RÉPUBLIQUE ALGÉRIENNE DÉMOCRATIQUE ET POPULAIRE
MINISTÈRE DE L'ENSEIGNEMENT SUPÉRIEUR
ET DE LA RECHERCHE SCIENTIFIQUE

UNIVERSITÉ ABOU BEKR BELKAÏD DE TLEMCEN
FACULTÉ DES SCIENCES
DÉPARTEMENT DE CHIMIE

LABORATOIRE DE THERMODYNAMIQUE APPLIQUÉE ET MODÉLISATION MOLÈCULAIRE(LTA2M)

THÈSE

En vue de l'obtention du titre de
Docteur en Science
Option : Chimie Théorique et Modélisation Moléculaire

Présentée par
Mr. Abdelkrim GUENDOUZI

Élaboration des modèles QSPR prédictifs des propriétés physico-chimiques à l'aide des descripteurs moléculaires.

Soutenu publiquement le : 28 / 09 / 2015 devant le jury composé de :

Mr Benamar DAHMANI	Professeur	Université A. Belkaïd - Tlemcen	Président
Mr Lotfi BELKHIRI	Professeur	Université Constantine 1.	Examinateur
Mr Djamel-Eddine KHATMI	Professeur	Université 8 mai 45 - Guelma	Examinateur
Mlle Latifa NEGADI	Professeur	Université A. Belkaïd - Tlemcen	Examinateur
Mr Sidi Mohamed MEKELLECHE	Professeur	Université A. Belkaïd - Tlemcen	Directeur de Thèse

بسم الله الرحمن الرحيم

*** و قل ربي زدني علما ***

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

Je dédie la présente thèse :

* A ma chère mère pour son dévouement et sa compréhension.

* A mes frères et mes soeurs.

* A ma femme et à mon fils Berrezoug.

* A mes chèrs amis, Ouici, Brahim, Hadji, Ariche....

* A mes collègues de laboratoire LTA2M, Charif, Messouadi, Berrahoui, Taki, Bellifa, Benchouk, Chemouri, Mahi,

* A la mémoire de mon père, et mon frère Berrezoug.

* A la mémoire de défunt BENHABIB Réda

REMERCIEMENTS

Les travaux du présent manuscrit ont été réalisés au Laboratoire de Thermodynamique Appliquée et Modélisation Moléculaire (LATA2M), de l'université Abou Bekr Belkaid de Tlemcen,

Mes premiers remerciements sont adressés à mon directeur de thèse, le professeur Sidi Mohamed Mekelleche. D'abord, pour m'avoir accueilli dans son équipe et m'avoir permis de travailler sur un sujet aussi passionnant.Ensuite pour m'avoir orienté et m'encadré avec efficacité tout en me laissant l'initiative et pour ses multiples conseils durant toute la durée qu'a nécessité la réalisation de ce projet de recherche. Enfin je ne le remercierai jamais assez pour la confiance qu'il m'a accordée et surtout pour sa grande patience .

J'exprime également ma profonde gratitude à monsieur Benamar Dahmani professeur à l'université Abou Bekr Belkaïd de Tlemcen, pour nous avoir fait l'honneur de présider le jury de cette thèse.

Mes vifs remerciements vont également à monsieur Lotfi Belkhiri professeur à l'université Constantine-1. et à monsieur Djameleddine Khatmi, professeur à l'université 8 mai 45 Guelma, pour l'honneur qu'ils nous ont fait en acceptant d'examiner notre travail, et de faire le déplacement à Tlemcen.

J'adresse mes sincères remerciements à mademoiselle Latifa NEGADI, professeur à l'université Abou Bekr Belkaïd de Tlemcen, pour l'honneur qu'elle nous a fait d'accepter d'évaluer notre travail.

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

Un grand merci à mes chers amis d'avoir partagé avec moi d'agréables moments. Je tiens à présenter ma reconnaissance et mes remerciements à ma famille, qui est ma source d'inspiration et mon plus grand soutien.

Table des matières

Table des figures

Liste des tableaux

ABRÉVIATIONS

B3LYP - Becke 3-Parameter Lee-Yang-Parr
CLOA - Combinaison Linéaire d'Orbitales Atomiques
CGTO - Contracted Gaussian Type Orbital Configuration
DFT - Density Functional Theory
FMO - Frontier Molecular Orbital
GGA - Generalized Gradient Approximation
GTO - Gaussian Type Orbital
HF - Hartree-Fock
HOMO - Highest Occupied Molecular Orbital
KS - Kohn, Sham
LDA - Local Density Approximation
LUMO - Lowest Unoccupied Molecular Orbital
MLR - Multiple Linear Regression
NAO - Natural Atomic Orbitals
NBO - Natural Bond Orbitals
NPA - Natural Population Analysis
OA - Orbitale Atomique
OM - Orbitale Moléculaire
PCM - Polarizable Continuum Model
PM6 - Parametric Method 6
QSPR - Quantitative Structure-Property Relationships
QSAR - Quantitative Structure-Activity Relationships
SCF - Self Consistent Field
STO - Slater Type Orbital
SCRF - Self-Consistent Reaction Field

INTRODUCTION GÉNÉRALE

Le développement spectaculaire au cours de ces dernières décennies des moyens informatiques et des machines de calcul (vitesse, espace disque, mémoire vive, . . .) alliés aux avancées des méthodes de chimie computationnelle [1], permettent aujourd'hui à la modélisation moléculaire de traiter de nombreux types de problèmes. La modélisation moléculaire est une technique permettant, non seulement de représenter les structures en deux ou trois dimensions mais aussi à aider tout chimiste théoricien, de réaliser des études exhaustives et précises des systèmes moléculaires, et de représenter d'une façon explicite les modèles sous une forme mathématique.

Actuellement, la modélisation moléculaire est largement utilisée [2] pour élaborer des modèles fiables permettant de prédire les propriétés physico-chimiques et les activités biologiques. L'une de ces techniques est la modélisation QSAR/QSPR *(Quantitative Structure-Activity/Property Relationships)* [3] qui permet de prédire les propriétés/activités des systèmes chimiques à partir de leurs structures moléculaires.

Les premiers développements de ces méthodes sont anciens. En **1868**, *Crum-Brown* et *Fraser* [4] ont mis en evidence la relation qui lie certaines activités physiologiques et les structures chimiques. En **1937**, *Hammett* a élaboré ces modèles qui caractérisent les vitesses de réactions pour des composés organiques [5]. En **1964**, *Hansch* [6], et *Free* [7] ont développé de nouvelles méthodes pour prédire l'activité biologique de certains composés avec les propriétés hydrophobes, électroniques et stériques à l'échelle moléculaire.

Dans de nombreux domaines, la prédiction des propriétés physico-chimiques ou activités biologiques des molécules présentent un enjeu industriel important car elle permet de réduire les délais et les coûts de productions.

Les méthodes QSAR/QSPR sont communément employées dans la littérature et représentent un sous-domaine important de la *chemo-informatique*. Ces techniques servent à prédire plusieurs propriétés et activités biologiques tels que :

a) Des propriétés physico-chimiques :

• Prédiction de la solubilité aqueuse, points de fusion, températures d'ébullition, températures critiques, indices de rétention, indices de réfraction, viscosité, coefficient de

partage eau-air, constantes d'acidité [8–14].

• Prédiction de l'indice de réfraction des polymères [15].

• Prédiction de la concentration micellaire critique des composées anioniques [16].

• Prédiction des points de fusion des benzènes substitués [17].

• Prédiction des hyperpolarisabilités [18].

• Prédiction des points de fusion et la viscosité des liquides ioniques [20].

• Prédiction des absorbances spectrales ultraviolets [21].

• Prédiction de la solubilité aqueuse de médicaments [22].

• Prédiction de déplacements chimiques ^{13}C [23].

• Prédiction des coefficients de partage « air-to-blood » [24].

b) Des activités biologiques ;

• Prédictions de la toxicité [25].

• Prédictions de l'activité anti-HIV [26], anti-malaria [27], anti-inflammatoire [28], anti-Alzheimer [29], Anti-cancer [30], anti-microbien [31], antibactérienne [32], anti-tumorale [33], anti-prolifératif [34] et anti-hépatocellulaire [35] de différentes familles de composés chimiques.

L'objectif principal de ce travail est d'appliquer la modélisation QSPR pour développer des modèles fiables pour prédire deux propriétés physico-chimiques telles que le point de fusion des acides gras [36] et la constante d'acidité des acides benzoïques.

Les acides gras sont des molécules organiques appartenant à la catégorie des lipides, leurs fonctions biologiques sont très variées chez l'être humain [37]. Les acides gras se caractérisent par la longueur de leur chaîne carbonée (nombre d'atomes de carbone), ainsi que par le nombre et la position des doubles liaisons carbone-carbone non saturées sur cette chaîne, les variations de ces caractéristiques permettent de distinguer plusieurs types d'acides gras et répondent à une nomenclature précise.

Les acides gras remplissent plusieurs fonctions dans l'organisme. Ils constituent la matière grasse des êtres vivants. Ce sont une source d'énergie importante pour les organes. Ils sont stockés sous forme de triglycérides dans les tissus adipeux, et dans certains cas, ils peuvent servir à la synthèse d'autres lipides, notamment les phospholipides qui forment les membranes autour des cellules et des organites, et d'autre part, ils sont des précurseurs de plusieurs messagers intra - et extracellulaires [38]. Les lipides peuvent se présenter à l'état solide comme les cires, ou bien liquide comme les huiles, et leur nature amphiphile conduit les molécules de certains lipides à s'organiser en vésicules, liposomes et micelles lorsqu'elles se trouvent en milieu

aqueux [39–41].
On note T_f la température de fusion, la température à laquelle une substance passe de l'état solide à l'état liquide. Cette propriété est à la base du vivant, permettant la formation de structures biologiques. Diverses techniques permettent de mesurer cette propriété, une des plus courantes est l'utilisation d'un $banc\ Kofler$. Cet appareil est constitué d'une plaque chauffante avec un gradient de température et d'une échelle de température. La mesure est rapide et précise à plus ou moins un degré Celsius mais ne convient que pour des substances dont la température T_f varie entre 50 et 250 °C. Pour faire face à cet inconvénient, c'est a dire déterminer des points de fusion aux composés sur laquelle ces valeurs sont dessous de 50 °C, nous avons développé dans la première application des modèles (QSPR) pour prédire les températures de fusion d'une série constituée de 62 acides gras dont les valeurs expérimentales comprises entre -65.00 et 96.00 °C [42]. Nous avons utilisé la méthode dite Best Multi-Linear Regression (BMLR) implémentée dans le programme CODESSA [43] pour développer des modèles avec cinq descripteurs moléculaires [44].

Dans la deuxième application, des modèles QSPR fiables ont été élaborés pour corréler les valeurs expérimentales aqueuses de pKa de 52 acides benzoïques [42] avec la variation de l'énergie libre de Gibbs de déprotonation en phase aqueuse ΔG_{aq}. L'acide benzoïque est un acide organique qui contient un groupe carboxyle lié directement à un noyau de benzène (C_6H_5COOH). Les acides benzoïques sont présents dans la plupart des fruits, en particulier les baies et les canneberges. Ces acides sont naturellement présents dans certaines plantes médicinales qui sont utilisées en pharmacie. L'acide benzoïque est essentiellement utilisé comme conservateur et comme additif alimentaire qui empêche la croissance de la levure et de certaines bactéries. L'acide benzoïque est produit de manière industrielle à partir du toluène. La détermination des valeurs de constantes d'acidité des acides benzoïques est essentielle pour expliquer de nombreuses réactions fondamentales de la biochimie, Dans cette application, nous avons utilisé la théorie de la fonctionnelle de la densité au niveau B3LYP/6-311++G(d, p) pour calculer les énergies libres de Gibbs de déprotonation en solution aqueuse de 52 acides benzoïques et leurs anions correspondants. Nous avons étudié deux cycles thermodynamiques pour prédire les constantes d'acidité. Le pouvoir prédictif des modèles obtenues a été testé avec succès sur une série de tests constituée de 25 acides non inclus dans la série d'apprentissage.

Le travail présenté dans ce manuscrit est divisé en cinq chapitres :

Dans le premier chapitre, nous présenterons des généralités sur les méthodes QSPR et quelques considérations historiques. Nous rappellerons également les principaux types de descripteurs, ainsi que les techniques d'apprentissage et de validation du modèle QSPR.

Dans le second chapitre, nous décrirons les méthodes statistiques d'analyses de données employées dans ce travail.

Dans le troisième chapitre, nous présenterons brièvement les différentes méthodes de la chimie quantique.

Dans le quatrième et le cinquième chapitre, nous présenterons les résultats obtenues et leurs discussions

Enfin, nous clôturons cette recherche par une conclusion générale et des perspectives de recherche.

Bibliographie

[1] Andrew L., Molecular Modelling : Principles and Applications, (**2001**), Prentice Hall, 2nd Ed.

[2] Frank Jensen., Introduction to Computational Chemistry, (**2007**), Wiley, 2nd edition.

[3] Charton M., B. Charton., Advances in Quantitative Structure-Property Relationships, (**1999**), Vol 2. JAI PRESS INC.

[4] Crum-Brown A., Thomas R. Fraser. J. Anat. Physiol., 2(2), (**1868**), 224.

[5] Hammett Louis P. J. Am. Chem. Soc., 59 (1), (**1937**), 96.

[6] Hansch C., Fujita, T., J. Am. Chem. Soc., 86, (**1964**), 1616.

[7] Free SM, Jr Wilson J. W. J Med Chem., 7, (**1964**), 395.

[8] Katritzky A. R., P. D. T. Huibers, J. Chem. Inf. Comput. Sci., 38, (**1998**), 283.

[9] Goodarzi M., T. Chen, Matheus P. Freitas, Chem Int Lab Sys., 104, (**2010**), 2.

[10] Danial A, M. A. Sobati, Int. J. Ref., 40, (**2014**), 282.

[11] Sobati M. A., Danial A., Thermochimica Acta, 602, (**2015**), 53.

[12] Cristian R., P. R. Duchowicz, Reinaldo P. D. Chem. Int. Lab. Sys., 140, (**2015**), 15, 126.

[13] Bor-Kuan C., Ming-Jyh L., Tzi-Yi Wu, H. Paul Wang, F. Ph. Eq., 350, (**2013**), 37

[14] Li L., Qiang W., Xinghua Q., Yian D., Shenglan J. J. Haz. Mat., 276, (**2014**), 278

[15] Katritzky A. R., M. Karelson, J. Chem. Inf. Comp. Sci., 38 (6), (**1998**), 1171.

[16] Katritzky A. R., P. D. T. Huibers, V.S. Shah, M. Karelson, J. Col. Int. Sci., 187, (**1997**), 113.

[17] Katritzky A. R., U. Maran, V.S. Lobanov, J. Chem. Inf. Comput. Sci., 37, (**1997**), 913.

[18] Katritzky A. R., L. Pacureanu, D. Dobchev, M. Karelson, J. Mol. Mod., 13, (**2007**), 9, 951

[19] Yu G., L. Wen, D. Zhao, C. Asumana, X .Chen., J. Mol. Liq., 184, (**2013**), 51.

[20] Farahani N., F. Gharagheizi, S. A. Mirkhani, Thermochimica Acta., 549, (**2012**), 10, 17.

[21] Fitch W. L., M. McGregor, A. R. Katritzky, J. Chem. Inf. Comput. Sci., 42 (4), (**2002**), 830.

