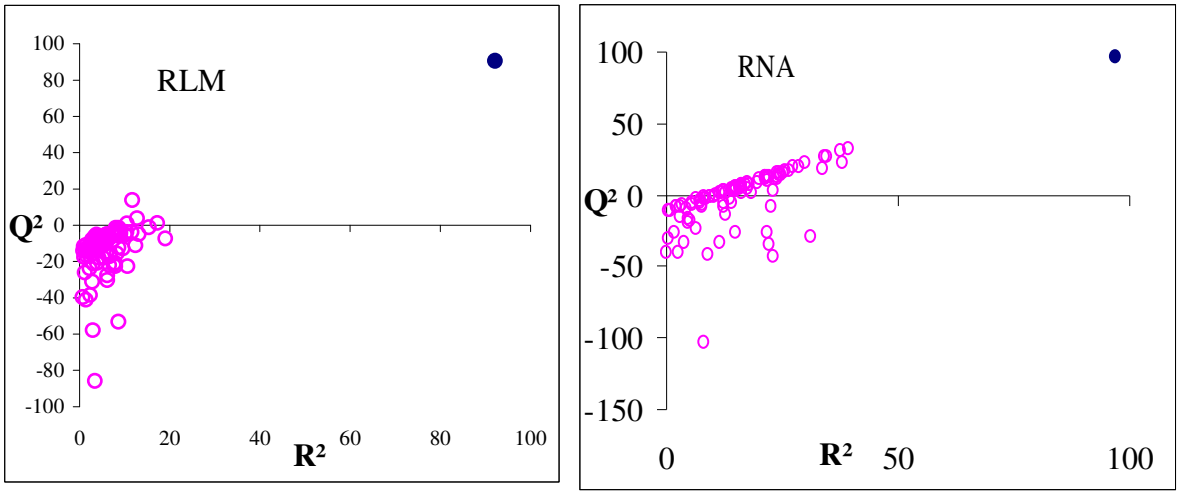


Figure VI-5 : *Tests de randomisation.*

Dans le tableau suivant nous confrontons les résultas statistiques condensés des deux méthodes.

Tableau VI-6 : Récapitulatif des résultats.

	RLM	RNA
R^2	92,24 %	96,91 %
Q^2	90,68 %	96,62 %
Q^2_{ext}	93,92 %	96,45 %
EQMC	0,037	0,023
EQMP	0.040	0,024
$EQMP_{ext}$	0,033	0,024

Pour les valeurs de ce tableau la comparaison se fait toujours à l'avantage de la régression non linéaire, car les écarts qui doivent être faibles (EQMC/P) le sont encore plus avec cette méthode et les coefficients de corrélation et de prédiction (qui doivent avoir de grandes valeurs) sont encore plus grands.

CONCLUSION GENERALE

La méthodologie QSPR a été utilisée pour relier le facteur acentrique, d'un mélange hétérogène de composés organiques, à des descripteurs moléculaires calculés à l'aide de logiciels spécialisés.

Les 119 composés ont été divisés en trois groupes pour chacun desquels on a établi un modèle distinct usant de corrélations linéaires. Puis à des fins de comparaison, des corrélations linéaire et non linéaire ont été imposées pour la totalité des composés. Le premier type de corrélation consiste en une régression multilinéaire et le second est un réseau de neurones artificiels standard à trois couches (les entrées, une couche cachée et une couche de sortie) avec algorithme d'apprentissage par rétropropagation du gradient (Levenberg-Marquardt).

La taille de chaque modèle a été fixée par la valeur optimale de la fonction FIT de Kubinyi et la sélection des variables explicatives a été réalisée par algorithme génétique dans le logiciel MOBY DIGS en maximisant la valeurs du coefficient de prédiction Q^2_{LOO} .

Les statistiques des modèles obtenus sont reproduites dans le tableau suivant :

		Calibration						Validation		
	p	n	R^2	EQMC	Q^2	EQMP	F	n_{ext}	Q^2_{ext}	$EQMP_{ext}$
1^{er} groupe	2	18	91,53 %	0,0191	88,15 %	0,0226	81,04	-	-	-
$2^{ème}$ groupe	4	51	92,05 %	0,0223	91,08 %	0,0236	133,19	13	91,28 %	0,019
$3^{ème}$ groupe	3	30	95,29 %	0,016	93,49 %	0,019	175,27	7	95,26 %	0,016
RLM	5	96	92,24 %	0,037	90,68 %	0.040	213,88	23	93,92 %	0,033
RNA			96,91 %	0,023	96,62 %	0,024	501,62		96,45 %	0,024

Ces statistiques permettent de nous assurer de la qualité de l'ajustement, de la robustesse interne et externe, des capacités prédictives et de la possibilité d'extension suffisante de chaque modèle. Ainsi le facteur acentrique peut être prédit à partir de la seule structure des molécules par des modèles linéaires ou non linéaires, ces derniers se révélant plus performants (les deux dernières lignes du tableau).

Ce travail peut être étendu à un nombre plus important de composés pour chaque famille chimique et le choix ou l'éclatement des données en deux ensembles disjoints (calibration et validation) pourrait se faire d'une manière plus réfléchie.

Enfin on doit rechercher la ou les causes possibles des aberrations relevées pour certains modèles.

REERENCES BIBLIOGRAPHIQUES

1- M.Karelson. Molecular descriptors in QSAR/QSPR. Wiley- Interscience, p. 385 (2000).

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

3- L. Eriksson, E. Vohannson, N. Kettaneh- Wold. Multi and Megavariate Data Analysis- Principles and Applications. Umetricsacademy, Umeå (2001).

4- S. Wold, A. Ruhe, H.Wold, W. Dunn. SIAMJ. Sci. Stat. Comput., 5, 735 (1984).

5- S. Wold. Chemometrics: Mathematics and Statistics in Chemistry. Reidel, Dordrecht, The Netherlands (1984).

6- P. Gelada, B. R. Kowalski, Anal. Chim. Acta, 185, 1 (1986).

7- A. Höskuldsson, J. Chemometrics, 2, 211 (1988).

8- J. A. Burns, G. M. Whiteside. Chem. Rev., 93, 2583 (1993).

9- L. S. Anker, P. C. Jurs. Anal. Chem., 64, 1157 (1992).

10- T. Aoyama, Y. Suzuki, H. Ichikawa. J. Med. Chem., 33, 2583 (1990).

11- T. A. Andrea, J. Med. Chem., 34, 2824 (1991).

12- P. C. Jurs, Computer Software Applications in Chemistry. Second Edition, J. Wiely (1996).

13- A. R. Katritzky, V. S. Lobanov, M. Karelson. CODESSA Reference Manual. University of Florida, Gainesville (1994).

14- V. Y. Nalimov. The Application of Mathematical Statistics to Chemical Analysis, Addison- Wesley, Reading, MA (1962).

15- R. Calcutt, R. Body. Statistics for Analytical Chemists. Champman & Hall, New York (1983).

16- J. C. Miller, J. N. Miller. Statistics for Analytical Chemistry. Ellis Horwood, New York (1988).

17- P. C. Meier, R. E. Zund. Statistical Methods in Analytical Chemistry. Wiley, New York (1993).

18- P. Dagnélie. Statistique Théorique Et Appliquée. Tomes 1 et 2. De Boeck & Larcier s. a. (1998).

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

20- R. Wehrens, H. Putter, L. M. C. Buydens. Chemom. Int. Lab. Syst., 54, 35- 52 (2000).