[22] Shayanfar A., M. A. A. Fakhree, A. Jouyban, J Drug Deliv Sci Tec., 20, (**2010**), 6, 467

[23] Goodarzi M., M. P. Freitas, T. C. Ramalho, Spectroc. Acta A., 74, (**2009**), 2, 1, 563

[24] Konoz E., H. Golmohammadi., Anal. Chim. Acta., 619, (**2008**), 2, 7 157

[25] Katritzky A. R., D. B. Tatham, U. Maran., J. Chem. Inf. Comput. Sci., 41, (**2001**), 1162.

[26] Darnag R., E.L. M. Mazouz, D. Villemin, Eur. J. Med. Chem., 45, (**2010**), 4, 1590.

[27] Apilak W., Chanin N., C. I. Ayudhya, Virapong P., Chem Pap., 67, (**2013**), 11 , 1462

[28] Lokwani D. K., Santosh N. M., Devanand B. S., Eur. J. Med. Chem., 73, (**2014**), 233.

[29] Kamlendra S. B., Mukesh C., Smita S., Shailesh V., Arab. J. Chem., 7, (**2014**), 6, 924.

[30] Sabet R., M. Mohammad, A .Sadeghi, A. Fassihi. Eur.J.Med.Chem., 45, (**2010**), 3, 1113.

[31] Xing J., Y. Liu, Y. Li, H. Gong, Y. Zhou. Chemometr. Intell. Lab., 137, (**2014**), 82

[32] Narasimhan B., V. Mourya, A .Dhake Bioorg. Med. Chem. Lett., 16, (**2006**), 11, 3023

[33] González-Díaz H., Á. Sánchez-González., J. Inorg. Biochem., 100, (**2006**), 7, 1290

[34] Nikolic K., D. Agababa., J. Mol. Graph., 27, (**2009**), 7, 777

[35] Rui M.V. Abreu, Isabel C.F.R. Ferreira, Ricardo C. Calhelha, Raquel T. Lima, M. Helena Vasconcelos, Maria-João R.P. Queiroz., Eur. J. Med. Chem., 46, (**2011**), 12, 5800

[36] Guendouzi A., Mekelleche S. M. Chem. Phys. Lip., 165, (**2012**), 1, 1.

[37] Fahy E,. Subramaniam, S,. Brown, A,. J Lipid Res, 46, (**2005**), 839.

[38] Guesnet P,. Alessandri J. M,. Astorg P,. Corps gras et Lipides, 12, (**2005**), 333.

[39] Bailey A.E., Melting and Solidification of Fats., (**1950**), Interscience Pub, Inc.,New-York.

[40] Moziar C., deMan, J.M., deMan, L., . Can. Inst. Food Sci. Technol. J. 22, (**1989**), 238-242.

[41] Ghotra B.S., Dyal, S.D., Narine, S.S., Lipid shortenings : a review. Food Res. Int., 35, (**2002**), 1015.

[42] David R. Lide, CRC Handbook of Chemistry and Physics, 90th (**2010**), Edition CRC Press.

[43] Katritzky A. R., Lobanov, V.S., Karelson, M., Codessa : Comprehensive Descriptors for Structural and Statistical Analysis, User Manual. University of Florida, Gainesville, (**1997**), Florida.

[44] Roberto T., V, Consonni., Handbook of Molecular Descriptors, (**2000**), WILEY-VCH.

CHAPITRE 1

Modélisation QSPR

Les relations entre les structures moléculaires et leurs propriétés sont généralement établies à l'aide de méthodes de modélisation par apprentissage statistique. Les techniques usuelles reposent sur la caractérisation des molécules par un ensemble de nombres réels mesurés ou calculés à partir des structures moléculaires "descripteurs". Ce chapitre est dédié à l'étude bibliographique des méthodologies QSPR..... [1–5]

1.1 Introduction

Les premiers essais de modélisation d'activités de molécules datent de la fin du **19ème** siècle, lorsque *Crum-Brown* et *Frazer* [4] postulèrent que l'activité biologique d'une molécule est une fonction de sa constitution chimique. Mais ce n'est qu'en **1964** que furent développés les modèles de "contribution de groupes", qui constituent le début réel de la modélisation QSPR avec les travaux de *Hansch* [5]. Depuis, l'essor de nouvelles techniques de modélisation par apprentissage, linéaires d'abord, puis non linéaires, ont permis la mise en place de nombreuses méthodes ; elles reposent pour la plupart sur la recherche d'une relation entre un ensemble de nombres réels, descripteurs de la molécule, et la propriété ou l'activité que l'on souhaite prédire. Dans les dernières années, l'utilisation des méthodes QSAR/QSPR n'a cessé de progresser. Elle est même devenue indispensable en chimie pharmaceutique et pour la conception de médicaments [6, 7]. Leur développement dans une gamme plus large d'applications leur ouvre d'ailleurs de grandes perspectives [8]. Les informations extraites à partir des résultats d'études QSAR/QSPR peuvent être utilisées pour obtenir une meilleure connaissance des structures moléculaires et probablement le mode d'action au niveau moléculaire. Ces informations peuvent alors être utilisées pour prédire les propriétés physicochimiques et les activités biologiques de nouveaux composés ainsi que pour concevoir de nouvelles structures.

1.2 Relation Quantitative Structure Propriété

1.2.1 Principe et Théorie

Le principe des méthodes QSAR/QSPR est de mettre en œuvre une relation mathématique reliant de manière quantitative des descripteurs moléculaires (constitutionnels, topologiques, géométriques, électrostatiques, quantiques, thermodynamiques, spectroscopiques,. . .) avec une observable macroscopique (propriété physicochimique ou activité biologique) pour une série de composés chimiques similaires à l'aide de méthodes statistiques d'analyse de données.
La forme générale d'un modèle QSPR/QSAR est :

(Propriété, Activité) = f(descripteurs)

L'objectif de ces méthodes est donc d'analyser les données structurales afin de détecter les facteurs déterminants pour la propriété ou l'activité mesurée. Pour ce faire, différents types d'outils peuvent être employés : régressions linéaires simples (SLR (Voir Section 2.2) (page 30)) et multiples (MLR (Voir Section 2.3) (page 36)) [9, 10], régression sur composantes principales PCA [11, 12], régressions aux moindres carrés partiels (PLS) [13], régression par réseaux de neurones [14, 15], et régression par algorithmes génétiques [16, 17].

Régressions PLS et PCA

Les régressions PLS et PCA sont des méthodes statistiques qui permettent de trouver par une transformation linéaire, les axes qui représentent au mieux les données dans l'espace. En d'autres termes, ces méthodes vont permettre de trouver les axes qui expliquent au mieux la dispersion du nuage de points. Si les données sont représentées en fonction de n descripteurs, PLS et PCA vont donc permettre de trouver au maximum n axes classés en fonction de la variance qu'ils représentent. Ces méthodes consistent à remplacer une matrice des données prédictives X comprenant n lignes et m colonnes, par une nouvelle matrice, dérivée de X, qu'on design par T, comprenant le même nombre de lignes (observations) que X, mais un nombre de colonnes k très inférieur à m. On impose, de plus, que les colonnes de la matrice T soient des combinaisons linéaires des variables d'origine. Sous forme matricielle, la relation s'écrire :

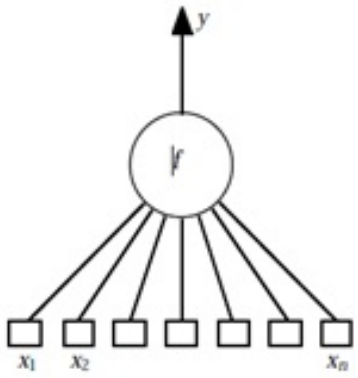
$$T = XW$$

Avec, W la matrice de dimensions $m * k$ des coefficients définissant les combinaisons linéaires. T est donc une nouvelle matrice dont les colonnes forment des « variables artificielles », obtenues par combinaison linéaire des variables d'origine. Après cette transformation, la régression linéaire multiple est appliquée sur le tableau T à la place de X. Le problème est donc de déterminer W de manière à avoir une matrice de variables prédictives T plus adaptée au calcul de la régression que la matrice X d'origine. La différence principale entre les deux méthodes PLS et ACP résident dans leur manière de calculer la matrice W.

Régression par réseaux de neurones

Les réseaux de neurones formels [14, 15] sont devenus en quelques années des outils précieux dans plusieurs domaines. Néanmoins, ils n'ont pas encore atteint leur

plein développement, pour des raisons plus psychologiques que techniques, liées aux connotations biologiques du terme,
Un "neurone formel" (ou simplement "neurone") est une fonction algébrique non linéaire et bornée, dont la valeur dépend de paramètres appelés coefficients ou poids. Les variables de cette fonction sont habituellement appelées "entrées" du neurone, et la valeur de la fonction est appelée sa "sortie". Un neurone est donc avant tout un opérateur mathématique, dont on peut calculer la valeur numérique par quelques lignes de logiciel. On a pris l'habitude de représenter graphiquement un neurone comme indiqué sur la figure suivante.
Les neurones les plus fréquemment utilisés sont ceux pour lesquels la fonction f est



une fonction non linéaire (généralement une tangente hyperbolique) d'une combinaison linéaire des entrées :

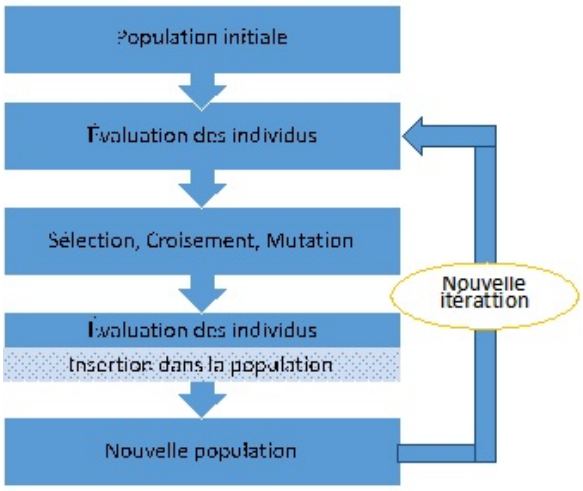
$$y = tanh\left[\sum_{i=0}^{n} c_i x_i\right]$$

Les x_i sont les variables (ou entrées) du neurone, les c_i sont des paramètres ajustables. Un neurone formel ne réalise donc rien d'autre qu'une somme pondérée suivie d'une nonlinéarité. C'est l'association de tels éléments simples sous la forme de réseaux qui permet de réaliser des fonctions utiles pour des applications industrielles.

Régression par algorithmes génétiques

Les algorithmes génétiques (AGs) [16, 17] initiés dans les années **1970** par *John Holland*, sont des algorithmes d'optimisation stochastiques itérés s'appuyant sur des techniques dérivées de la génétique et des mécanismes d'évolution de la nature : croisement, mutation, sélection. Les algorithmes génétiques fournissent des solutions

aux problèmes n'ayant pas de solutions calculables en temps raisonnable de façon analytique ou algorithmique. Les algorithmes génétiques sont présentés par les cinq éléments suivants :



1.2.2 Méthodologie de la modélisation QSPR

En pratique, le développement d'un modèle débute par la collecte de données expérimentales fiables et en nombre le plus important possible. Il s'agit ensuite de développer une série de descripteurs qui caractérisent les structures moléculaires électroniques et géométriques des composés de la base de données en vue de les relier à la propriété expérimentale étudiée. Des outils d'analyse de données sont alors employés pour aider à choisir les descripteurs adéquats et mettre en œuvre le modèle. Afin de s'assurer que le modèle QSPR développé est fiable, il doit alors être validé en termes de corrélation (sur le jeu de données d'entraînement). Pour estimer son pouvoir prédictif, il est ensuite nécessaire de disposer de données expérimentales supplémentaires pour déterminer la capacité du modèle à prédire ces valeurs.
Une fois cette relation mise en place et validée sur un jeu de validation, elle peut alors être employée pour la prédiction de la propriété de nouvelles molécules pour lesquelles la valeur expérimentale n'est pas disponible, ou pour des molécules encore non synthétisées. Et dans certains cas, peut être utilisés pour mieux appréhender les phénomènes moléculaires mis en jeu dans la propriété d'intérêt.

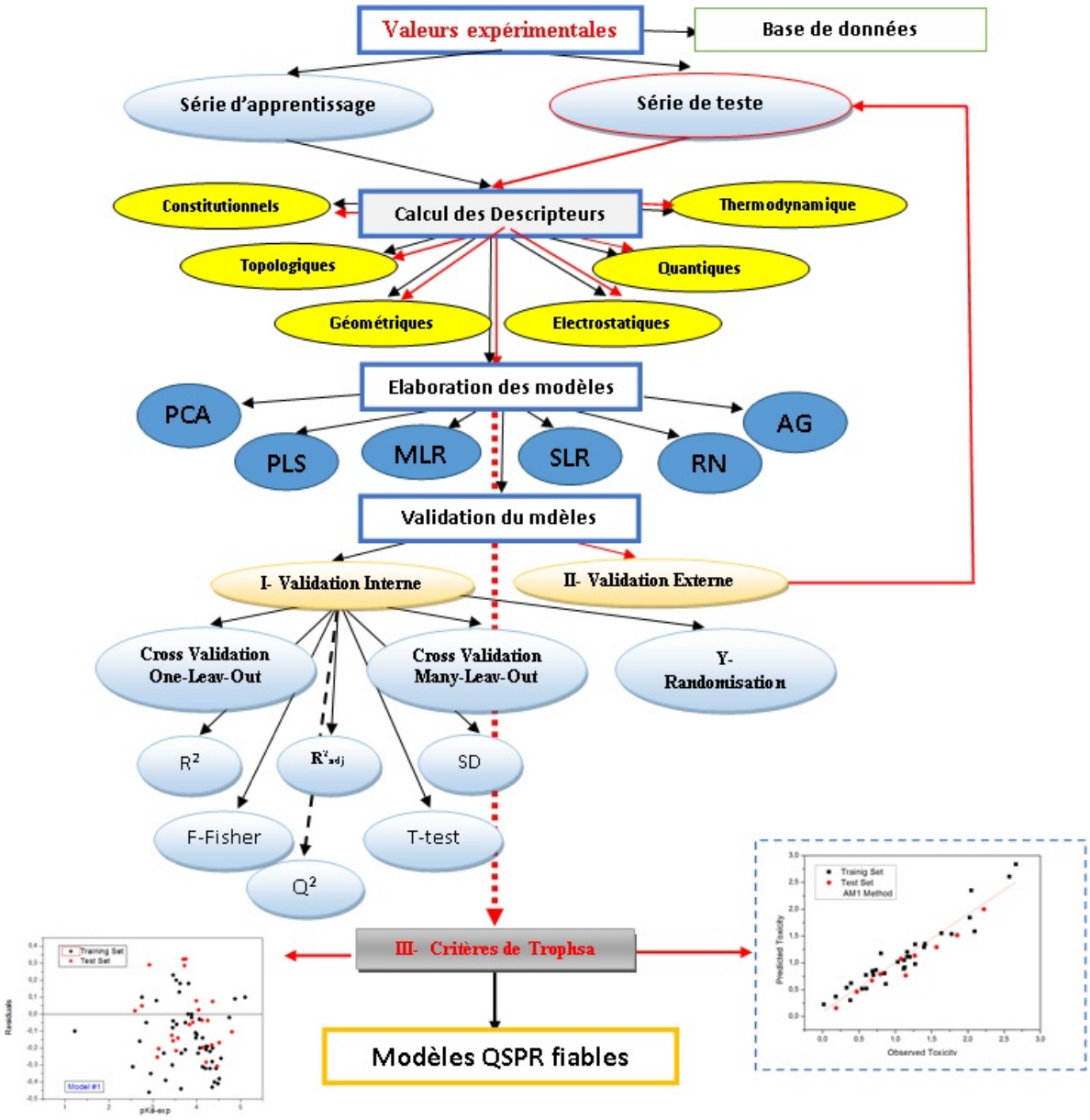


FIGURE *1.1: Méthodologie QSPR*

1.2.3 Importance de la base de données

Un modèle QSPR est dépendant des données expérimentales de référence, le choix de la base de données est un point critique dans le développement de ces modèles. Dans la plupart des cas, les données expérimentales sont issues de la littérature, et pour être de qualité, une base de données doit être composée de données expérimentales aussi fiables que possible, puisque les barres d'erreurs sur celles-ci se propageront dans le modèle final, étant donné que les paramètres de ce dernier sont ajustés par rapport à ces données. Il est donc important de choisir des données présentant des incertitudes faibles, afin de limiter les barres d'erreur expérimentales. De plus, les données doivent être obtenues suivant un protocole expérimental unique. En effet, les conditions expérimentales ont une forte influence sur les valeurs obtenues. La définition de la propriété en termes de conditions expérimentales est d'ailleurs un point important de la démarche.

1.3 Les descripteurs moléculaires

De nombreuses recherches ont été menées, au cours des dernières décennies, pour trouver la meilleure façon de représenter l'information contenue dans la structure des molécules, et ces structures elles-mêmes, en un ensemble de nombres réels appelés descripteurs moléculaires, ces descripteurs ont pour but de décrire de manière numérique la structure d'une molécule, une fois que ces nombres sont disponibles, il est possible d'établir une relation entre ceux-ci et une propriété moléculaire, à l'aide d'outils de modélisation classiques. Ces descripteurs numériques réalisent de ce fait un codage de l'information chimique en un vecteur de réels. On en dénombre aujourd'hui plus de 6000 types [2], qui quantifient des caractéristiques physico-chimiques ou structurelles de molécules. Ils peuvent être obtenus de manière empirique ou non-empirique, ils permettent en effet d'effectuer des prédictions sans avoir à synthétiser les molécules.
Nous allons présenter succinctement les descripteurs moléculaires les plus courants, en commençant par les descripteurs les plus simples, qui nécessitent peu de connaissances sur la structure moléculaire. Nous verrons ensuite comment les progrès de la modélisation moléculaire ont permis d'accéder à la structure 3D de la molécule, et de calculer des descripteurs à partir de cette structure.

1.3.1 Descripteurs constitutionnels

Les descripteurs constitutionnels sont directement liés à la formule brute de la molécule, à l'aide de la composition moléculaire, c'est-à-dire les atomes qui le constituent, Il s'agit de :

• La masse molaire

• Les nombres absolus et relatifs d'atomes ($C, H, O, S, N, F, Cl, Br, I, P\ldots$).

• Les nombres absolus et relatifs de groupes fonctionnels ($NH_2, COOH, OH\ldots$).

• Les nombres absolus et relatifs de liaisons (simples, doubles, aromatiques...).

• Les nombres absolus et relatifs de cycles (aromatiques ou non).

Ces descripteurs sont très utilisés du fait de leur extrême simplicité non seulement d'un point de vue conceptuel mais surtout calculatoire. On peut remarquer que ces descripteurs ne permettent pas de distinguer les isomères de constitution. c-a-d, si on développe des modèles avec ce type de descripteurs seulement, alors que ces derniers peuvent poser problème pour l'interprétation des mécanismes d'interaction mis en jeu pour la propriété étudiée.

1.3.2 Descripteurs topologiques

Les descripteurs topologiques "ou indices topologiques", décrivent la connectivité atomique dans la molécule, ils sont obtenus à partir de la structure 2D de la molécule, et donnent des informations sur sa taille, sa forme globale et ses ramifications. Ces descripteurs s'inspirent de la théorie des graphes appliquée à la table de connectivité qui n'est autre qu'une représentation compacte de la connectivité interatomique au sein de la molécule. Les indices topologiques les plus fréquemment utilisés sont l'indice de *Wiener* [18]. l'indice de *Randic* [19]. l'indice de connectivité de valence de *Kier-Hall* [20] et l'indice de *Balaban* [21].

L'indice *Wiener* est exprimé par la matrice de distance. Cette matrice de distance est une matrice carrée $(N_{SA}xN_{SA})$, et ces entrées $d_{i,j}$ correspond au le plus petit nombre de liaisons séparant ces deux atomes i et j. L'indice de *Wiener* W est égal à la demi-somme de toutes les entrées de la matrice de distance :

$$W = \frac{1}{2}\sum_{i,j}^{N_{sa}} d_{ij}$$

Les indices de *Randic* et *Kier & Hall* sont des descripteurs les plus utilisés.
La formule générale pour le calcul de ces indices est la suivante :

$$^{n}\chi = \sum_{i=1}^{N_s} \prod_{k=1}^{m+1} \left(\frac{1}{\delta_k^v} \right)^{\frac{1}{2}}$$
$$\delta_k^v = \frac{Z_k^v - H_k}{Z_k - Z_k^v - 1}$$

Z_k est le nombre total d'électrons dans l'atome de k
Z_k^v est le nombre d'électrons de valence à l'atome de k
H_k est le nombre d'atomes d'hydrogène liés directement à l'atome k non-hydrogène
L'indice *Balaban* est un des indices topologiques les plus importants. Il est défini par la formule suivante

$$j = \frac{q}{\mu + 1} \sum_{ij} (S_i S_j)^{-1/2} \tag{1.1}$$

En général, ce type de descripteurs simplifie la représentation de la connectivité chimique au sein de la molécule puisqu'ils ne prennent pas en compte les différences de distances, d'angles et d'ordres de liaison ni même la nature des atomes dans la molécule. Ces descripteurs reflètent bien les propriétés physiques dans la plupart des cas, mais sont insuffisants pour expliquer de façon satisfaisante certaines propriétés, ce sont souvent considérés comme des descripteurs convenables d'un point de vue numériques.

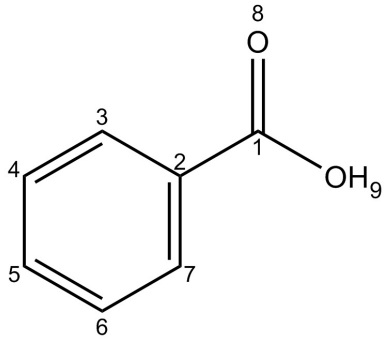


FIGURE *1.2: Distances topologiques au sein de la molécule d'acide benzoïque*

1.3.3 Descripteurs géométriques

Les descripteurs géométriques d'une molécule sont évalués à partir des positions relatives de ses atomes dans l'espace, et décrivent des caractéristiques plus complexes ; leurs calculs nécessitent de connaître, la géométrie 3D de la molécule, par modélisation moléculaire empirique ou ab initio, Ces descripteurs s'avèrent donc relativement coûteux en temps de calcul, mais apportent davantage d'informations, et sont nécessaires à la modélisation de propriétés qui dépendent de la structure 3D. On distingue plusieurs descripteurs importants, le volume moléculaire, la surface accessible au solvant, le moment d'inertie. Les distances, les angles de liaisons ou angles dièdres dans la molécule.
Le volume moléculaire est le volume occupé par la molécule en appliquant une grille 3D de cubes dans la boîte parallélépipédique dont les dimensions X_{max}, Y_{max} et Z_{max}.
La surface accessible au solvant SAS, ou la zone de surface accessible est la surface d'une molécule qui est accessible à un solvant, généralement mesuré en unités d'angströms carrés, SAS été décrite par *Lee et Richards* en **1971** et est parfois appelé la surface moléculaire de Lee-Richards [22, 23].

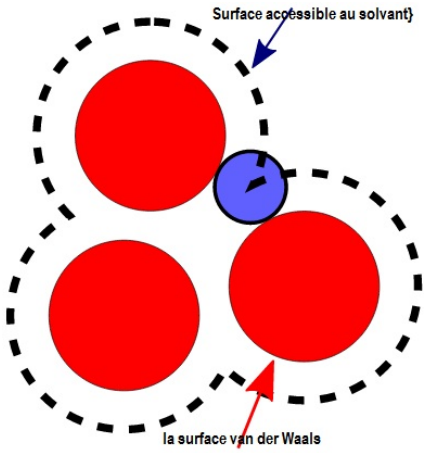


FIGURE *1.3: La surface accessible au solvant*

Le moment d'inertie est une grandeur physique qui caractérise la distribution de masse dans la molécule. Dans l'approximation des rotateurs rigide, en général pour une molécule quelconque il y a trois moments d'inertie I_A, I_B, et I_C liés à trois axes orthogonaux A, B et C avec le point d'intersection au centre de masse du système. Le moment

d'inertie est calculé comme suit :

$$I_A = \sum_i m_i r_{ix}^2$$

$$I_B = \sum_i m_i r_{iy}^2$$

$$I_C = \sum_i m_i r_{iz}^2$$

où m_i sont les masses atomiques et r_{ix}, r_{iy} et r_{iz} indiquent la distance de la i-ème noyau atomique et les principaux axes de rotation, x, y et z, de la molécule.

1.3.4 Descripteurs électrostatiques

Ces descripteurs reflètent les caractéristiques de la distribution de charge de la molécule. Les charges partielles empiriques dans la molécule sont calculées en utilisant l'approche proposée par *Zefirov* [24, 25]. Cette méthode est basée sur l'échelle d'électronégativité, Sur la base de ces charges partielles les descripteurs électrostatiques suivantes sont calculés :

• Les charges partielles minimales et maximales dans la molécule (q_{min}, q_{max})

• Les charges partielles minimales et maximales pour les atome (C, N ,O ,...)

• Les indices électroniques topologiques sont calculés selon les formules suivantes [26], pour toutes les paires d'atomes T_1^E et pour l'ensemble des laissons d'atomes liés T_2^E :

$$T_1^E = \sum_{i<j}^{N_B} \frac{|q_i - q_j|}{r_{ij}^2}$$

$$T_2^E = \sum_{i<j}^{N} \frac{|q_i - q_j|}{r_{ij}^2}$$

Avec q_i est la charge partielle pour l'atome i, r_{ij} est la distance entre les atomes i et j.

• Les charges partielles de la zone du surface (Charged partial surface area (CSPA)) ont été développés par *Jurs et al.* [27, 28], ces descripteurs sont responsable sur des interactions entre les molécules polaires. Dans CODESSA [29], un ensemble de 26 descripteurs est calculé comme une combinaison des contributions des charges partielles atomiques sur la zone accessible au solvant.

1. TMSA - total molecular surface area.
2. PPSA-1 - partial positive surface area.
3. PPSA-2 - total charge weighted PPSA.

4. PPSA-3 - atomic charge weighted PPSA.
5. PNSA-1 - partial negative surface area.
6. PNSA-2 - total charge weighted PNSA.
7. PNSA-3 - atomic charge weighted PNSA.
8. DPSA-1 - difference in CPSAs (PPSA1-PNSA1).
9. DPSA-2 - difference in CPSAs (PPSA2-PNSA2).
10. DPSA-3 - difference in CPSAs (PPSA3-PNSA3).
11. FPSA-1 - fractional CPSA (PPSA-1/TMSA).
12. FPSA-2 - fractional CPSA (PPSA-2/TMSA).
13. FPSA-3 - fractional CPSA (PPSA-3/TMSA).
14. FNSA-1 - fractional CPSA (PNSA-1/TMSA).
15. FNSA-2 - fractional CPSA (PNSA-2/TMSA).
16. FNSA-3 - fractional CPSA (PNSA-3/TMSA).
17. WPSA-1 - surface weighted CPSA (PPSA-1*TMSA/1000).
18. WPSA-2 - surface weighted CPSA (PPSA-2*TMSA/1000).
19. WPSA-3 - surface weighted CPSA (PPSA-3*TMSA/1000).
20. WNSA-1- surface weighted CPSA (PNSA-1*TMSA/1000).
21. WNSA-2 - surface weighted CPSA (PNSA-2*TMSA/1000).
22. WNSA-3 - surface weighted CPSA (PNSA-3*TMSA/1000).
23. RPCG - relative positive charge.
24. RNCG - relative negative charge.
25. RPCS - relative positive charged surface area.
26. RNCS - relative negative charged surface area.

1.3.5 Descripteurs quantiques

Les descripteurs de chimie quantique donnent des informations importantes pour la molécule. Ces descripteurs permettent de quantifier différents types d'interactions inter- et intramoléculaires, de grande influence sur des propriétés physicochimiques de molécules. Le calcul de la plupart de ces descripteurs nécessite la recherche de la géométrie pour laquelle l'énergie est minimale, et fait appel aux approches de chimie quantique. On cite toutes les données énergétiques, vibrationnelles et orbitalaires du système par exemple, les énergies de la plus haute orbitale moléculaire occupée HOMO et de la plus basse vacante LUMO, le moment dipolaire, la polarisabilité, le

potentiel d'ionisation. . . .
Les descripteurs quantiques sont classés comme suit,

a- Descripteurs lie à la distribution de charge, ces descripteurs représentent ou dépendent directement aux calculs de chimie quantique de la distribution de charge dans les molécules, qui décrire donc les interactions entre les molécules polaires ou leur réactivité chimique.
• Nombre de niveaux électroniques doublement remplis (n_f)
• Les valeurs (minimum et maximum) des charges partielles sur les atomes présentés dans la molécule, par exemple, $\delta_H(min)$ est la charge minimum (négatif) partielle sur un atome d'hydrogène dans la molécule donnée.
Dans le cadre de la théorie LCAO-MO, les populations de Mulliken fournissent une méthode de calcul des charges atomiques. qui a été donnée comme suit [30] :

$$\delta_A = Z_A - \sum_{i \in A} P_{ii}$$

où Z_A est la charge nucléaire (noyau) de l'atome i, et P_{ii} la somme de populations de Mulliken sur tous orbitales atomiques de valence.
• Le moment dipolaire total de la molécule (μ), et son point de charge (μ_c).

b- Descripteurs liés à l'énergie, Ces descripteurs caractérisent l'énergie totale de la molécule et la distribution d'énergie intramoléculaire en utilisant différents schémas de partitionnement.
• La chaleur de formation de la molécule (ΔH_F) donne l'énergie de la molécule dans l'état standard (T à 298.15K et P à 1 atm).
• L'énergie totale de la molécule (E_{tot}).
• Le potentiel du premier et la deuxième ionisation de la molécule.

$$IP(1) = -\varepsilon_{HOMO}$$
$$IP(2) = -\varepsilon_{HOMO-1}$$

• Le premier et la deuxième affinité électronique de la molécule.

$$EA(1) = -\varepsilon_{LUMO}$$
$$EA(2) = -\varepsilon_{LUMO+1}$$

• HOMO - LUMO, La différence d'énergie entre ces deux niveaux HOMO et LUMO, appelé "$gap_{HOMO-LUMO}$" est un bon indicateur de la stabilité de la molécule.