21- L. Eriksson, J. Jaworska, A. P. Worth, M. T. D. Cronin, R. M. Mc Dowell, P. Gramatica. Methods for Reliability and Uncertainty Assessment and for Applicability Evaluations of Classification and Regression Based QSARs. Environmental Health Perspectives 111, 1361-1375 (2003).

22- A. R. Leach. Modeling. Principales and Applications, 2 nd ed. Pearson, Printice Hall, London (2001).

23- A. R. Leach, V. L. Gillet. An Introduction to Chemoinformatics: Revised Edition; Springer, Dordrecht (2007).

24- L. Chambers. Practical Handbook of Genetic Algorithms. Lewis Publishing (1995).

25- R. Todeschini, D. Ballabio, V. Consonni, A. Mauri, M. Pavan. Moby Digs Software for Multilinear Regression Analysis and Variable Subset Selection by Genetic Algorithm. Release 1.1 for Windows, Milano (2009).

26- H. Kubinyi. Variable selection in QSAR Studies: I. An Evolutionary Algorithm. Quant. Struct.- Act. Relat. 13, 285-294 (1994).

27- J.A. Burns, G. M. Whiteside. Chem. Rev. 93, 2583 (1993).

28- B. J. Wythoff. Chemometrics Intell. Lab. Syst. 19, 115 (1993).

29- J. Zupan, J. Gasteiger. Neural Networks in Chemistry and Drug Design. Second Edition. Wiley VCH, New York (1999).

30- R. Hechet- Nielson. Neurocomputing. Addison- Wesly Publishing Company (1990).

31- MATLAB. Version 7.0.0.19920 (Release 14). The Language of Technical Computing. The Math Works, Inc. May 06 (2004).

32- Hyperchem ^{TM} Release 6.03 for windows, Molecular Modeling System (2000).

33- R. Todeschini, V. Consonni, M. Pavan. DRAGON, Software for the Calculation of Molecular Descriptors. Release 5.3 for windows, Milano (2005).

34- K. S. Pitzer, *J. Am. Chem. Soc.* 77:3427 (1955).

35- K. S. Pitzer, D. Z. Lippman, R. F. Curl, C. M. Huggins, and D. E. Petersen, *J. Am.Chem. Soc.* 77:3433 (1955)

36- W. C. Edmister, *Petrol. Refiner.* 37:173 (1958).

37- T. W. Leland, Jr. and P. S. Chappelear, *Ind. Eng. Chem.* 60:15 (1968).

38- G. D. Fisher and T. W. Leland Jr., *Ind. Eng. Chem. Fundam.* 9:537 (1970).

39- L. I. Stiel, *Chem. Eng. Sci.* 27:2109 (1972).

40- A. S. Teja, *AIChE J.* 26:337 (1980).

41- G. Z. A. Wu and L. I. Stiel, *AIChE J.* 31:1632 (1985).

42- V. L. Bhirud, *AIChE J.* 24:880 (1978).

43- V. L. Bhirud, *AIChE J.* 24:1127 (1978).

44- B. Armstrong, *J. Chem. Eng. Data* 26:168 (1981).

45- A. S. Teja, S. I. Sandler, and N. C. Patel, *Chem. Eng. J.* 21:21 (1981).

46- J. Nath, *Ind. Eng. Chem. Fundam.* 18:297 (1979).

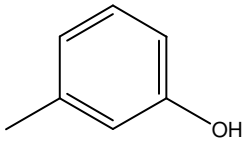
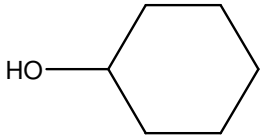
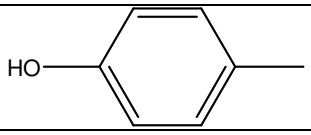
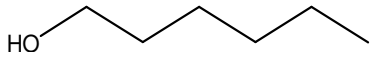
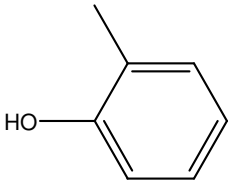
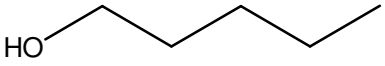
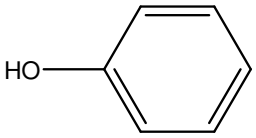
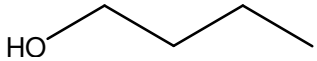
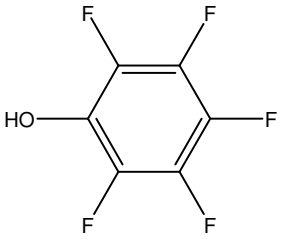
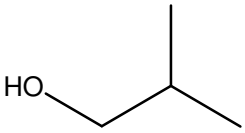
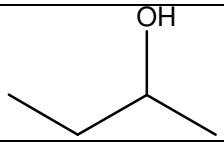
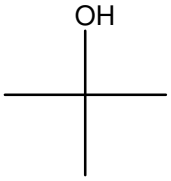
47- C. Tsonopoulos, *AIChE J.* 20:263 (1974).

48- C. Van Ness and M. M. Abbott, *Classical Thermodynamics of Nonelectrolyte Solu*tions (McGraw–Hill, New York, 1982), pp. 126–133.

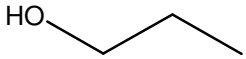
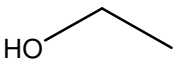
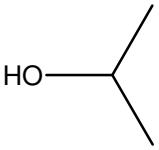
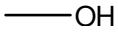
49- R.C. Reid, J.M. Prausnitz, T.K. Sherwood, *The properties of gases and liquids*, 3 ed., McGraw-Hill, New York (1977).

ANNEXES

Annexe 1a : Composés du 1er groupe

n°	Formule chimique et CAS	Formule développée	n°	Formule chimique et CAS	Formule développée
1	C_7H_8O 108-39-4		8	$C_6H_{12}O$ 108-93-0	
2	C_7H_8O 106-44-5		9	$C_6H_{14}O$ 111-27-3	
3	C_7H_8O 95-48-7		10	$C_5H_{12}O$ 123-51-3	
4	C_6H_6O 108-95-2		11	$C_4H_{10}O$ 71-36-3	
5	C_6HF_5O 771-61-9		12	$C_4H_{10}O$ 78-83-1	
6	$C_8H_{18}O$ 111-87-5		13	$C_4H_{10}O$ 78-92-2	
7	$C_7H_{16}O$ 111-70-6		14	$C_4H_{10}O$ 75-65-0	