1.3.6 Descripteurs thermodynamiques

Les descripteurs thermodynamiques sont calculés sur la base de la fonction de partition totale Q de la molécule [32–34], La fonction de partition commode la façon avec laquelle l'énergie d'un système de molécules est répartie parmi les individus moléculaires. Sa valeur dépend du poids moléculaire, de la température, du volume moléculaire, des distances inter nucléaires, des mouvements moléculaires et des forces intermoléculaires. La fonction de partition est le point le plus commode entre les propriétés microscopiques des molécules individuelles (niveaux d'énergie, moments d'inertie) avec les propriétés macroscopiques (chaleur spécifique, entropie).
La molécule peut accroître son énergie de translation, de vibration, de rotation de façon pratiquement indépendante. On peut donc écrire :

$$Q = Q_{el}Q_{tr}Q_{vib}Q_{rot}$$

Avec

$$EA(1) = -\varepsilon_{LUMO}$$
$$EA(2) = -\varepsilon_{LUMO+1}$$

$$Q_{el} = exp(-\frac{E_{el}}{kT})$$
$$Q_{vib} = \prod_{j=1}^{\alpha} \frac{exp(-h\eta_j/2kT)}{1-exp(-h\eta_j/2kT)}$$
$$Q_{rot} = \frac{\pi^{1/2}}{\sigma} \prod_{j=1}^{3} \left(\frac{8\pi^2 I_j kT}{h^2}\right)$$

Dans ces formules, E_{el} est l'énergie électronique de la molécule ;
• n_j est les fréquences de vibration de la molécule ;
• α est le nombre de degrés de liberté vibrationnels dans la molécule ;
• I_j sont les moments d'inertie de la molécule ;
• σ est le nombre de symétrie de la molécule ;
• h est la constante de $Planck$, k est la constante de $Boltzmann$.
• La chaleur de formation de la molécule a 300K.

1.4 Validation des modèles QSPR

1.4.1 Réduction le nombre de variables

Dans un calcul quanto-chimique un grand nombre de descripteurs sont collectés pour la modélisation d'une grandeur donnée, car les facteurs déterminants du processus étudié ne sont a priori pas connus. Cependant, les descripteurs envisagés n'ont pas tous une influence significative sur la grandeur modélisée, et les variables ne sont pas toujours mutuellement indépendantes. De plus, le nombre de descripteurs, c'est-à-dire la dimension du vecteur d'entrée, détermine la dimension du vecteur des paramètres à ajuster. Si cette dimension est trop importante par rapport au nombre d'exemples de la base d'apprentissage, le modèle risque d'être sur ajusté à ces exemples, et incapable de prédire la grandeur modélisée sur de nouvelles observations. Il est donc nécessaire de réduire la dimensionnalité des variables d'entrée, plusieurs approches sont possibles pour résoudre ce problème.

• Réduire la dimension de l'espace des entrées.

• Remplacer les variables corrélées par de nouvelles variables synthétiques, obtenues à partir de leurs combinaisons.

• Sélectionner les variables les plus pertinentes.

1.4.2 Validation interne

La technique la plus employée pour déterminer la stabilité du modèle prédictif est de tester l'influence de chaque échantillon sur le modèle final. Pour ce faire, on emploie une technique de validation croisée (cross validation ou CV). Ce processus consiste à extraire un certain nombre k de molécules du jeu initial à N molécules et à construire un nouveau modèle avec les N-k molécules restantes à l'aide des descripteurs choisis (seules les constantes de la régression changent), ce nouveau modèle est alors utilisé pour la phase de prédiction sur les k molécules retirées. Ce processus est ensuite réitéré pour retirer et prédire les valeurs de toutes les molécules du jeu d'entraînement, en fonction du nombre de molécules retirées à chaque itération, on parlera de Leave-One-Out (LOO) ou de Leave-Many-Out (LMO) [35] selon qu'une ou plusieurs molécules est (sont) retirée(s). D'une manière générale, ces techniques de validation interne permettent l'évaluation de la robustesse du modèle, autrement dit la stabilité des paramètres du modèle QSPR vis-à-vis des molécules du jeu d'entraînement.

Cela dit, elles ne permettent en aucun cas de démontrer le pouvoir prédictif des modèles [36, 37].

Y- Randomization

Afin d'évaluer la part de chance dans les modèles QSRR construits, et à cause de la complexité des outils de chimiométrie employés, qui peuvent constituer une source de corrélations fortuites, dans le but d'établir que le modèle n'est pas dû au hasard, une technique appelée scrambling ou Y-randomization [38, 39] est utilisée. Ce test consiste à générer un vecteur de la propriété étudiée par permutation aléatoire des composantes du vecteur réel. On calcule alors sur le vecteur obtenu un modèle QSRR, selon la méthode habituelle. Ce procédé est répété plusieurs fois. Si les modèles sont affectés par des corrélations fortuites on devrait voir apparaître de bonnes performances en validation croisée malgré le fait d'avoir mélangé la propriété au hasard. Dans le cas contraire la performance des modèles devrait être équivalente à celle d'un modèle faisant des prédictions au hasard.

1.4.3 Validation externe

Afin de tester de manière fiable le pouvoir prédictif, l'utilisation d'un jeu de validation externe, non employé pour le développement du modèle, est nécessaire, pour que le jeu de données initial soit suffisamment large, cette validation est caractérisée par les paramètres $R^2(test)$ $R^2_{cv}(test)$. Par conséquent, d'autres paramètres doivent être vérifiés pour cet objectif. Ces paramètres sont connus sous le nom « critères de validation externe » ou souvent appelés « *critères de Tropsha* » (*Tropsha criteria*) [36, 37].

Critères de validation Externe (série de test)

- $R^2 > 0.7$
- $R^2_{cv} > 0.6$
- $\frac{R^2 - R_0^2}{R^2} < 0.1$ et $0.85 \leq k \leq 1.15$
- $\frac{R^2 - R_0'^2}{R^2} < 0.1$ et $0.85 \leq k' \leq 1.15$
- $|R^2 - R_0^2| \leq 0.3$

Avec

R^2 Coefficient de corrélation pour les molécules de la série de test.

R_0^2 coefficient de corrélation entre les valeurs prédites et expérimentales pour la série de test.
$R_0'^2$ coefficient de corrélation entre les valeurs expérimentales et prédites pour la série de test.
k : est la constante de la droite (à l'origine) de corrélation (valeurs prédites en fonction des valeurs expérimentales)
k' : est la constante de la droite (à l'origine) de corrélation (valeurs expérimentales en fonction des valeurs prédites).

Bibliographie

[1] Marvin Charton, Advances in Quantative Structure Property Relationships, (**2002**), Vol 3. 1st Ed, JAI PRESS INC.

[2] Roberto T., V. Consonni, Handbook of Molecular Descriptors, (**2000**), WILEY-VCH.

[3] Gasteiger J., H. Kubinyi, Handbook of Chemoinformatics, (**2008**), WILEY-VCH.

[4] Crum Brown A., Fraser, T. R. On the connection between chemical constitution and physiological action. On the physiological action of the salts of the ammonium bases, derived from strychnia, brucia. thebaia. codeia. morphia. and nicotia. J Anat Physiol., 2(2),(**1868**), 224.

[5] Hansch C., Fujita, T. ρ-σ-π analysis. A method for the correlation of biological activity and chemical structure. J. Am. Chem. Soc., 86, (**1964**), 1616.

[6] Grover M. B., Singh, M. Bakshi, S. Singh, Quantitative structure-property relationships in pharmaceutical research-Part 1, Pharm. Sci. Tech. Today., 3, (**2000**), 28.

[7] Grover M. B., Singh, M. Bakshi, S. Singh, Quantitative structure-property relationships in pharmaceutical research-Part 2. Pharm. Sci. Tech. Today., 3, (**2000**), 50.

[8] Katritzky A.R., D.C. Fara, R.O. Petrukhin, D.B. Tatham, U. Maran, A. Lomaka, M. Karelson, The present utility and future potential for medicinal chemistry of QSAR/QSPR with whole molecule descriptors, Curr. Top. Med. Chem., 2, (**2002**), 1333.

[9] Alvin C., Rencher and G. Bruce Schaalje, Linear Models in Statistics, Second Edition, (**2008**), John Wiley & Sons, Inc.

[10] Weisberg S., Applied Linear Regression. (**1985**), New York, Wiley

[11] Jolliffe I. T., Principal Component Analysis, Series, Springer Series in Statistics, 2nd ed., (**2002**), Springer, NY,

[12] Abdi H., Williams, L.J. "Principal component analysis.". Wiley Interdisciplinary Reviews, Computational Statistics, 2, (**2010**),433.

[13] Michel Tenenhaus„ La régression PLS , Théorie et Pratique, Paris, Éditions Technip, (**1998**), 254

[14] François Blayo., Michel Verleysen, "Les réseaux de neurones artificiels", (**1996**), 1re édition,

[15] Léon Personnaz., Isabelle Rivals, Réseaux de neurones formels pour la modélisation, la commande et la classification, (**2003**), CNRS Éditions,

[16] David Goldberg., Genetic Algorithms in Search, Optimization, and Machine Learning, (**1989**), Addison-Wesley Professional,

[17] Jean-Marc A., Thomas S., Intelligence Artificielle et Informatique Théorique, (**1994**), Éditions Cepadues,

[18] Wiener H., Structural determination of paraffin boiling points. Journal of Chemical Information and Computer Sciences, 69, (**1947**), 17.

[19] Randic M., On characterization of molecular branching. Journal of the American Chemical Society, 97, (**1975**), 6609.

[20] Kier L.B., L.H. Hall, Derivation and significance of valence molecular connectivity, J. Pharm., Sci., 70, (**1981**), 583.

[21] Balaban A.T., Highly discriminating distance-based topological index. Chemical Physics Letters., 89, (**1982**), 399.

[22] Lee B., Richards, F. M., The interpretation of protein structures, estimation of static accessibility, J Mol Biol., 55, (**1971**), 3, 379.

[23] Shrake A., Rupley J. A., Environment and exposure to solvent of protein atoms. Lysozyme and insulin, J Mol Biol., 79 (2), (**1973**), 351.

[24] Zefirov N. S., Kirpichenok, M.A., Izmailov, F.F., Trofimov, M.I. Dokl. Akad. Nauk SSSR, 296, (**1987**), 883.

[25] Kirpichenok, M. A., Zefirov, N.S. Zh. Org. Khim., 23, (**1987**), 4.

[26] Osmialowski K., Halkiewicz, J., Kaliszan, R. J. Chromatogr. 361, (**1986**), 63.

[27] Stanton D. T., Jurs, P.C. Anal. Chem., 62, (**1990**), 2323.

[28] Stanton D. T., Egolf, L.M., Jurs, P.C., Hicks, M.G. J. Chem. Inf. Comput. Sci., 32, (**1992**), 306.

[29] Katritzky A. R., Lobanov, V.S., Karelson, M., Codessa, Comprehensive Descriptors for Structural and Statistical Analysis, User Manual. University of Florida, Gainesville, (**1997**), Florida.

[30] Pople J. A., Beveridge, D.L. Approximate Molecular Orbital Theory, (**1970**), McGraw Hill, New York.

[31] Fukui K., Theory of Orientation nad Stereoselection, (**1975**), Springer-Verlag, Berlin.

[32] McQuarrie D. A., Statistical Thermodynamics, (**1973**), Harper Row Publ., New York.

[33] Akhiezer A. I., Peltminskii, S.V. Methods of Statistical Physics, (**1981**), Pergamon Press, Oxford.

[34] Atkins P. W. Physical Chemistry, Ch. 20, 2nd edition, W.H. (**1982**), Freeman and Co, San Francisco.

[35] Allen D. M., The relationship between variable selection and data augmentation and a method for prediction. Technometrics, 16, (**1974**), 125.

[36] Golbraikh A., A. Tropsha, Beware of q2 !, J. Mol. Graph. Model., 20,(**2002**), 269.

[37] **(a)** Tropsha A. P. Gramatica, and V.K. Gombar, The importance of being earnest, validation is the absolute essential for successful application and interpretation of QSPR models, QSAR Comb. Sci. 22, (**2003**), 69. **(b)** Tropsha A. Best practices for QSAR model development, validation, and exploitation. Molecular Informatics, 29, (**2010**), 476.

[38] Rucker C., G. Rucker, and M. Meringer, y-Randomization and its variants in QSPR/QSAR. Journal of Chemical Information and Modeling, 47(6), (**2007**), 2345.

[39] Lindgren F., B. Hansen, W. Karcher, M. Sjöström, and L. Eriksson, Model validation by permutation tests, Applications to variable selection, J. Chemometrics, 10 (**1996**), 521.

CHAPITRE 2

Statistiques et méthodes d'analyse de données

La notion de régression est fondamentale dans toutes les sciences appliquées puisqu'elle consiste à analyser une relation entre deux ou plusieurs variables quantitatives et à l'exploiter pour estimer la valeur inconnue de l'une à l'aide de la valeur connue de l'autre. Elle est couramment utilisée dans les domaines de la chimie et de la biologie, Nous formalisons ici la démarche utilisée dans ce chapitre pour calculer l'équation de régression linéaire simple et multiple [1–8].

2.1 Introduction

La méthode de régression linéaire est classée parmi les méthodes d'analyses multivariées qui traitent des données quantitatives. Cette méthode de calcul date de la fin du 18^{me} siècle. Elle est essentiellement, due à *A. M. Legendre* **1805** et. *C. F. Gauss* **1809** [9, 10]. Le mot régression vient, de la fin du 19^{me} siècle, et est dû à *F. Galton* **1886** [11]. *Galton* a introduit le mot régression pour expliquer la diminution progressive des écarts par rapport à la moyenne, d'une génération à la suivante, dans des problèmes d'hérédité, Le terme régression a ensuite été utilisé d'une manière tout à fait générale, L'analyse de la régression est une méthode statistique qui permet d'étudier le type de relation pouvant exister entre une ou plusieurs variables explicatives (notée Xi) et une variable à expliquer (notée Y). En d'autres termes, l'analyse de la régression permet d'étudier les variations de la variable dépendante en fonction des variations connues des variables indépendantes.

2.2 La régression linéaire simple

La régression linéaire dite simple si elle permet de prédire les valeurs d'une variable (expliquée (Y)) à partir des valeurs prises par une autre variable (explicative (X)). Lorsque cette dépendance est exacte, la liaison entre les deux variables est «fonctionnelle » : à chaque valeur de X correspond une et une seule valeur possible de Y [12, 13].

2.2.1 Modélisation

Le modèle de régression est simplement une équation censée représenter cette relation entre les deux variables. Il s'écrit

$$Y = f(X) + \varepsilon \tag{2.1}$$

La variable Y est donc supposée approximativement égale à une fonction f(x), le terme ε caractérisant la marge d'erreur ou d'imprécision du modèle.
Dans un modèle linéaire simple, Nous savons pertinemment que toutes les observations mesurées ne sont pas sur la droite, et d'une part, il est irréaliste de croire que la

variable X dépend linéairement de Y, mais cette liaison est perturbée par un « bruit ». Nous supposons en fait que les données suivent le modèle suivant :

$$Y = \beta_1 + \beta_2 X + \varepsilon \tag{2.2}$$

L'équation (2.2) est appelée modèle de régression linéaire simple. Les β_j, appelés les constantes de régression (ou coefficients de régression), sont fixes mais inconnus, et nous voulons les estimer, la quantité notée ε est appelée bruit, ou erreur.
Afin d'estimer les paramètres inconnus du modèle, nous mesurons dans le cadre de la régression simple une seule variable explicative ou variable exogène X et une variable à expliquer ou variable endogène Y, la variable X est souvent considérée comme non aléatoire au contraire de Y. Nous mesurons alors n observations de la variable X, notées x_i, où i varie de 1 à n, et n valeurs de la variable à expliquer Y notées y_i.
Nous supposons que nous avons collecté n couples de données (x_i, y_i) où y_i est la réalisation de la variable aléatoire Y_i, suivant le modèle (2.2), nous pouvons écrire

$$y_i = \beta_1 + \beta_2 x_i + \varepsilon_i \qquad i = 1, ..., n \tag{2.3}$$

Où

- Les x_i sont des valeurs connues non aléatoires ;
- Les paramètres β_j, j = 1, 2 du modèle sont inconnus ;
- Les ε_i sont les réalisations inconnues d'une variable aléatoire ;
- Les y_i sont les observations d'une variable aléatoire.