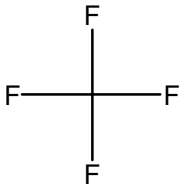
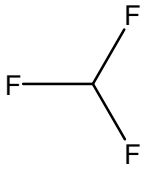
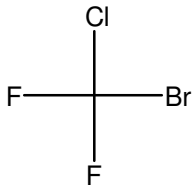
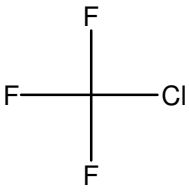
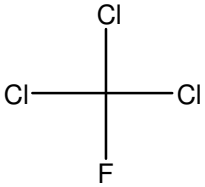
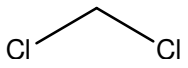
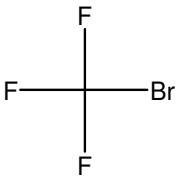
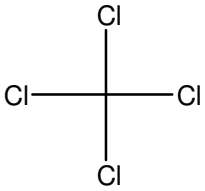
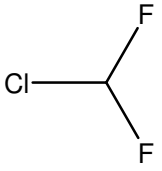
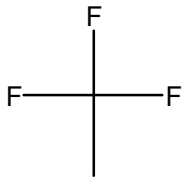
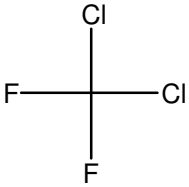
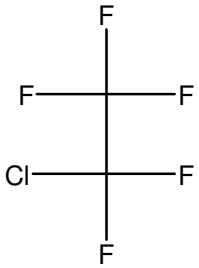
Annexe 1a : suite et fin

n°	Formule chimique et CAS	Formule développée	n°	Formule chimique et CAS	Formule développée
15	C_3H_8O 71-23-8		17	C_2H_6O 64-17-5	
16	C_3H_8O 67-63-0		18	CH_4O 67-56-1	

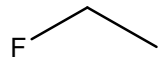
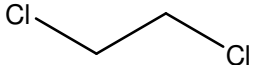
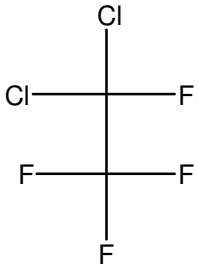
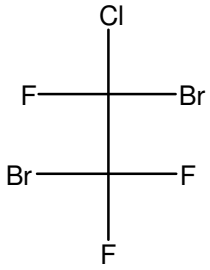
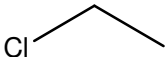
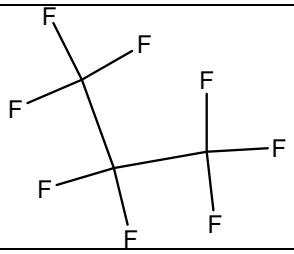
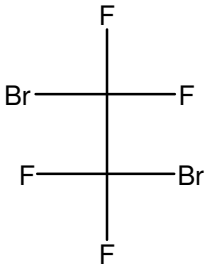
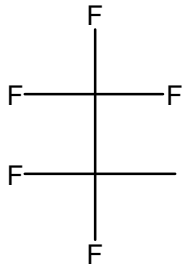
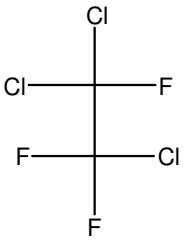
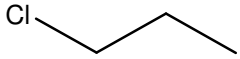
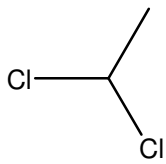
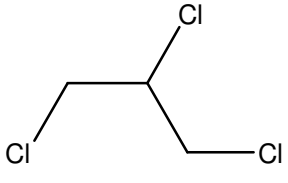
Annexe 1b : Valeurs des descripteurs du 1er groupe

n°	Composé	ω_{obs}	MATS6p	HATS3e
1	3-Méthylphénol	0,454	0,824	0,428
2	4-Méthylphénol	0,505	0,195	0,478
3	2-Méthylphénol	0,433	0,824	0,403
4	Phénol	0,438	0,919	0,513
5	2,3,4,5,6-Pentafluorophénol	0,502	0,822	0,939
6	Octan-1-ol	0,587	-0,022	0,738
7	Heptan-1-ol	0,56	-0,033	0,805
8	Cyclohexanol	0,528	0,487	1,247
9	Hexan-1-ol	0,56	-0,057	0,912
10	Pentan-1-ol	0,579	0,246	1,032
11	Butan-1-ol	0,593	0,392	1,237
12	2-Méthyl-propan-1-ol	0,592	0	1,083
13	Butan-2-ol	0,577	0	1,205
14	2-Méthyl-propan-2-ol	0,612	0	1,119
15	Propan-1-ol	0,623	0	1,546
16	Propan-2-ol	0,665	0	1,488
17	Ethanol	0,644	0	1,863
18	Méthanol	0,556	0	0,911

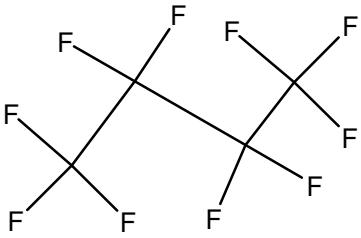
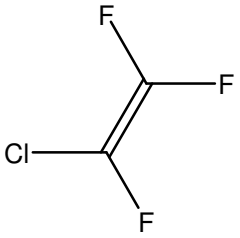
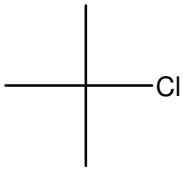
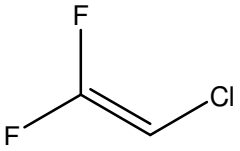
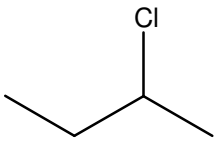
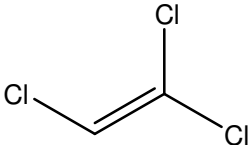
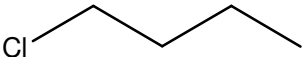
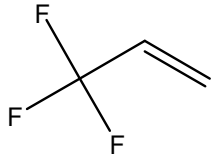
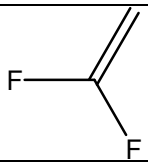
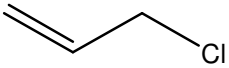
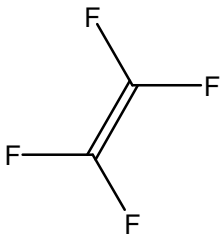
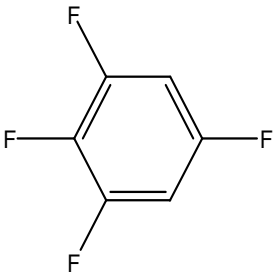
Annexe 2a : Composés du 2ème groupe.

n°	Formule chimique et CAS	Formule développée	n°	Formule chimique et CAS	Formule développée
1	CF_4 75-73-0		8	CH_3Cl 74-87-3	
2	CHF_3 75-46-7		9	$CBrClF_2$ 353-59-3	
3	$CClF_3$ 75-72-9		10	CCl_3F 75-69-4	
4	CH_3F 593-53-3		11	CH_2Cl_2 75-09-2	
5	$CBrF_3$ 75-63-8		12	CCl_4 56-23-5	
6	$CHClF_2$ 75-45-6		13	$C_2H_3F_3$ 420-46-2	
7	CF_2Cl_2 75-71-8		14	C_2ClF_5 76-15-3	