2.2.2 Estimateurs des moindres carrés

On appelle estimateurs des moindres carrés (MC) de β_1 et β_2, les estimateurs $\widehat{\beta}_1$ et $\widehat{\beta}_2$ obtenus par minimisation de la quantité

$$S(\beta_1, \beta_2) = \sum_{i=1}^{n} \varepsilon_i = \sum_{i=1}^{n} (y_i - \beta_1 - \beta_2 x_i)^2 = \|Y - \beta_1 \mathbf{1} - \beta_2 X\|^2 \tag{2.4}$$

où $\mathbf{1}$ est le vecteur de $\Re^n$ dont tous les coefficients valent 1, les estimateurs peuvent s'écrire sous la forme de $\arg\min$ [i] :

$$(\widehat{\beta}_1, \widehat{\beta}_2) = \underset{(\beta_1, \beta_2) \in \Re x \Re}{\arg\min} \; S(\beta_1, \beta_2) \tag{2.5}$$

i. $\arg\min$ est l'argument du minimum, est l'ensemble des points en lesquels une expression atteint sa valeur minimum.

2.2.3 Calcul des estimateurs de β_j :

La fonction $S(\beta_1, \beta_2)$ est strictement convexe. Si elle admet un point singulier, celui-ci correspond à l'unique minimum. Nous obtenons un système d'équations :

$$\frac{\partial S(\widehat{\beta}_1, \widehat{\beta}_2)}{\partial \beta_1} = -2 \sum_{i=1}^{n} (y_i - \widehat{\beta}_1 - \widehat{\beta}_2 x_i) = 0 \tag{2.6}$$

$$\frac{\partial S(\widehat{\beta}_1, \widehat{\beta}_2)}{\partial \beta_2} = -2 \sum_{i=1}^{n} x_i (y_i - \widehat{\beta}_1 - \widehat{\beta}_2 x_i) = 0 \tag{2.7}$$

La première équation (2.6) donne

$$\widehat{\beta}_1 n + \widehat{\beta}_2 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i \tag{2.8}$$

Et nous avons un estimateur de l'ordonnée à l'origine

$$\widehat{\beta}_1 = \overline{y} - \widehat{\beta}_2 \overline{x} \tag{2.9}$$

Où

$$\overline{x} = \sum_{i=1}^{n} x_i / n \tag{2.10}$$

La seconde équation (2.7) donne

$$\widehat{\beta}_1 \sum_{i=1}^{n} x_i + \widehat{\beta}_2 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i \tag{2.11}$$

En remplaçant $\widehat{\beta}_1$ par son expression (2.9) nous avons une première écriture de

$$\widehat{\beta}_2 = \frac{\sum x_i y_i - \sum x_i \overline{y}}{\sum x_i^2 - \sum x_i \overline{x}} \tag{2.12}$$

Et en utilisant astucieusement la nullité de la somme $\sum (x_i - \overline{x})$, nous avons d'autres écritures pour l'estimateur de la pente de la droite

$$\widehat{\beta}_2 = \frac{\sum x_i (y_i - \overline{y})}{\sum x_i (x_i - \overline{x})} = \frac{\sum (x_i - \overline{x})(y_i - \overline{y})}{\sum (x_i - \overline{x})(x_i - \overline{x})} = \frac{\sum (x_i - \overline{x}) y_i}{\sum (x_i - \overline{x})^2} = \frac{cov(x,y)}{var(x)} \tag{2.13}$$

En remplaçant l'expression de $\widehat{\beta}_2$ dans l'équation (2.9) on obtient :

$$\widehat{\beta}_1 = \overline{Y} - \frac{cov(x,y)}{var(x)} \overline{X} \tag{2.14}$$

Pour obtenir ce résultat, nous supposons qu'il existe au moins deux points d'abscisses différentes. Cette hypothèse notée $\mathbf{H_1}$ s'écrit $x_i \neq x_j$ pour au moins deux individus.

Elle permet d'obtenir l'unicité des coefficients estimés $\widehat{\beta}_1, \widehat{\beta}_2$. Une fois déterminés les estimateurs $\widehat{\beta}_1, \widehat{\beta}_2$, nous pouvons estimer la droite de régression par la formule

$$\widehat{Y} = \widehat{\beta}_1 + \widehat{\beta}_2 X \tag{2.15}$$

Si nous évaluons la droite aux points x_i ayant servi à estimer les paramètres, nous obtenons des $\widehat{y}_i$ et ces valeurs sont appelées les valeurs ajustées. Si nous évaluons la droite en d'autres points, les valeurs obtenues seront appelées les valeurs prévues ou prévisions, représentons les points initiaux et la droite de régression estimée.
La droite de régression passe par le centre de gravité du nuage de points $(\overline{x}, \overline{y})$ comme l'indique l'équation (2.9).

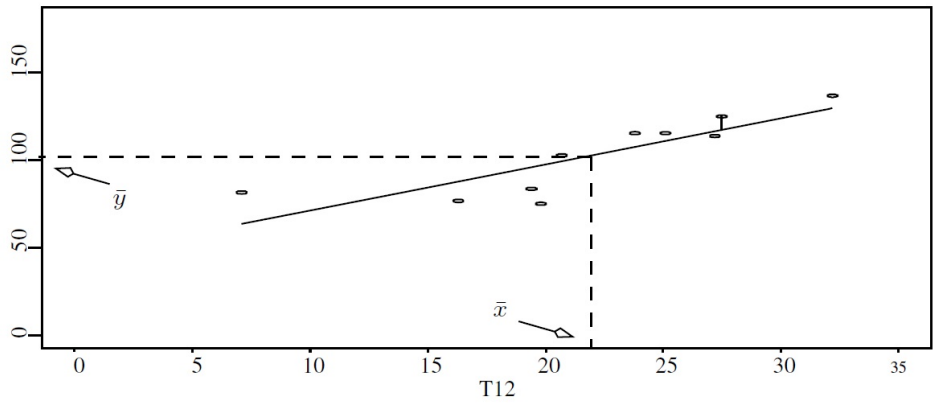


FIGURE *2.1: Nuage de points, droite de régression et centre de gravité*

Afin d'estimer les variances de $\widehat{\beta}_1$ et $\widehat{\beta}_2$, nous allons tout d'abord donner un rappel sur quelques notions de base de statistique, espérance mathématique, variance, écart-type, et covariance.

Espérance mathématique

Si la variable X prend les valeurs $x_1, x_2, ..., x_n$ avec P les probabilités $p_1, p_2, ..., p_n$, l'espérance de X est définie par :

$$E(X) = \sum_{i}^{n} x_i P_i$$

Comme la somme des probabilités est égale à un, l'espérance peut être considérée comme la moyenne des x_i pondérée par les p_i

$$E(X) = \frac{x_1 P_1 + x_2 P_2 + ... + x_n P_n}{P_1 + P_2 + ... + P_n}$$

Variance

La variance est une mesure servant à caractériser la dispersion d'un échantillon. Elle indique de quelle manière la série statistique ou la variable aléatoire se disperse autour de sa moyenne ou son espérance. Elle est donnée par :

$$V(X) = \sum_{i}^{n} (X_i - \overline{X})^2$$

Ecart-type

Ecart-type est une mesure de la dispersion d'une variable aléatoire, c'est la racine carrée de la variance.

$$\sigma = \sqrt{V(X)}$$

Covariance

La covariance entre deux variables aléatoires est un nombre permettant de quantifier leurs écarts conjoints par rapport à leurs espérances respectives. Elle s'utilise également pour deux séries de données numériques (écarts par rapport aux moyennes).

$$Cov(X,Y) = \sum_{i}^{n} (X_i - \overline{X})(Y_i - \overline{Y})$$

2.2.4 Variances de $\widehat{\beta}_1$ et $\widehat{\beta}_2$

Cette proposition nous permet d'envisager la précision de l'estimation en utilisant la variance. Plus la variance est faible, plus l'estimateur sera précis, pour avoir des variances petites, il faut avoir un numérateur petit et (ou) un dénominateur grand, les estimateurs seront donc de faibles variances lorsque :

$$V(\widehat{\beta}_2) = \frac{\sigma^2}{\sum (x_i - \overline{x}_i)^2} \tag{2.16}$$

$$V(\widehat{\beta}_1) = \frac{\sigma^2 \sum x_i^2}{n \sum (x_i - \overline{x}_i)^2} \tag{2.17}$$

$$Cov(\widehat{\beta}_1, \widehat{\beta}_2) = \frac{\sigma^2 \overline{x}}{\sum (x_i - \overline{x}_i)^2} \tag{2.18}$$

• La variance σ^2 est faible, cela signifie que la variance de Y est faible et donc les mesures sont proches de la droite à estimer ;

• La quantité $\sum (x_i - \overline{x})^2$ est grande, les mesures x_i doivent être dispersées autour de

leur moyenne ;

• La quantité $\sum x^2$ ne doit pas être trop grande, les points doivent avoir une faible moyenne en valeur absolue.

2.2.5 Résidus et variance résiduelle

Nous avons estimé β_1 et β_2, la variance σ^2 des ε_i est le dernier paramètre inconnu à estimer, pour cela, nous allons utiliser les résidus : ce sont des estimateurs des erreurs inconnues ε_i.

Les résidus sont définis par

$$\widehat{\varepsilon}_i = y_i - \widehat{y}_i \tag{2.19}$$

Où $\widehat{y}_i$ est la valeur ajustée de y_i par le modèle, c'est-à-dire $\widehat{y}_i = \widehat{\beta}_1 + \widehat{\beta}_2 x_i$

Dans un modèle de régression linéaire simple, la somme des résidus est nulle.

2.2.6 Interprétations géométriques

Pour chaque individu, ou observation, nous mesurons une valeur x_i et une valeur y_i. Une observation peut donc être représentée dans le plan, nous dirons alors que $\Re^2$ est l'espace des observations.$\widehat{\beta}_1$ correspond à l'ordonnée à l'origine alors que $\widehat{\beta}_1$ représente la pente de la droite ajustée. Cette droite minimise la somme des carrés des distances verticales des points du nuage à la droite ajustée.

Les couples d'observations (x_i, y_i) avec i = 1,..., n ordonnées suivant les valeurs croissantes de x sont notés $(x_{(i)}, y_{(i)})$. Nous avons représenté la neuvième valeur de x et sa valeur ajustée $\widehat{y}_{(9)} = \widehat{\beta}_1 + \widehat{\beta}_2 x_{(9)}$ sur le graphique, ainsi que le résidu correspondant $\widehat{\varepsilon}_{(9)}$.

A partir des équations (2.9) et (2.15) on obtient :

$$\widehat{Y}_i = \overline{Y} + \widehat{\beta}_1 (X_i - \overline{X}) \tag{2.20}$$

$$\widehat{Y}_i - \overline{Y} = \widehat{\beta}_1 (X_i - \overline{X}) \tag{2.21}$$

$$Y_i - \widehat{Y}_i = (Y_i - \overline{Y}) - (\widehat{Y}_i - \overline{Y}) \tag{2.22}$$

On élève les deux membres au carré et on somme sur les observations i on obtient

$$\sum_{i=1}^{n} (Y_i - \widehat{Y}_i)^2 = \sum_{i=1}^{n} (Y_i - \overline{Y})^2 + \sum_{i=1}^{n} (\widehat{Y}_i - \overline{Y})^2 - 2 \sum_{i=1}^{n} (Y_i - \overline{Y})(\widehat{Y}_i - \overline{Y}) \tag{2.23}$$

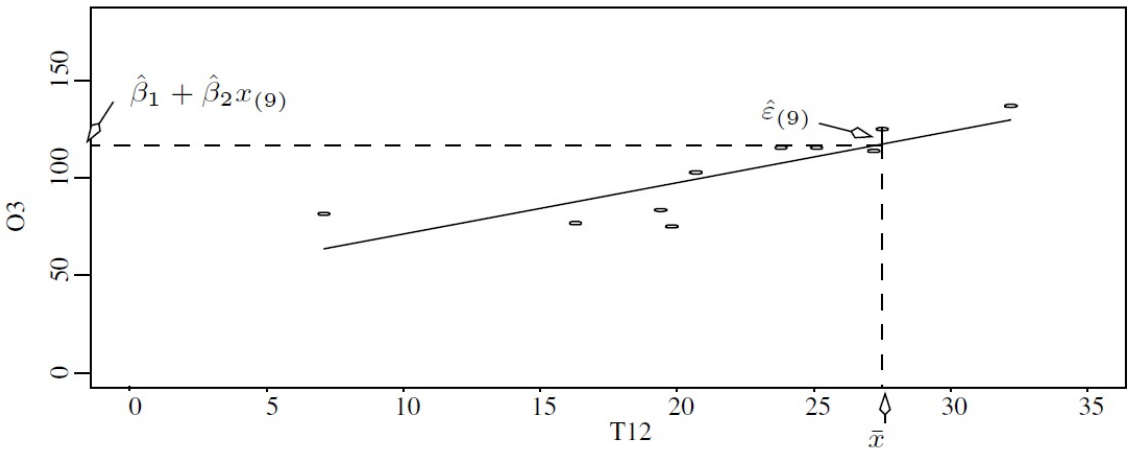


FIGURE *2.2: Représentation des individus*

Avec l'équation (2.22), on obtient,

$$\sum_{i=1}^{n}(Y_i-\widehat{Y}_i)^2=\sum_{i=1}^{n}(Y_i-\overline{Y})^2+\sum_{i=1}^{n}(\widehat{Y}_i-\overline{Y})^2-2\widehat{\beta}_1\sum_{i=1}^{n}(Y_i-\overline{Y})(X_i-\overline{X}) \tag{2.24}$$

Et avec l'équation (2.9), on obtient,

$$\sum_{i=1}^{n}(Y_i-\widehat{Y}_i)^2=\sum_{i=1}^{n}(Y_i-\overline{Y})^2+\sum_{i=1}^{n}(\widehat{Y}_i-\overline{Y})^2-2\widehat{\beta}_1^2\sum_{i=1}^{n}(X_i-\overline{X})^2 \tag{2.25}$$

$$\sum_{i=1}^{n}(Y_i-\widehat{Y}_i)^2=\sum_{i=1}^{n}(Y_i-\overline{Y})^2+(\widehat{Y}_i-\overline{Y})^2-2\sum_{i=1}^{n}(\widehat{Y}_i-\overline{Y})^2 \tag{2.26}$$

Finalement on obtient la relation fondamentale de l'analyse de la variance

$$\sum_{i=1}^{n}(Y_i-\overline{Y})^2=\sum_{i=1}^{n}(\widehat{Y}_i-\overline{Y})^2+\sum_{i=1}^{n}(Y_i-\widehat{Y}_i)^2 \tag{2.27}$$

$$SCT = SCE + SCR \tag{2.28}$$

Où SCT (respectivement SCE et SCR) représente la somme des carrés totale (respectivement expliquée par le modèle et résiduelle).