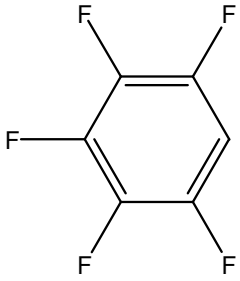
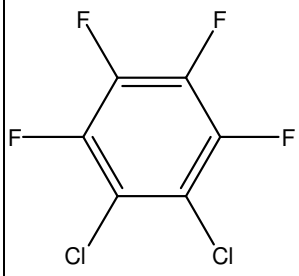
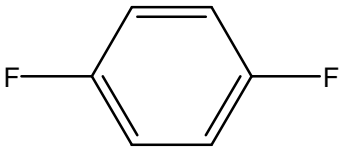
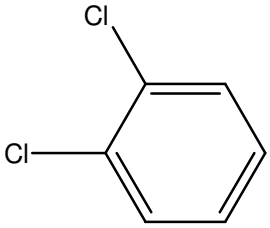
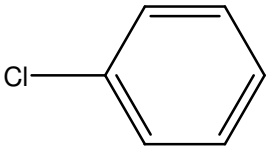
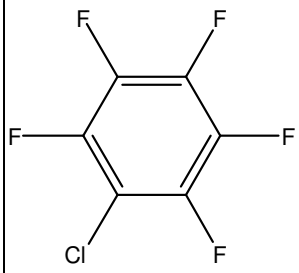
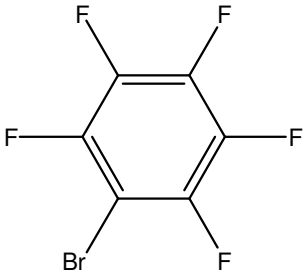
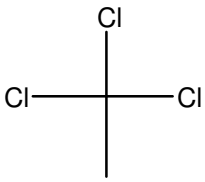
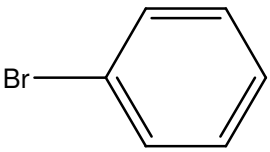
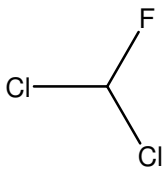
Annexe 2a : suite

n°	Formule chimique et CAS	Formule développée	n°	Formule chimique et CAS	Formule développée
15	C_2H_5F 353-36-6		21	$C_2H_4Cl_2$ 107-06-2	
16	$C_2Cl_2F_4$ 374-07-2		22	$C_2Br_2ClF_3$ 354-51-8	
17	C_2H_5Cl 75-00-3		23	C_3F_8 76-19-7	
18	$C_2Br_2F_4$ 124-73-2		24	$C_3H_3F_5$ 1814-88-6	
19	$C_2Cl_3F_3$ 76-13-1		25	C_3H_7Cl 540-54-5	
20	$C_2H_4Cl_2$ 75-34-3		26	$C_3H_5Cl_3$ 7789-89-1	

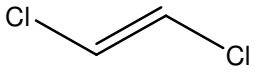
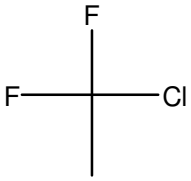
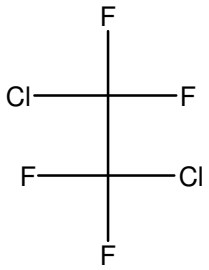
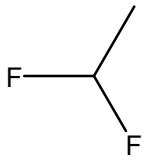
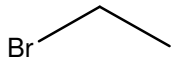
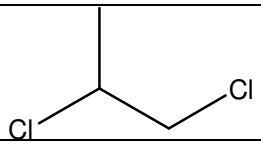
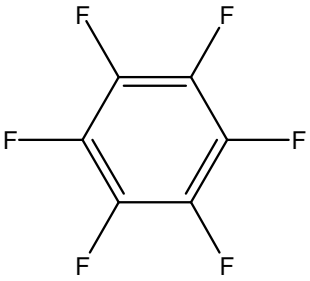
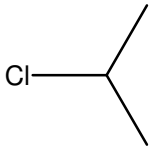
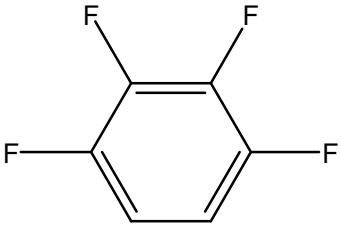
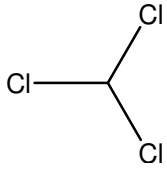
Annexe 2a: suite

n°	Formule chimique et CAS	Formule développée	n°	Formule chimique et CAS	Formule développée
27	C_4F_{10} 355-25-9		33	C_2ClF_3 79-38-9	
28	C_4H_9Cl 507-20-0		34	C_2HClF_2 359-10-4	
29	C_4H_9Cl 78-86-4		35	C_2HCl_3 79-01-6	
30	C_4H_9Cl 109-69-3		36	$C_3H_3F_3$ 677-21-4	
31	$C_2H_2F_2$ 75-38-7		37	C_3H_5Cl 107-05-1	
32	C_2F_4 116-14-3		38	$C_6H_2F_4$ 2367-82-0	

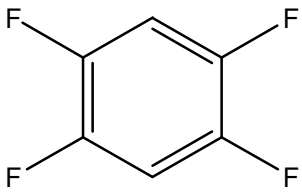
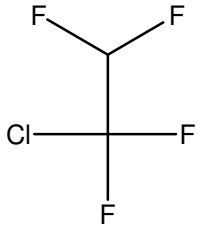
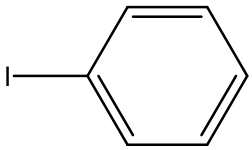
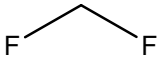
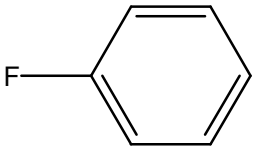
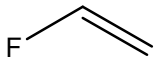
Annexe 2a : suite

n°	Formule chimique et CAS	Formule développée	n°	Formule chimique et CAS	Formule développée
39	C_6HF_5 363-72-4		44	$C_6Cl_2F_4$ 1198-59-0	
40	$C_6H_4F_2$ 540-36-3		45	$C_6H_4Cl_2$ 95-50-1	
41	C_6H_5Cl 108-90-7		46	C_6ClF_5 344-07-0	
42	C_6BrF_5 344-04-7		47	$C_2H_3Cl_3$ 71-55-6	
43	C_6H_5Br 108-86-1		48	$CHCl_2F$ 75-43-4	

Annexe 2a : suite

n°	Formule chimique et CAS	Formule développée	n°	Formule chimique et CAS	Formule développée
49	$C_2H_2Cl_2$ 156-60-5		54	$C_2H_3ClF_2$ 75-68-3	
50	$C_2Cl_2F_4$ 76-14-2		55	$C_2H_4F_2$ 75-37-6	
51	CH_3Br 74-83-9		56	$C_3H_6Cl_2$ 78-87-5	
52	C_6F_6 392-56-3		57	C_3H_7Cl 75-29-6	
53	$C_6H_2F_4$ 551-62-2		58	$CHCl_3$ 67-66-3	

Annexe 2a : suite et fin

n°	Formule chimique et CAS	Formule développée	n°	Formule chimique et CAS	Formule développée
59	$C_6H_2F_4$ 327-54-8		62	C_2HClF_4 354-25-6	
60	C_6H_5I 591-50-4		63	CH_2F_2 75-10-5	
61	C_6H_5F 462-06-6		64	C_2H_3F 75-02-5	

Annexe 2b: Valeurs des descripteurs du 2ème groupe.

n°	Composé	ω_{obs}	Log P	SNar	PCR	Mor23m
1	Tétrafluorométhane	0,177	2,198	1,386	1	-0,265
2	Trifluorométhane	0,26	1,054	1,099	1	-0,262
3	Chlorotrifluorométhane	0,198	2,49	1,386	1	0,09
4	Fluorométhane	0,187	0,338	0	1	-0,034
5	Bromotrifluorométhane	0,171	1,856	1,386	1	-0,809
6	Chlorodifluorométhane	0,221	1,24	1,099	1	0,121
7	Dichlorodifluorométhane	0,204	2,782	1,386	1	0,407
8	Chlorométhane	0,153	0,881	0	1	0,079
9	Bromochlorodifluorométhane	0,184	2,148	1,386	1	-0,534
10	Trichlorofluorométhane	0,189	3,074	1,386	1	0,722
11	Dichlorométhane	0,199	1,15	0,693	1	0,211
12	Tétrachlorométhane	0,193	3,366	1,386	1	1,071
13	1,1,1-Trifluoroéthane	0,251	1,481	1,386	1	-0,181
14	1-Chloro-1,1,2,2,2-pentafluoroéthane	0,279	2,314	2,773	1	-0,154
15	Fluoroéthane	0,215	0,68	0,693	1	-0,004

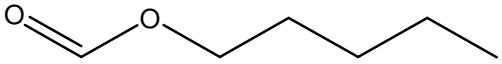
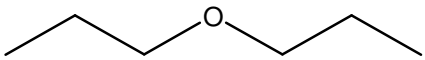
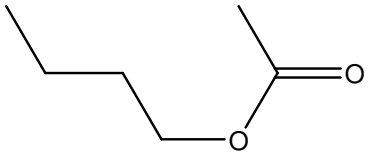
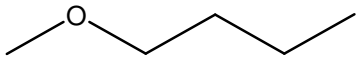
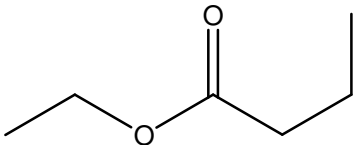
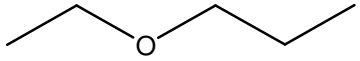
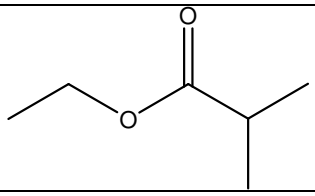
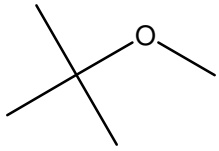
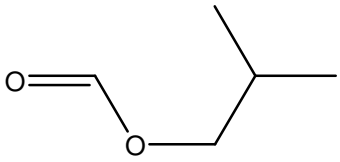
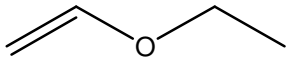
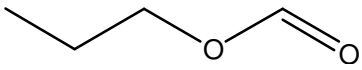
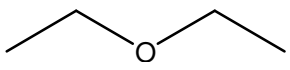
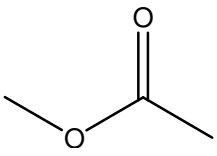
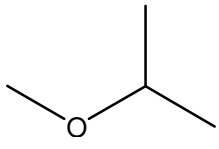
Annexe 2b: suite

n°	Composé	ω_{obs}	Log P	SNar	PCR	Mor23m
16	1,1-Dichlorotétrafluoroéthane	0,263	2,501	2,773	1	-0,052
17	Chloroéthane	0,191	1,224	0,693	1	0,055
18	1,2-Dibromo-1,1,2,2-tétrafluoroéthane	0,245	2,436	2,773	1	-0,771
19	1,1,2-Trichloro-1,2,2-trifluoroéthane	0,256	2,687	2,773	1	0,04
20	1,1-Dichloroéthane	0,24	1,163	1,099	1	0,187
21	1,2-Dichloroéthane	0,278	1,585	1,386	1	0,181
22	1,2-Dibromo-1-chloro-1,2,2-trifluoroéthane	0,248	2,622	2,773	1	-0,912
23	1,1,1,2,2,3,3,3-Octafluoropropane	0,325	2,962	4,159	1	-0,796
24	1,1,1,2,2-Pentafluoropropane	0,308	2,03	2,773	1	-0,292
25	1-Chloropropane	0,235	1,693	1,386	1	0,043
26	1,2,3-Trichloropropane	0,31	2,359	2,485	1	-0,163
27	1,1,1,2,2,3,3,4,4,4-Decafluorobutane	0,374	3,797	5,545	1	-1,057
28	2-Chloro-2-méthypropane	0,19	1,715	1,386	1	0,01
29	2-Chlorobutane	0,3	2,106	1,792	1	0,041
30	1-Chlorobutane	0,218	2,089	2,079	1	0,016
31	1,1-Difluoroéthène	0,14	1,24	1,099	1,093	-0,091
32	1,1,2,2-Tétrafluoroéthène	0,223	1,354	2,197	1,105	-0,084
33	1-Chloro-1,2,2-trifluoroéthène	0,252	1,646	2,197	1,105	0,052
34	2-Chloro-1,1-difluoroéthène	0,22	1,124	1,792	1,125	-0,14
35	1,1,2-Trichloroéthène	0,213	1,708	1,792	1,125	0,272
36	3,3,3-Trifluoropropène	0,238	1,893	2,079	1,074	-0,173
37	3-Chloropropène	0,13	1,622	1,386	1,128	0,032
38	1,2,3,5-Tétrafluorobenzène	0,346	2,605	5,781	1,303	-0,261
39	1,2,3,4,5-Pentafluorobenzène	0,373	2,744	6,186	1,278	-0,312
40	1,4-Difluorobenzène	0,299	2,326	4,97	1,32	-0,185
41	Chlorobenzène	0,249	2,565	4,564	1,369	-0,164
42	1-Bromo-2,3,4,5,6-pentafluorobenzène	0,355	3,536	6,592	1,255	-0,545
43	Bromo-benzène	0,251	2,838	4,564	1,369	-0,313
44	1,2-Dichloro-3,4,5,6-tétrafluorobenzène	0,622	-1,25	6,592	1,255	0,433
45	1,2-Dichlorobenzène	0,272	3,083	4,97	1,378	-0,048
46	Chloropentafluorobenzène	0,4	3,262	6,592	1,255	-0,246
47	1,1,1-Trichloroéthane	0,217	2,04	1,386	1	0,541
48	Dichloro-fluorométhane	0,21	1,427	1,099	1	0,35
49	Trans-1, 2-dichloroéthylene	0,232	0,894	1,386	1,159	0,103
50	1,2-Dichloro-1,1,2,2-Tétrafluoroéthane	0,246	2,501	2,773	1	-0,153
51	Bromoéthane	0,229	1,287	0,693	1	-0,091
52	1,2,3,4,5,6-hexafluorobenzène	0,396	2,884	6,592	1,255	-0,35
53	1,2,3,4-Tétrafluorobenzène	0,344	2,605	5,781	1,305	-0,287
54	1-Chloro-1,1-difluoroéthane	0,251	1,667	1,386	1	0,137
55	1,1-Difluoroéthane	0,256	0,477	1,099	1	-0,04
56	1,2-Dichloropropane	0,24	1,998	1,792	1	-0,026
57	2-Chloropropane	0,232	1,637	1,099	1	0,041
58	Trichlorométhane	0,218	1,613	1,099	1	0,445
59	1,2,4,5-Tétrafluorobenzène	0,355	2,605	5,781	1,303	-0,265