2.3 La régression linéaire multiple

2.3.1 Modélisation

Le modèle de régression multiple est une généralisation du modèle de régression simple lorsque les variables explicatives sont en nombre fini [14–23]. Nous ne suppo-

sons donc que les données collectées suivent le modèle suivant :

$$y_i = \beta_1 x_{i1} + \beta_2 x_{21} + ...\beta_p x_{ip} + \varepsilon_i, \qquad i = 1, ..., n \tag{2.29}$$

où

• Les x_{ij} sont des nombres connus, non aléatoires. La variable x_{i1} peut valoir 1 pour tout i variant de 1 à n. Dans ce cas, β_1 représente la constante (intercept). En statistiques, cette colonne de 1 est presque toujours présente ;

• Les paramètres à estimer β_j du modèle sont inconnus.

• Les ε_i sont des variables aléatoires inconnues.

En utilisant l'écriture matricielle de (2.29), nous obtenons le modèle de régression linéaire, défini par l'équation :

$$Y_{nx1} = X_{nxp}\beta_{px1} + \varepsilon_{nx1} \tag{2.30}$$

Où :

• Y est un vecteur aléatoire de dimension n ;

• X est une matrice de taille nxp connue, appelée matrice du plan d'expérience,

X est la concatenation des p variables X_j : $X = (X_1|X_2|...|X_p)$. Nous noterons la i^e ligne du tableau X par le vecteur ligne $x_i^{'} = (x_{i1}, ..., x_{ip})$;

• β est le vecteur de dimension p des paramètres inconnus du modèle ;

• ε est le vecteur centré, de dimension n, des erreurs.

Nous supposons que la matrice X est de plein rang. Cette hypothèse sera notée $\mathbf{H_1}$. Comme, en général, le nombre d'individus n'est plus grand que le nombre de variables explicatives p, le rang de la matrice X vaut p.

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

$$\begin{pmatrix} y_i \\ y_2 \\ y_3 \\ ... \\ y_p \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & x_{12} & ... & x_{1p} \\ 1 & x_{21} & x_{22} & ... & x_{2p} \\ 1 & x_{31} & x_{32} & ... & x_{3p} \\ ... & ... & ... & ... & ... \\ 1 & x_{n1} & x_{n2} & ... & x_{np} \end{pmatrix} \begin{pmatrix} \beta_i \\ \beta_2 \\ \beta_3 \\ ... \\ \beta_p \end{pmatrix} + \begin{pmatrix} \varepsilon_i \\ \varepsilon_2 \\ \varepsilon_3 \\ ... \\ \varepsilon_p \end{pmatrix} \tag{2.31}$$

Il est naturel dans nombre de problèmes de penser que des interactions existent entre les variables explicatives. Pour modéliser cette interaction, nous écrivons en général un modèle avec un produit entre les variables explicatives qui interagissent. Ainsi, pour deux variables, nous avons la modélisation suivante :

$$y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i1} x_{i2} + \varepsilon_i, \qquad i = 1, ..., n \tag{2.32}$$

Les produits peuvent s'effectuer entre deux variables définissant des interactions d'ordre 2, entre trois variables définissant des interactions d'ordre 3, etc.

Cependant, ce type de modélisation rentre parfaitement dans le cadre de la régression multiple. Les variables d'interaction sont des produits de variables connues et sont donc connues.

2.3.2 Estimateurs des moindres carrés

On appelle estimateur des moindres carrés (noté MC) $\widehat{\beta}$ de β la valeur suivante :

$$\widehat{\beta} = \underset{(\beta_1,..,\beta_p)}{\arg\min} \sum_{i=0}^{n} \left(y_i - \sum_{j=1}^{p} \beta_j x_{ij}\right)^2 = \underset{\beta \in \Re^p}{\arg\min} \|Y - X\beta\|^2 \tag{2.33}$$

Les valeurs des β_i qui minimisent ce critère seront les solutions β_1, β_2, ... β_p du système.

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

En premier lieu, les variables indépendantes Xi, comme leur nom l'indique, sont supposées indépendantes entre elles et leur incertitude est négligeable. Ensuite, les différents échantillons Y_i sont supposés indépendants entre eux. Enfin, par nature, la dépendance de Y vis-à-vis des Xi est supposée linéaire.La valeur prédite de la variable dépendante Y (estimé par le modèle de régression) s'écrit :

$$\widehat{Y} = X\widehat{\beta} = X\beta \tag{2.34}$$

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

$$Y_i - \widehat{Y}_i = \widehat{\varepsilon} \tag{2.35}$$

Interprétation géométrique :

Comme pour le modèle linéaire simple, les hypothèses de régression linéaire doivent être vérifiées pour un modèle de régression multiple. Le théorème de Pythagore donne directement l'égalité suivante :

$$\begin{aligned} \|Y\|^2 &= \|\widehat{Y}\|^2 + \|\widehat{\varepsilon}\|^2 \\ &= \|X\widehat{\beta}\|^2 + \|Y - X\widehat{\beta}\|^2 \end{aligned}$$

Si la constante fait partie du modèle, alors nous avons toujours par le théorème de Pythagore

$$\|Y - \widehat{y}\mathbf{1}\|^2 = \|\widehat{Y} - \widehat{y}\mathbf{1}\|^2 + \|\widehat{\varepsilon}\|^2$$

$$SCtotal = SCexplique \ \ par \ \ le \ \ modele + SCresiduelle$$

$$SCT = SCE + SCR. \tag{2.36}$$

Ces formules montrent que les variations de y de sa moyenne, c-a-d SCT peuvent être expliquées par el modèle grâce SCE modèle et ce que ne pas être expliqué par le modéle est SCE.

2.4 Validation du modèle

2.4.1 L'analyse de la variance

L'analyse de la variance (terme souvent abrégé par le terme anglais *(ANOVA)* : *(ANalysis OF VAriance)*) est un test statistique permettant de vérifier que plusieurs échantillons sont issus d'une même population.
Ce test s'applique lorsque l'on mesure une ou plusieurs variables explicatives catégorielles qui ont de l'influence sur la distribution d'une variable continue à expliquer. Nous présentons dans le tableau suivant l'analyse de variance.
Le degré de liberté (ddl) désigne le nombre de variables aléatoires.

Source de variation	Degrés de liberté	Somme des carrés	Moyenne des carrées
Modèle	P	$\sum(\widehat{Y}_i - Y)^2$	$\sum(\widehat{Y}_i - Y)^2/p$
Résidus	n-1-p	$\sum(Y_i - \widehat{Y}_i)^2$	$\sum(Y_i - \widehat{Y}_i)^2/(n-p-1)$
Totale	n-1	$\sum(Y_i - \overline{Y}_i)^2$	

2.4.2 Hypothèses de l'analyse de régression linéaire

2.4.2.1 Normalité des résidus

Loi normale En théorie des probabilités, on dit qu'une variable aléatoire réelle x suit une loi normale (ou loi normale gaussienne, loi de Laplace-Gauss) d'espérance μ et

d'écart type σ strictement positif (donc de variance σ^2) si cette variable aléatoire réelle x admet pour densité de probabilité la fonction p(x) définie, pour tout nombre réel x, par :

$$p(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp -\frac{1}{2}(\frac{x-\mu}{\sigma})^2 \tag{2.37}$$

Loi normale centrée réduite Cette loi est un cas particulier de la loi normale, ou la variable x est centrée réduite. Une variable centrée réduite a

- Une moyenne nulle
- Une variance égale à 1
- Un écart type égal à 1

La fonction de densité de probabilité devient alors

$$p(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp -\frac{1}{2}(x)^2 \tag{2.38}$$

2-2

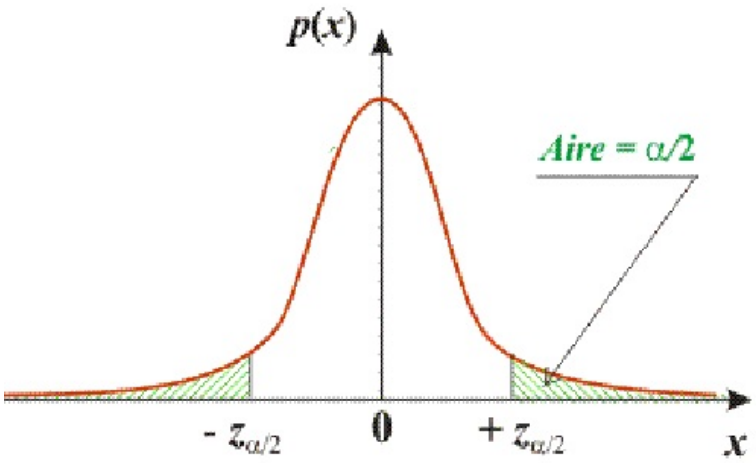


FIGURE *2.3: Représentation graphique de la fonction de densité de la loi normale centrée réduite*

2.4.3 Coefficient de Regression R^2

Un modèle, que l'on qualifiera de bon, possédera des estimations $\widehat{y}_i$ proches des vraies valeurs de y_i Les deux quantités SCT et SCE modèle sont des sommes de carrés donc toujours positives ou nulles et telles que : SCE < SCT. Le rapport des ces

deux quantités est compris entre 0 et 1 que l'on appelle le coefficient de détermination R^2 qui défini par

$$R^2 = \frac{\|\widehat{Y} - \overline{y}\mathbf{1}\|^2}{\|Y - \overline{y}\mathbf{1}\|^2} \tag{2.39}$$

$$R^2 = \frac{\sum(\widehat{Y}_i - \overline{y})^2}{\sum(Y_i - \overline{y})^2} = \frac{b_i \sum(x_i - \overline{x})^2}{\sum(Y_i - \overline{y})^2} = \frac{cov(x_i, y_i)^2}{var(x_i)var(y_i)} \tag{2.40}$$

Le R^2 peut aussi s'écrire en fonction des résidus :

$$R^2 = \frac{SCE}{SCT} \tag{2.41}$$

On a :

$$SCT = SCE + SCR \tag{2.42}$$

$$R^2 = \frac{SCE}{SCT} = 1 - \frac{SCR}{SCT} \tag{2.43}$$

$$R^2 = 1 - \frac{\|\widehat{\varepsilon}\|^2}{\|Y - \widehat{y}\|^2} \tag{2.44}$$

Interprétation de R^2 : R^2 qui varie entre 0 et 1, mesure la proportion de variation totale de y autour de la moyenne expliquée par la régression. Plus la valeur de R^2 sera proche de 1 (cas idéal) et plus les valeurs prédites et observées sont corrélées. Un R^2 faible signifie que le modèle à un faible pouvoir explicatif et les descripteurs sont sans effet sur la réponse.

Il n'est pas recommandé d'utiliser R^2 pour comparer des modèles, par exemple, si on hésite entre $y = \beta_o + \beta_i x_i$ et $y = \beta_o + \beta_1 x_1 + \beta_2 X_2$, le coefficient R^2 nous dira toujours de choisir le second modèle car son R^2 sera plus important (on projette sur un espace plus grand), même si la variable X_2 est sans rapport avec la réponse y.
Pour comparer des modèles, en pénalisant les modèles les plus complexes, il existe de nombreux indicateurs. Parmi ceux-ci, le coefficient de détermination ajusté découle simplement de notre table d'analyse de variance.

2.4.4 Coefficient de Regression adjusté R^2_{adj}

Le coefficient de détermination ajusté R^2_{adj} est défini par

$$R^2_{adj} = 1 - \frac{n}{n-p}\frac{\|\varepsilon\|^2}{\|Y\|^2} \tag{2.45}$$

Et,

$$R^2_{adj} = 1 - \frac{n-1}{n-p}\frac{\|\varepsilon\|^2}{\|Y - \widehat{y}\mathbf{1}\|^2} \tag{2.46}$$

L'ajustement correspond à la division des normes au carré par leur degré de liberté (ou dimension du sous-espace auquel le vecteur appartient) respectif.

$$R^2_{adj} = 1 - \frac{n - intercept}{n-p-1}(1-R^2) \tag{2.47}$$

Avec Intercept=0, si il n'y a pas de constante β_0 a l'origine si non Intercept=1. Cette formule indique notamment que R^2_{adj} est toujours inférieur à R^2, et ceci d'autant plus que le modèle contient un grand nombre de prédicteurs.

2.4.5 Déviation standard (SD)

La fiabilité de la prédiction de la variable dépendante peut être évaluée également parla valeur de l'erreur type d'estimation« s »ou déviation standard (SD).

$$S^2 = MSE = \frac{\sum_i (Y_i - \widehat{Y}_i)^2}{n-p-1} \tag{2.48}$$

$$RootMSE = SD = s \tag{2.49}$$

L'estimation de l'erreur-type appelé aussi déviation standard SD est une mesure de la dispersion des valeurs observées de la variable dépendante sur la droite de régression(de surface). Les petites valeurs de SD signifient un bon ajustement statistique du modèle et une forte fiabilité de la prédiction.

2.4.6 Critère de validation croisée : PRESS

La somme des erreurs de prédiction « PREdiction Sum of Squares » (PRESS) est définie par

$$PRESS = \sum_i \varepsilon_i^2 \tag{2.50}$$

Ce critère permet de sélectionner les modèles ayant un bon pouvoir prédictif. (On cherche toujours le PRESS le plus petit).

2.4.7 Test de Fisher-Snedecor

Ce test à surtout un intérêt dans le cadre de la régression multiple, c'est à dire avec p régresseurs.

$$\mu_i = E(y_i) = \beta_0 + \beta_1 x_1 + ... + \beta_p x_p \tag{2.51}$$

Ce test permet de connaître l'apport global de l'ensemble des variables $x_1, x_2, ..., x_p$ à la détermination de y. On calcul la statistique de test

$$F = \frac{MS\ \ Model}{MS\ \ error} \tag{2.52}$$

avec

$$MS\ \ model = \frac{SS\ \ Model}{P} \tag{2.53}$$

et

$$MS\ \ error = \frac{SS\ \ error}{n-P-1} \tag{2.54}$$

Les expressions (2.53) et (2.54) représentent respectivement la somme des carrés des écarts moyens pour le modéle et pour l'erreur.

Le test de Fisher mesure le rapport entre la variance de la variable dépendante expliquée et non expliquée par le modèle de régression. En d'autres termes le test Fisher-Senedecor permet de tester l'hypothèse nulle selon laquelle chaque β est significativement différent de zéro, ce qui est signe d'une relation évidente entre la variable expliquée et les variables explicatives.
Intuitivement, nous rejetterons l'hypothèse nulle lorsque la somme des carrés expliquée par la régression est grande. En d'autres termes, la région critique de ce test est de la forme $(F > seuil)$. Si la quantité F observée dépasse le seuil, on rejette l'hypothèse **H0** dans le cas contraire, on conserve **H1**.
Pour la régression simple, ce test porte uniquement sur le paramètre. Ce test fournit un moyen d'apprécier la régression dans son ensemble, ce qui ne signifie pas que chacun des coefficients de la régression soit significativement différent de 0. La statistique F est liée au coefficient de détermination par la relation suivante

$$F = \frac{R^2}{1-R^2}\frac{n-p-1}{p} \tag{2.55}$$

2.4.8 Test de Student (t-test)

Le test de Student, ou t-test, est un ensemble de tests d'hypothèse paramétriques où la statistique calculée suit une loi de Student lorsque l'hypothèse nulle est vraie.