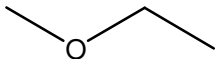
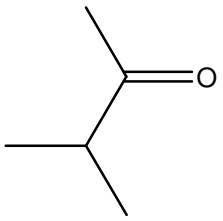
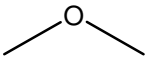
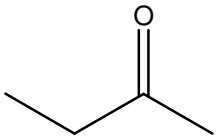
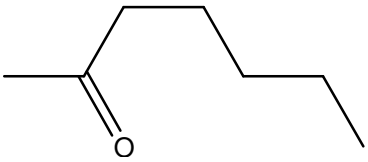
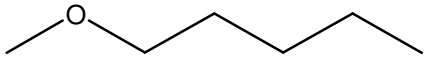
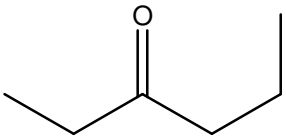
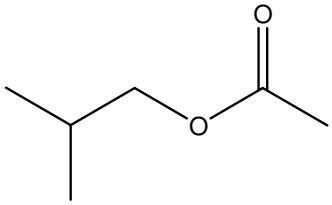
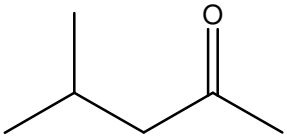
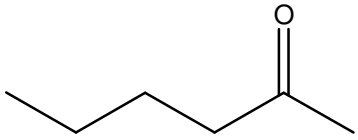
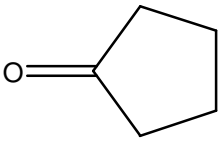
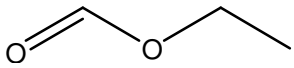
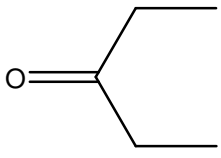
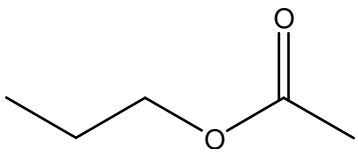
Annexe 2b: suite et fin

n°	Composé	ω_{obs}	Log P	SNar	PCR	Mor23m
60	Iodobenzène	0,249	2,114	4,564	1,369	0,005
61	Fluorobenzène	0,244	2,186	4,564	1,369	-0,188
62	1-Chloro-1,1,2,2-tétrafluoroéthane	0,281	1,596	2,485	1	-0,155
63	Difluorométhane	0,271	0,464	0,693	1	-0,11
64	Fluoroéthène	0,157	0,632	0,693	1,145	-0,053

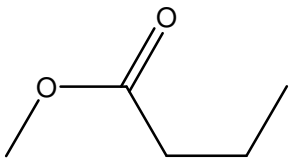
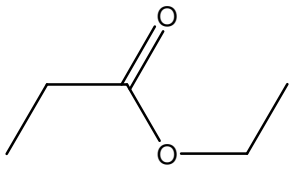
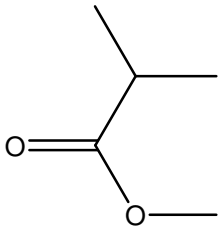
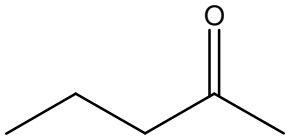
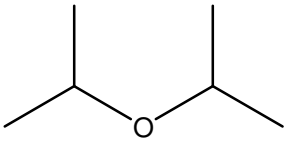
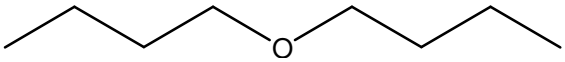
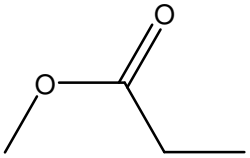
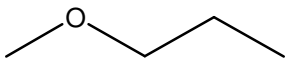
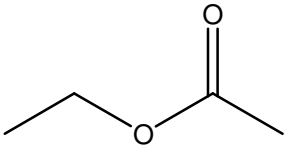
Annexe 3a : Composés du 3ème groupe.

n°	Formule chimique et CAS	Formule développée	n°	Formule chimique et CAS	Formule développée
1	$C_6H_{12}O_2$ 638-49-3		8	$C_6H_{14}O$ 108-20-3	
2	$C_6H_{12}O_2$ 123-86-4		9	$C_5H_{12}O$ 628-28-4	
3	$C_6H_{12}O_2$ 105-54-4		10	$C_5H_{12}O$ 628-32-0	
4	$C_6H_{12}O_2$ 97-62-1		11	$C_5H_{12}O$ 1634-04-4	
5	$C_5H_{10}O_2$ 542-55-2		12	C_4H_8O 109-92-2	
6	$C_4H_8O_2$ 110-74-7		13	$C_4H_{10}O$ 60-29-7	
7	$C_4H_8O_2$ 110-74-7		14	$C_4H_{10}O$ 598-53-8	

Annexe 3a : suite

n°	Formule chimique et CAS	Formule développée	n°	Formule chimique et CAS	Formule développée
15	C_3H_8O 540-67-0		22	$C_5H_{10}O$ 563-80-4	
16	C_2H_6O 115-10-6		23	C_4H_8O 78-93-3	
17	$C_7H_{14}O$ 110-43-0		24	$C_6H_{14}O$ 628-80-8	
18	$C_6H_{12}O$ 589-38-8		25	$C_6H_{12}O_2$ 110-19-0	
19	$C_6H_{12}O$ 108-10-1		26	$C_6H_{12}O$ 591-78-6	
20	C_5H_8O 120-92-3		27	$C_3H_6O_2$ 109-94-4	
21	$C_5H_{10}O$ 96-22-0		28	$C_5H_{10}O_2$ 109-60-4	

Annexe 3a : suite et fin

n°	Formule chimique et CAS	Formule développée	n°	Formule chimique et CAS	Formule développée
29	$C_5H_{10}O_2$ 623-42-7		34	$C_5H_{10}O_2$ 105-37-3	
30	$C_5H_{10}O_2$ 623-42-7		35	$C_5H_{10}O$ 107-87-9	
31	$C_6H_{14}O$ 108-20-3		36	$C_8H_{18}O$ 142-96-1	
32	$C_4H_8O_2$ 554-12-1		37	$C_4H_{10}O$ 557-17-5	
33	$C_4H_8O_2$ 141-78-6				

Annexe 3b : Valeurs des descripteurs du 3ème groupe.

n°	Composé	ω_{obs}	GMTIV	BELm2	Mor20v
1	Formiate de pentyle	0,538	678	1,541	-0,119
2	Ethanoate de n-butyle	0,417	574	1,552	0,002
3	Butanoate d'éthyle	0,461	510	1,642	0,023
4	Isobutyrate d'éthyle	0,431	470	1,635	0,026
5	Formiate d'isobutyle	0,396	445	1,261	-0,049
6	Formiate de propyle	0,314	328	1,259	0,003
7	Ethanoate de méthyle	0,326	183	1,366	0,128
8	1-Propoxypropane	0,369	218	1,707	0,027
9	1-Méthoxybutane	0,316	153	1,487	0,013
10	1-Ethoxypropane	0,333	137	1,62	0,077
11	2-Méthoxy-2-méthylpropane	0,269	101	1,385	0,064
12	Ethoxyéthène	0,268	115	1,487	0,133
13	Ethoxyéthane	0,281	76	1,562	0,12
14	2-Méthoxypropane	0,266	68	1,381	0,113
15	Méthoxyéthane	0,244	39	1,372	0,125
16	Méthoxy-méthane	0,2	14	1,249	0,128
17	Heptan-2-one	0,483	438	1,65	-0,234
18	Hexan-3-one	0,378	261	1,637	-0,188
19	4-Méthylpentan-2-one	0,385	265	1,542	-0,207
20	Cyclopentanone	0,35	228	1,616	-0,024
21	Pentan-3-one	0,344	168	1,562	-0,145
22	3-Méthylbutan-2-one	0,331	160	1,437	-0,116
23	Butan-2-one	0,32	107	1,412	-0,088
24	1-Méthoxypentane	0,347	250	1,567	-0,051
25	Ethanoate d'isobutyle	0,455	534	1,54	-0,017
26	Hexan-2-one	0,392	293	1,586	-0,186
27	Formiate d'éthyle	0,285	215	1,116	0,057
28	Ethanoate de n-propyle	0,391	401	1,533	0,058
29	Butanoate de méthyle	0,38	385	1,449	0,048
30	Isobutanoate de méthyle	0,362	349	1,425	0,055
31	2-Isopropoxypropane	0,331	170	1,666	0,132
32	Propionate de méthyle	0,35	264	1,419	0,106
33	Ethanoate d'éthyle	0,362	272	1,504	0,11
34	Propanoate d'éthyle	0,391	369	1,618	0,08
35	Pentan-2-one	0,346	184	1,511	-0,131
36	1-Butoxybutane	0,502	472	1,787	-0,109
37	1-Méthoxypropane	0,271	84	1,426	0,077