Un test de Student peut être utilisé notamment pour tester statistiquement l'hypothèse d'égalité de l'espérance de deux variables aléatoires suivant une loi normale et de variance inconnue. Il est aussi très souvent utilisé pour tester la nullité d'un coefficient dans le cadre d'une régression linéaire.

Le principe du test de Student est de déterminer si la valeur d'espérance μ d'une population de distribution normale et d'écart type σ non connu est égale à une valeur déterminée μ_0. Pour ce faire, on tire de cette population un échantillon de taille n dont on calcule la moyenne $\overline{x}$ et l'écart-type empirique s.
Selon l'hypothèse nulle, la distribution d'échantillonnage de cette moyenne se distribue elle aussi normalement avec un écart type $s/\sqrt{n}$.

La variable :

$$t = (\overline{x} - \mu_0)/(s/\sqrt{n}), \tag{2.56}$$

Suit alors une loi de Student avec n-1 degrés de liberté. où :

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2} \tag{2.57}$$

2.4.9 La valeur p "P-Value"

Pour éviter de raisonner seulement sur F, les programmes destinés à l'analyse statistique fournit la p-value associée au F observé. La p-value est le niveau de significativité du test de Fisher-Snedecor, c'est-à-dire la probabilité de dépasser le F observé si l'hypothèse nulle est vraie.
On compare la p-value au risque α choisi (par exemple $\alpha = 0.05$), si P-value $\leq \alpha$, alors on rejette l'hypothèse nulle $\beta_1 = ... = \beta_p = 0$.

Bibliographie

[1] Antoniadis A., Berruyer J., Carmona R. Règression non linéaire et applications. (**1992**), éditions Economica.

[2] Birkes D., Dodge Y. Alternative methods of regression. (**1993**), Wiley.

[3] Brown P., Fearn T., Vannucci M. Bayesian wavelet regression on curves with application to a spectroscopic calibration problem. J. Am. Stat. Assoc., 96, (**2001**), 398.

[4] Cleveland W.S. Robust locally weighted regression and smoothing scatterplots. J. Am. Stat. Assoc., 74, (**1979**), 829.

[5] Cook R.D. Detection of influential observation in linear regression. Technometrics, 19, (**1977**), 15.

[6] De Jong, S. PLS shrinks. J. Chemometrics, 9, (**1995**), 323.

[7] Dodge Y., Rousson V. Analyse de régression appliquée. (**2004**), Dunod.

[8] Byun K., Mo Y., Gao J., New insight on the origin of the unusual acidity of Meldrum's acid from ab initio and combined QM/MM simulation study. J. Am. Chem. Soc., 123, (**2001**), 3974.

[9] Legendre A. M. (**1805**) Nouvelles Méthodes pour la Détermination des Orbites des Comètes. Courcier, Paris

[10] Gauss C. F. (**1809**) Theoria Motus Corporum Coelestium in Sectionibus Conicis Solem Ambientium Frid. Perthes et I H Besser, Hamburgi.

[11] Francis Galton, Regression Towards Mediocrity in Hereditary Stature, Journal of the Anthropological Institute, 15, (**1886**), 246.

[12] Alvin C. Rencher and G. Bruce Schaalje, Linear Models in Statistics, Second Edition, (**2008**), John Wiley & Sons, Inc.

[13] Weisberg S. Applied Linear Regression. (**1985**), New York : Wiley

[14] Morrison D. F. Applied Linear Statistical Methods. Englewood Cliffs, NJ : (**1983**), Prentice-Hall.

[15] **(a)** Myers R. H. Classical and Modern Regression with Applications (2nd ed.), (**1990**), Boston : Duxbury Press. **(b)** Myers, R. H. and J. S. Milton, A First Course in the Theory of Linear Statistical Models, (**1991**), Boston : PWS-Kent.

[16] Montgomery D. C. and E. A. Peck, Introduction to Linear Regresion Analysis (2nd ed.), (**1992**), New York : Wiley.

[17] Jørgensen B. The Theory of Linear Models. (**1993**), New York : Chapman & Hall.

[18] **(a)** Graybill F. A. Theory and Application of the Linear Model. (**1976**), North Scituate, MA : Duxbury Press. **(b)** Graybill, F. A. and H. K. Iyer, Regression Analysis : Concepts and Applications. North Scituate, (**1994**), MA : Duxbury Press.

[19] **(a)** Hocking R. R. The analysis and selection of variables in linear regression. Biometrics 32, (**1976**), 1–51. **(b)** The Analysis of Linear Models. Monterey, (**1985**), CA : Books/Cole. **(c)**, Methods and Applications of Linear Models. (**1996**), New York : Wiley.

[20] Ryan T. P. Modern Regression Methods, (**1997**), New York : Wiley.

[21] Fox J. Applied Regresion Analysis, Linear Models, and Related Methods. Thousand Oaks, (**1997**), CA : SAGE Publications.

[22] Kutner M. H., C. J. Nachtsheim., J. Neter., W. Li., Applied Linear Statistical Models, (5th ed.), (**2005**), New York, McGraw-Hill/Irwin.

[23] **(a)** Christensen R. Plane Answers to Complex Questions : The Theory of Linear Models (2nd ed.), (**1996**), New York : Springer-Verlag. **(b)** Log-Linear Models and Logistic Regression (2nd ed.), (**1997**), New York : Springer-Verlag.

CHAPITRE 3

Les méthodes de la chimie quantique

Dans le cadre de ce chapitre, nous effectuerons un bref rappel des notions théoriques dans le but de présenter la théorie de la fonctionnelle de la densité. Il ne s'agit pas d'une description exhaustive et le lecteur peut consulter de nombreux ouvrages spécialisés pour plus de détail. [1–5]

3.1 Notions de chimie quantique

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

$$H\psi = E\psi \tag{3.1}$$

La résolution exacte de l'équation (3.1) n'est possible que pour l'atome d'hydrogène et les systèmes hydrogénoïdes. Pour les systèmes poly-électroniques, il est nécessaire de faire appel aux méthodes d'approximation.

La première approximation en chimie quantique est de considérer l'équation de *Schrödinger* (3.1) [6] non relativiste indépendante du temps où l'hamiltonien est défini par :

$$H = -\frac{1}{2}\sum_i \Delta_i - \frac{1}{2}\sum_A \Delta_A - \sum_i \sum_A \frac{Z_A}{r_{iA}} + \sum_i \sum_{j>i} \frac{1}{r_{ij}} + \sum_A \sum_{B>A} \frac{Z_A Z_B}{r_{AB}} \tag{3.2}$$

Les indices qui apparaissent sous les symboles de sommation s'appliquent aux électrons (i et j) et aux noyaux (A et B). Dans l'expression (3.2), les deux premiers termes correspondent aux opérateurs associés à l'énergie cinétique des électrons et des noyaux, le troisième terme représente l'attraction coulombienne entre les noyaux et les électrons, tandis que les deux derniers décrivent la répulsion entre les électrons et entre les noyaux.

Trois autres approximations sont adoptées et employées : l'approximation de *Born-Oppenheimer*, l'approximation d'orbitales moléculaires et l'approximation C.L.A.O. (Combinaison Linéaire d'Orbitales Atomiques. *LCAO en anglais*). Nous allons décrire brièvement le principe des deux premières. La troisième sera présentée ensuite avec la méthode *Hartree-Fock*.

3.1.1 L'approximation de Born-Oppenheimer

L'approximation de *Born-Oppenheimer* [7] trouve son origine dans le fait que les noyaux possèdent une masse beaucoup plus importante que celle des électrons et qu'il est alors possible de considérer leur mouvement comme étant très lent par rapport à celui des électrons. On peut donc supposer que les électrons se déplacent dans un champ de noyaux fixes. Ainsi, dans le cadre de cette approximation, l'énergie cinétique des noyaux peut être supposée constante et nulle et la répulsion entre les différentes paires de noyaux considérées également comme constante, les termes restant de l'équation (3.1) permettent alors de définir l'hamiltonien électronique :

$$H_e = -\frac{1}{2}\sum_i \Delta_i - \sum_i \sum_A \frac{Z_A}{r_{iA}} + \sum_i \sum_{j>i} \frac{1}{r_{ij}} = \sum_i \mathbf{H}^c(i) + \sum_i \sum_{j>i} \frac{1}{r_{ij}} \tag{3.3}$$

Cet hamiltonien est alors utilisé pour résoudre l'équation de *Schrôdinger* électronique

$$\mathbf{H}_e \psi_e = E_e \psi_e \tag{3.4}$$

ψ_e est la fonction d'onde électronique. Elle dépend explicitement des coordonnées électroniques et paramétriquement des coordonnées nucléaires. Dans le cas d'un système multiélectronique, la fonction d'onde doit changer de signe lors de la permutation des coordonnées de deux électrons (principe de *Pauli* [8])
$E_e$représente l'énergie électronique, pour obtenir l'énergie totale E' dans un champ de noyaux fixes, on ajoute un terme de répulsion nucléaire à l'énergie électronique :

$$E' = E_e + \sum_A \sum_{A>B} \frac{Z_A Z_B}{r_{AB}} \tag{3.5}$$

Dans la suite de ce manuscrit les symboles H, ψ et E désignerons respectivement l'hamiltonien électronique, la fonction d'onde électronique et l'énergie totale calculée pour des positions fixes des noyaux.

3.1.2 L'approximation d'Orbitales Moléculaires

La fonction d'onde la plus simple qui respecte le principe de *Pauli* peut s'écrire sous la forme d'un déterminant, appelé déterminant de *Slater* [9] . Ce déterminant pondéré par un facteur de normalisation est construit à partir d'un ensemble de fonctions monoélectroniques, ou spinorbitales ϕ définies comme le produit d'une fonction spatiale, ou orbitale moléculaire (OM)ψ . par une fonction de spin α ou β :

$$\Psi = \frac{1}{\sqrt{N!}} \begin{vmatrix} \phi_1(1) & \phi_1(2) & \dots & \phi_1(N) \\ \phi_2(1) & \phi_2(2) & \dots & \phi_2(N) \\ \vdots & \vdots & \ddots & \vdots \\ \phi_N(1) & \phi_N(2) & \dots & \phi_N(N) \end{vmatrix} \tag{3.6}$$

Et

$$\phi_i(x) = \psi_i(x).\alpha_i(x) \tag{3.7}$$

$$\phi_j(x) = \psi_j(x).\alpha_j(x) \tag{3.8}$$

Ce déterminant peut également s'écrire plus simplement :

$$\psi = |\phi_1\phi_2\phi_3.....\phi_n| \tag{3.9}$$

3.2 La Méthode Variationnelle

Pour des systèmes multi-électroniques, l'équation de $Schrôdinger$ indépendante du temps est solutionnée avec la méthode variationnelle qui garantit que l'énergie du système déterminée par :

$$E|\psi| = \langle \psi|\widehat{H}|\psi\rangle \tag{3.10}$$

Sera supérieure ou égale à l'énergie exacte de l'état fondamental E_o, c'est-à-dire que

$$E|\psi| = E_0 \tag{3.11}$$

Selon la méthode variationnelle, les paramètres de la fonction d'essai sont variés jusqu'à ce que la valeur attendue de l'énergie atteigne un minimum ce qui correspond à minimiser la fonctionnelle $E|\psi|$ par rapport à toutes les N fonctions d'onde multi-électroniques. Mathématiquement, ceci se traduit par

$$E_0 = minE|\psi| \tag{3.12}$$

En pratique, on détermine les extremums de $\langle \psi|\widehat{H}|\psi\rangle$, avec la contrainte $\langle \psi|\psi\rangle = 1$, en solutionnant

$$\delta\left[\langle \psi|\widehat{H}|\psi\rangle - E\langle \psi|\psi\rangle\right] = 0 \tag{3.13}$$

Où E est employée comme multiplicateur de $Lagrange$. Cette équation permet d'aller de N et $\upsilon(r)$ vers ψ et de là, avec l'équation variationnelle (3.10), à l'énergie du système. L'énergie totale du système E est donc une fonctionnelle de N et de $\upsilon(r)$.

3.3 La Méthode Hartree-Fock(HF)

3.3.1 Approximation du champ moyen de Hartree

L'approximation du champ moyen, proposée par *Hartree* [10] en **1927**, consiste à remplacer l'interaction d'un électron avec les autres électrons par l'interaction de celui-ci avec un champ moyen créé par la totalité des autres électrons ; ce qui permet de remplacer le potentiel biélectronique qui exprime la répulsion entre l'électron i et les autres électrons j, ($j \neq i$) par un potentiel monoélectronique moyen de l'électron i de la forme U(i). Par conséquent et en se basant sur le théorème des électrons indépendants, nous pouvons écrire la fonction d'onde totale comme le produit de fonctions d'onde mono-électroniques, dont la forme abrégée pour un système à couches fermées est n=2m :

$$\Psi(1,2,..,n) = \frac{1}{(n!)^{1/2}} |\phi_1(1)\overline{\phi_1}(2)...\phi_m(2m-1)\phi_m(2m)| \tag{3.14}$$

avec ;

$$\phi_1(1) = \psi_1(1).\alpha_1(1)$$
$$\phi_1(2) = \psi_1(2).\alpha_1(2)$$

Pour décrire la méthode de *Hartree-Fock*, nous nous placerons dans le cas d'un système à couches fermées. Pour les systèmes à couches ouvertes, la même démarche mathématique peut être adoptée en traitant séparément les électrons α et les électrons β. Celle-ci conduit à un ensemble d'équations pratiquement analogues à celles décrites dans la suite de ce paragraphe.
La théorie HF utilise le principe variationnel et se base sur le fait que l'énergie calculée pour un état électronique donné d'un système décrit par une fonction d'onde de type variationnelle est toujours supérieure à l'énergie que l'on obtiendrait pour ce même état en utilisant une fonction d'onde exacte.

3.3.2 Équations de Hartree-Fock

La fonction d'onde polyélectronique de *Hartree* (3.14) ne vérifie ni le principe d'indiscernabilité des électrons ni le principe d'exclusion de *Pauli* [8]. Pour tenir compte de ces deux principes, *Fock* [11] a proposé d'écrire la fonction d'onde totale ψ sous forme

de déterminant de *slater* [9].
La fonction d'onde $\psi = |\phi_1\phi_2\phi_3.....\phi_n|$ construite sous la forme d'un déterminant de *slater* est utilisée pour résoudre l'équation (3.1-3.4) et calculer l'énergie électronique correspondante qui se décompose en une somme de termes mono et bi-électroniques :

$$E^{HF} = 2\sum_i h_{ii} + \sum_i \sum_{j>i} (2J_{ij} - K_{ij}) \tag{3.15}$$

Avec :

$$h_{ii} = \int \phi_i^*(1)\widehat{h}_1\phi_i(1)d\overrightarrow{r}_1$$
$$J_{ij} = \int \phi_i^*(1)\phi_j^*(2)\frac{1}{r_{12}}\phi_i(1)\phi_j(2)d\overrightarrow{r}_1 d\overrightarrow{r}_2$$
$$K_{ij} = \int \phi_i^*(1)\phi_j^*(2)\frac{1}{r_{12}}\phi_j(1)\phi_i(2)d\overrightarrow{r}_1 d\overrightarrow{r}_2$$

Où

$$\widehat{h}_1 = -\frac{1}{2}\Delta_1^2 - \sum_k \frac{Z_k e^2}{r_{1k}}$$

Dans cette expression, J et K sont respectivement des intégrales de Coulomb et d'échange, qui caractérisent les répulsions entre électrons. Les intégrales d'échange résultent de la nature antisymétrique de la fonction d'onde multiélectronique.
En s'appuyant sur le principe variationnel, il s'agir de trouver les meilleures spinorbitales, et par conséquent les meilleures orbitales moléculaires, c'est-à-dire celles rendant l'énergie **E** la plus basse possible. Si on fait l'hypothèse que ces orbitales correspondent à un minimum de l'énergie, on doit vérifier qu'une petite modification apportée à une orbitale quelconque, qui n'en modifie ni la norme ni l'orthogonalité aux autres orbitales, n'entraîne pas de variation de l'énergie. Cette condition impose aux orbitales d'être fonctions propres d'un opérateur F appelé *opérateur de Fock*.
Les équations de *Hartree-Fock* [12] correspondantes déterminent ces orbitales :

$$F(1)\phi_i(1) = \varepsilon_i\phi_i(1) \tag{3.16}$$

$$F(1) = \widehat{h}_i(1) + V^{eff}(1) = \widehat{h}_i(1) + \sum(2\widehat{J}_j(1) - \widehat{k}_j(1)) \tag{3.17}$$

V^{eff} est appelé potentiel effectif formé par les noyaux et le champ moyen des électrons. ε_i est l'énergie de l'orbitale i correspondante.