Annexe 4 : Valeurs des descripteurs pour la comparaison des méthodes.

N_{Grp}	n°	Composé	ω_{obs}	Hy	HIC	RDF050m	ESpm09d	J
2-1	1	Tétrafluorométhane	0,177	-0,18	2	0	13,557	3,024
2-2	2*	Trifluorométhane	0,26	-0,215	2,012	0	11,301	2,324
2-3	3	Chlorotrifluorométhane	0,198	-0,18	2,003	0	13,582	3,024
2-4	4	Fluorométhane	0,187	-0,315	2,003	0	3,733	1
2-5	5*	Bromotrifluorométhane	0,171	-0,18	2,007	0	13,542	3,024
2-6	6	Chlorodifluorométhane	0,221	-0,215	2,013	0	11,344	2,324
2-7	7	Dichlorodifluorométhane	0,204	-0,18	2,005	0	13,254	3,024
2-8	8	Chlorométhane	0,153	-0,315	2,018	0	4,02	1
2-9	9	Bromochlorodifluorométhane	0,184	-0,18	2,008	0	13,567	3,024
2-10	10*	Trichlorofluorométhane	0,189	-0,18	2,006	0	12,529	3,024
2-11	11	Dichlorométhane	0,199	-0,264	2,021	0	7,097	1,633
2-12	12	Tétrachlorométhane	0,193	-0,18	2	0	12,086	3,024
2-13	13	1,1,1-Trifluoroéthane	0,251	-0,359	2,743	0	12,96	3,024
2-14	14	1-Chloro-1,1,2,2,2-pentafluoroéthane	0,279	-0,237	2,748	0	14,133	4,02
2-15	15*	Fluoroéthane	0,215	-0,528	2,765	0	6,276	1,633
2-16	16	1,1-Dichlorotétrafluoroéthane	0,263	-0,237	2,737	0	13,972	4,02
2-17	17	Chloroéthane	0,191	-0,528	2,747	0	6,454	1,633
2-18	18	1,2-Dibromo-1,1,2,2-tétrafluoroéthane	0,245	-0,237	2,75	0,177	14,104	4,02
2-19	19	1,1,2-Trichloro-1,2,2-trifluoroéthane	0,256	-0,237	2,743	0	13,987	4,02
2-20	20*	1,1-Dichloroéthane	0,24	-0,431	2,715	0	9,557	2,324
2-21	21	1,2-Dichloroéthane	0,278	-0,431	2,764	0	7,882	1,975
2-22	22	1,2-Dibromo-1-chloro-1,2,2-trifluoroéthane	0,248	-0,237	2,734	0	14,117	4,02
2-23	23	1,1,1,2,2,3,3,3-Octafluoropropane	0,325	-0,263	3,249	2,08	14,465	4,748
2-24	24	1,1,1,2,2-Pentafluoropropane	0,308	-0,355	3,242	0	13,847	4,02
2-25	25*	1-Chloropropane	0,235	-0,646	3,256	0,195	6,895	1,975
2-26	26	1,2,3-Trichloropropane	0,31	-0,46	3,212	0	10,007	2,754
2-27	27	1,1,1,2,2,3,3,4,4,4-Decafluorobutane	0,374	-0,278	3,595	1,046	14,723	5,3
2-28	28	2-Chloro-2-méthypropane	0,19	-0,719	3,606	0	11,341	3,024
2-29	29	2-Chlorobutane	0,3	-0,719	3,581	0,019	9,265	2,54
2-30	30*	1-Chlorobutane	0,218	-0,719	3,6	0,01	6,996	2,191
2-31	31	1,1-Difluoroéthène	0,14	-0,431	2,257	0	10,315	2,324
2-32	32	1,1,2,2-Tétrafluoroéthène	0,223	-0,307	2,253	0	11,605	2,993
2-33	33	1-Chloro-1,2,2-trifluoroéthène	0,252	-0,307	2,237	0	11,63	2,993
2-34	34	2-Chloro-1,1-difluoroéthène	0,22	-0,359	2,241	0	10,637	2,54
2-35	35*	1,1,2-Trichloroéthène	0,213	-0,359	2,202	0	10,024	2,54
2-36	36	3,3,3-Trifluoropropène	0,238	-0,46	2,798	0	13,284	3,168
2-37	37	3-Chloropropène	0,13	-0,646	2,8	0,051	7,72	1,975
2-38	38	1,2,3,5-Tétrafluorobenzène	0,346	-0,576	3,338	0,006	11,421	2,487
2-39	39	1,2,3,4,5-Pentafluorobenzène	0,373	-0,526	3,331	0,007	11,857	2,625
2-40	40*	1,4-Difluorobenzène	0,299	-0,71	3,358	0,014	10,198	2,192
2-41	41	Chlorobenzène	0,249	-0,802	3,341	0,125	9,57	2,123
2-42	42	1-Bromo-2,3,4,5,6-pentafluorobenzène	0,355	-0,484	3,294	0,119	12,169	2,76
2-43	43	Bromo-benzène	0,251	-0,802	3,329	1,063	9,45	2,123
2-44	44	1,2-Dichloro-3,4,5,6-tétrafluorobenzène	0,622	-0,484	3,272	17,408	12,196	2,76

Annexe 4 : suite

N_{Grp}	n°	Composé	ω_{obs}	Hy	HIC	RDF050m	ESpm09d	J
2-45	45*	1,2-Dichlorobenzène	0,272	-0,71	3,309	0,362	10,726	2,279
2-46	46	Chloropentafluorobenzène	0,4	-0,484	3,305	4,931	12,185	2,76
2-47	47	1,1,1-Trichloroéthane	0,217	-0,359	2,676	0	11,656	3,024
2-48	48	Dichloro-fluorométhane	0,21	-0,215	2,015	0	10,774	2,324
2-49	49	Trans-1, 2-dichloroéthylene	0,232	-0,431	2,248	0	7,882	1,975
2-50	50*	1,2-Dichloro-1,1,2,2-Tétrafluoroéthane	0,246	-0,237	2,753	0	14,146	4,02
2-51	51	Bromoéthane	0,229	-0,528	2,74	0	6,168	1,633
1-15	52	3-Méthylphénol	0,454	-0,158	3,432	0,801	9,293	2,231
1-2	53	4-Méthylphénol	0,505	-0,158	3,466	0,059	9,226	2,192
1-3	54	2-Méthylphénol	0,433	-0,158	3,458	0,029	9,756	2,279
1-5	55*	2,3,4,5,6-Pentafluorophénol	0,502	0,079	3,438	0,046	12,079	2,76
1-6	56	Octan-1-ol	0,587	-0,213	4,541	2,852	5,21	2,595
1-7	57	Heptan-1-ol	0,56	-0,158	4,368	2,444	5,21	2,53
1-8	58	Cyclohexanol	0,528	-0,088	4,11	0,031	8,613	2,123
1-9	59	Hexan-1-ol	0,56	-0,088	4,182	2,011	5,21	2,447
1-13	60*	Butan-1-ol	0,593	0,132	3,702	0,98	5,167	2,191
1-12	61	2-Méthyl-propan-1-ol	0,592	0,132	3,732	0	7,391	2,54
1-14	62	2-Méthyl-propan-2-ol	0,612	0,132	3,749	0	10,617	3,024
1-15	63	Propan-1-ol	0,623	0,323	3,371	0	4,875	1,975
1-16	64	Propan-2-ol	0,665	0,323	3,395	0	7,674	2,324
1-17	65*	Ethanol	0,644	0,638	2,938	0	3,782	1,633
1-18	66	Méthanol	0,556	1,262	2,32	0	0,229	1
3-1	67	Formiate de pentyle	0,538	-0,71	4,087	1,975	9,975	2,53
3-2	68	Ethanoate de n-butyle	0,417	-0,71	4,054	1,521	11,598	2,716
3-3	69	Butanoate d'éthyle	0,461	-0,71	4,107	1,071	11,729	2,92
3-4	70*	Isobutyrate d'éthyle	0,431	-0,71	4,125	0,101	11,959	3,171
3-5	71	Formiate d'isobutyle	0,396	-0,668	3,873	1,497	10,063	2,678
3-6	72	Formiate de propyle	0,314	-0,614	3,543	0,421	9,972	2,339
3-7	73	Ethanoate de méthyle	0,326	-0,539	3,166	0,009	11,569	2,54
3-8	74	1-Propoxypropane	0,369	-0,802	4,145	1,121	7,77	2,447
3-9	75*	1-Méthoxybutane	0,316	-0,767	3,948	1,434	7,1	2,339
3-10	76	1-Ethoxypropane	0,333	-0,767	3,917	0,596	7,635	2,339
3-11	77	2-Méthoxy-2-Méthylpropane	0,269	-0,767	4,014	0,015	11,011	3,168
3-12	78	Ethoxyéthène	0,268	-0,719	3,328	0,131	7,479	2,191
3-13	79	Ethoxyéthane	0,281	-0,719	3,634	0,054	7,479	2,191
3-14	80*	2-Méthoxypropane	0,266	-0,719	3,714	0,008	8,739	2,54
3-15	81	Méthoxyéthane	0,244	-0,646	3,348	0	6,776	1,975
3-16	82	Méthoxy-méthane	0,2	-0,528	2,983	0	5,589	1,633
3-17	83	Heptan-2-one	0,483	-0,828	4,226	0,937	11,432	2,716
3-18	84	Hexan-3-one	0,378	-0,802	4,011	0,499	11,565	2,832
3-19	85*	4-Méthylpentan-2-one	0,385	-0,802	4,053	0,477	11,468	2,953
3-20	86	Cyclopentanone	0,35	-0,767	3,61	0	11,604	2,184
3-21	87	Pentan-3-one	0,344	-0,767	3,749	0,275	11,553	2,754
3-22	88	3-Méthylbutan-2-one	0,331	-0,767	3,818	0,013	11,701	2,993

Annexe 4 : suite et fin

N_{Grp}	n°	Composé	ω_{obs}	Hy	HIC	RDF050m	ESpm09d	J
3-23	89	Butan-2-one	0,32	-0,719	3,47	0,025	11,418	2,54
3-24	90*	1-Méthoxypentane	0,347	-0,802	4,165	2,073	7,107	2,447
3-25	91	Ethanoate d'isobutyle	0,455	-0,71	4,104	1,082	11,617	2,928
3-26	92	Hexan-2-one	0,392	-0,802	4,019	0,512	11,432	2,678
3-27	93	Formiate d'éthyle	0,285	-0,539	3,155	0,003	9,957	2,191
3-28	94	Ethanoate de n-propyle	0,391	-0,668	3,817	0,434	11,598	2,678
3-29	95*	Butanoate de méthyle	0,38	-0,668	3,868	1,061	11,706	2,832
3-30	96	Isobutanoate de méthyle	0,362	-0,668	3,89	0,071	11,94	3,144
2-52	97	1,2,3,4,5,6-hexafluorobenzène	0,396	-0,484	3,324	0,008	12,175	2,76
2-53	98	1,2,3,4-Tétrafluorobenzène	0,344	-0,576	3,339	0,006	11,582	2,516
2-54	99	1-Chloro-1,1-difluoroéthane	0,251	-0,359	2,723	0	12,991	3,024
2-55	100	1,1-Difluoroéthane	0,256	-0,431	2,758	0	10,315	2,324
2-56	101	1,2-Dichloropropane	0,24	-0,539	3,219	0,245	9,597	2,54
2-57	102	2-Chloropropane	0,232	-0,646	3,231	0	9,011	2,324
2-58	103	Trichlorométhane	0,218	-0,215	2,012	0	9,363	2,324
2-59	104	1,2,4,5-Tétrafluorobenzène	0,355	-0,576	3,34	0,008	11,393	2,462
2-60	105	Iodobenzène	0,249	-0,802	3,301	0,732	9,176	2,123
2-61	106	Fluorobenzène	0,244	-0,802	3,366	0,015	9,495	2,123
2-62	107	1-Chloro-1,1,2,2-tétrafluoroéthane	0,281	-0,267	2,749	0	13,386	3,541
2-63	108	Difluorométhane	0,271	-0,264	2,008	0	8,283	1,633
2-64	109	Fluoroéthène	0,157	-0,528	2,27	0	6,276	1,633
1-10	110	Pentan-1-ol	0,579	0,004	3,954	1,626	5,21	2,339
1-4	111	Phénol	0,438	-0,088	3,474	0,014	8,613	2,123
1-13	112	Butan-2-ol	0,577	0,132	3,722	0,02	8,105	2,54
3-31	113	2-Isopropoxy-propane	0,331	-0,802	4,198	0,356	9,635	2,953
3-32	114	Propionate de méthyle	0,35	-0,614	3,56	0,011	11,696	2,754
3-33	115	Ethanoate d'éthyle	0,362	-0,614	3,5	0,018	11,594	2,627
3-34	116	Propanoate d'éthyle	0,391	-0,668	3,856	0,021	11,719	2,832
3-35	117	Pentan-2-one	0,346	-0,767	3,758	0,349	11,431	2,627
3-36	118	1-Butoxybutane	0,502	-0,848	4,5	2,99	7,825	2,595
3-37	119	1-Méthoxypropane	0,271	-0,719	3,666	0,491	7,043	2,191

La colonne N_{Grp} est la numérotation des composés dans leur groupe respectif par, exemple le Cyclopentanone a le n° 86 quand il est utilisé avec la totalité des composés, et le n° 20 quand il sert a la modélisation avec son groupe d' où le 3-20.

Les 20 composés afféctés d'un astérisque appartiennent à l'ensemble test du réseau de neuronnes artificiel.