3.3.3 L'approximation C.L.O.A.

Une des méthodes permettant de résoudre l'équation (3.16) consiste à développer les orbitales moléculaires en combinaisons linéaires de fonctions de base. Le choix de ces fonctions de base se porte généralement sur les orbitales atomiques (OA) du système et conduit â l'approximation C.L.O.A. (Combinaison Linéaire d'Orbitales Atomiques) :

$$\psi_i = \sum_\nu C_{\nu i}\phi_\nu \tag{3.18}$$

Les symboles latins servent à définir les OM alors que les symboles grecs sont utilisés pour représenter les OA. Dans le cadre de cette approximation, il s'agit de trouver les meilleurs coefficients $C_{\nu i}$ qui minimisent l'énergie électronique **E**.
En substituant l'équation (3.18) dans l'expression (3.16) et en multipliant chaque membre par ϕ_μ on aboutit aux *équations de Roothaan* [13] :

$$\sum_\nu \mathbf{F}_{\mu\nu}C_{\nu i} = \varepsilon_i \sum_\nu S_{\mu\nu}C_{\nu i} \tag{3.19}$$

Où $S_{\mu\nu}$ est un élément de la matrice de recouvrement et $\mathbf{F}_{\mu\nu}$, un élément de la matrice de *Fock* qui s'écrit :

$$\mathbf{F}_{\mu\nu} = H^c_{\mu\nu} + \sum_{\lambda\sigma} P_{\lambda\sigma}\left[<\mu\sigma|\nu\lambda> - \frac{1}{2}<\mu\sigma|\lambda\nu>\right] \tag{3.20}$$

$P_{\lambda\sigma}$ est un élément de la matrice densité définie par

$$P_{\lambda\sigma} + 2\sum_i^{occ} C^*_{\lambda i}C_{\sigma i} \tag{3.21}$$

Dans le cas du formalisme non restreint, des équations analogues aux équations de *Roothaan* sont construites. Elles portent le nom d'équations de *Berthier-Pople-Nesbet* [14]

3.4 Méthodes Post-SCF

La méthode *Hartree-Fock-Roothaan* présente l'inconvénient majeur de ne pas tenir compte de la corrélation électronique qui existe entre le mouvement des électrons. Ceci rend cette méthode relativement restreinte dans le calcul quantitatif des propriétés thermodynamiques telles que l'enthalpie d'activation, l'énergie de Gibbs de réactions, énergies de dissociation.

Ces propriétés peuvent être calculées d'une manière efficace par les méthodes Post-SCF en tenant compte de la corrélation électronique. Les deux familles importantes de méthodes qui ont été développées sont celles d'interaction de configurations (CI) [15, 16] et la théorie des perturbations *Møller-Plesset* d'ordre n (MPn) et les méthodes DFT. L'énergie de corrélation d'un système correspond à la différence entre l'énergie *Hartree-Fock* et l'énergie exacte non-relativiste du système : $E_{corr} = E_{HF} - E$
Les techniques Post-HF sont en général très efficaces pour retrouver l'énergie de corrélation, mais cependant à l'heure actuelle elles sont, pour la majeure partie d'entre-elles, trop lourdes pour être applicables à des systèmes dont le nombre d'atomes est grand. Il s'est ainsi parallèlement développé à ces techniques un modèle alternatif qui a atteint le statut de théorie à la fin des années **60**. La théorie de la fonctionnelle de la densité (DFT) est actuellement la seule permettant l'étude de systèmes chimiques de grande taille avec la prise en compte des effets de la corrélation électronique de manière satisfaisante.

3.5 La Théorie de la fonctionnelle de la densité (DFT)

La théorie de la fonctionnelle de la densité (DFT) s'est beaucoup développée ces dernières années. Dans cette approche l'énergie de l'état fondamental d'un système est une fonctionnelle d'une densité électronique tridimensionnelle. L'application du principe variationel donne les équations appelées équations de *Kohn-Sham* qui sont similaires aux équations de *Hartree-Fock*. En principe, il suffit de remplacer la contribution d'échange de l'opérateur de *Fock* par un potentiel d'échange et de corrélation qui correspond à la dérivation de la fonctionnelle d'énergie d'échange et de corrélation par rapport à la densité. Le point crucial en DFT est que l'énergie d'échange et de corrélation n'est pas connue de façon exacte. Néanmoins les formules approchées pour cette énergie donnent des résultats qui sont comparables ou meilleurs que ceux donnés par MP2 à un moindre coût de ressource informatique.
Les premières approximations de la DFT sont similaires à celles appliquées aux méthodes HF. L'équation de *Schrôdinger* est non-dépendante du temps et non-relativiste. A partir de l'approximation de *Born-Oppenheimer* le formalisme et les approximations divergent.

3.5.1 Théorèmes de Hohenberg et Kohn

Dans un système électronique le nombre d'électrons par unité de volume, dans un état donné, est appelé la densité électronique pour cet état [17]. Cette quantité est désignée par $\rho(\overrightarrow{r})$ et sa formule, en terme de ψ , pour l'électron 1, est :

$$\rho(\overrightarrow{r}) = \int .. \int |\psi(\overrightarrow{r_1}, \overrightarrow{r_2}, \overrightarrow{r_n})|^2 d\overrightarrow{s_1}, d\overrightarrow{r_2}, ...d\overrightarrow{r_n} \tag{3.22}$$

Avec s_1 comme coordonnée de spin.
Cela correspond à une simple fonction à trois variables x, y, et z intégrant sur le nombre total d'électrons. La densité électronique possède la propriété suivante :

$$\int \rho(\overrightarrow{r}) d\overrightarrow{r} = n \tag{3.23}$$

Hohenberg et Kohn ont prouvé [18] que l'énergie moléculaire fondamentale E_0, la fonction d'onde et toutes les autres propriétés électroniques sont uniquement déterminées par la connaissance de la densité électronique $\rho(\overrightarrow{r})$ en chaque point $\overrightarrow{r}$ du volume moléculaire. E_0 est une fonctionnelle de $\rho(\overrightarrow{r})$ et est représentée par $E_0[\rho]$ avec $\rho = \rho(\overrightarrow{r})$
En d'autres termes, les propriétés de l'état fondamental sont totalement déterminées par le nombre n et le potentiel externe dû au champ des noyaux $\upsilon(\overrightarrow{r})$.

3.5.1.1 Premier théorème de Hohenberg et Kohn

La densité électronique $\rho(\overrightarrow{r})$, pour l'état fondamental non dégénéré d'un système à n électrons, détermine $\upsilon(\overrightarrow{r})$ Autrement dit, $\rho(\overrightarrow{r})$ détermine de manière unique la fonction d'onde de l'état fondamental ψ et de là toutes les autres propriétés du système avec l'equation (3.22)
On peut alors utiliser la densité électronique comme variable de base pour la résolution de l'équation de *Schrôdinger* électronique. Étant donné que $\rho(r)$ est liée au nombre d'électrons du système, elle peut en effet également déterminer les fonctions propres de l'état fondamental ainsi que toutes les autres propriétés électroniques du système ; si n est le nombre d'électrons du système.
Rappelons l'expression de l'hamiltonien électronique d'un système polyélectronique :

$$H = -\frac{1}{2}\sum_i^n \Delta_i + \sum_{i>j}^n \frac{1}{r_{ij}} + \sum_i^n V(r_i) \tag{3.24}$$

Avec

$$V(r_i) = -\sum_{\alpha} \frac{Z_\alpha}{r_{i\alpha}}$$

$V(r_i)$: Potentiel externe de l'électron i.
Ce potentiel correspond à l'attraction de $l'e^-(i)$ avec tous les noyaux qui sont externes par rapport au système d'électrons.
Connaissant la densité électronique $\rho(r)$ d'un système, on a donc accès au nombre d'électrons, au potentiel externe, ainsi qu'à l'énergie totale E[$\rho(r)$]. Celle-ci peut s'écrire comme une somme de trois fonctionnelles :

$$E[\rho] = V_{ne}[\rho] + T[\rho] + V_{ee}[\rho] \tag{3.25}$$

$$V_{ne}[\rho] = \int \rho(r) v(r) dr \tag{3.26}$$

$$T[\rho] = \int \left[-\frac{1}{2} \nabla^2 \rho(r) \right] dr \tag{3.27}$$

Le terme $V_{ee}[\rho]$ est composé de deux parties ; la première correspond à l'interaction coulombienne classique $J[\rho]$, et la seconde partie dite non-classique est appelée « énergie d'échange et de corrélation $K[\rho]$.

$$J[\rho] = \frac{1}{2} \int \int \frac{1}{r_{12}} \rho(r_1) \rho(r_2) dr_1 dr_2$$
$$K[\rho] = \frac{1}{4} \int \int \frac{1}{r_{12}} \rho(r_1, r_2) \rho(r_1 r_2) dr_1 dr_2$$

Par conséquent, la fonctionnelle de l'énergie peut s'écrire :

$$E_0[\rho] = \int \rho_0(r) v(r) dr + F[\rho_0]$$

Où

$$F[\rho_0] = T[\rho_0] + V_{ee}[\rho_0]$$

est la fonctionnelle universelle de *Hohenberg et Kohn* $F[\rho_0]$ est une fonctionnelle prenant en compte tous les effets interélectroniques ; elle est indépendante du potentiel externe, et elle est donc valable quelque soit le système étudié. La connaissance de $F[[\rho]]$ permet l'étude de tous les systèmes moléculaires, malheureusement la forme exacte de cette fonctionnelle est à l'heure actuelle loin d'être connue, et il faut avoir recours à des approximations la fonctionnelle $F[[\rho]]$ est inconnue.

3.5.1.2 Second théorème de Hohenberg et Kohn

Le second théorème de *Hohenberg-Kohn* [18, 19] découle du premier théorème et reconsidère le principe variationnel d'énergie en fonction de la densité électronique. Il dit que pour une densité d'essai $\rho(r)$,tel que$\rho(r) > 0$ et $\int \rho(r)dr = n$,

$$E_0 \leq E_v[\widetilde{\rho}] \tag{3.28}$$

Où $E_\upsilon[\overrightarrow{\rho}]$ est la fonctionnelle d'énergie de $E_v[\widetilde{\rho}] = T[\widetilde{\rho}] + V_{ne}[\widetilde{\rho}] + V_{ee}[\widetilde{\rho}]$
La condition pour qu'une fonctionnelle telle que E_0 admette un extremum est que sa dérivée fonctionnelle s'annule. D'après la définition :

$$\delta E = \int \frac{\delta E}{\delta \rho} d\rho dr = 0 \tag{3.29}$$

La relation $\delta E = 0$ est donc vérifiée si $\delta E/\delta\rho = 0$:
La résolution du problème consiste dés lors à chercher à minimiser $E[\rho]$ avec la contrainte $\int \widetilde{\rho}(r)dr = n$.
Finalement on obtient l'équation fondamentale de la DFT :

$$\mu = \frac{\delta E[\rho]}{\delta \rho} = v(r) + \frac{\delta F_{HF}[\rho]}{\delta \rho} \tag{3.30}$$

Où la quantité μ est appelée « potentiel chimique » du système.
L'avantage de travailler avec ρ , bien que des expressions approchées pour $E_0(\rho)$ doivent être utilisées, réside dans la résolution plus facile, pour un niveau comparable de précision, des équations de la théorie de la fonctionnelle de la densité amenant à ρ par rapport aux méthodes ab initio correspondantes. De plus, les théorèmes de *Hohenberg et Kohn* fournissent les fondements théoriques pour l'obtention de méthodes de calcul toujours plus précises.

3.5.2 Les équations de Kohn-Sham

L'absence d'une expression analytique pour l'hamiltonien dans les théorèmes de *Hohenberg et Kohn* qui ne nous disent pas comment calculer E_0 à partir de ρ , ou comment trouver ρ sans trouver ψ en premier, a amené *Kohn et Sham* en **1965** à reformuler le problème en introduisant des orbitales moléculaires ϕ_i et en scindant l'hamiltonien en terme classique et résiduel [20], l'hamiltonien de système de référence d'un système à n-électrons peut-être écrite sans approximation comme [21, 22] :

$$\widehat{H}_s = \sum_{i=1}^{n} \left[-\frac{1}{2} \int \nabla_i^2 + \upsilon_s(r_1) \right] = \sum_{i=1}^{n} h_i^{KS} \tag{3.31}$$

Avec

$$h_i^{KS} = -\frac{1}{2}\int \nabla_i^2 + \upsilon_s(r_1)$$

Par conséquent, les équations de *Kohn et Sham*, pour l'électron i, peuvent s'écrire comme suit :

$$h_i^{KS}\Phi_i^{KS} = \varepsilon_i^{KS}\Phi_i^{KS} \tag{3.32}$$

Φ_i^{KS} sont les orbitales de Kohn et Sham de l'électron i.

3.5.3 Expression du terme d'échange et de corrélation E_{xc}

Soit ΔT la différence de l'énergie cinétique entre le système réel (électrons interagissants) et le système fictif (électrons non interagissant). La quantité ΔT étant cependant faible :

$$\Delta T = T[\rho] - T_s[\rho]$$

Donc

$$\Delta V = V_{eff}[\rho] - \frac{1}{2}\int\int \frac{\rho(r_1)\rho(r_2)}{r_{12}} dr_1 dr_2$$

ΔV est la différence entre la vraie répulsion électron-électron et la répulsion coulombienne entre deux distributions de charge ponctuelle.
L'énergie s'écrit alors :

$$E[\rho] = \int \rho(r)\upsilon(r)d_r + T_s[\rho] + \frac{1}{2}\int\int \frac{\rho(r_1)\rho(r_2)}{r_{12}} dr_1 dr_2 + \Delta T[\rho] + \Delta V_{ee}[\rho] \tag{3.33}$$

La fonctionnelle d'énergie d'échange-corrélation est définie comme suit :

$$E_{xc}[\rho] = \Delta T[\rho] + \Delta V_{ee}[\rho] \tag{3.34}$$

$$E[\rho] = \int \rho(r)\upsilon(r)d_r + T_s[\rho] + \frac{1}{2}\int\int \frac{\rho(r_1)\rho(r_2)}{r_{12}} dr_1 dr_2 + E_{xc}[\rho] \tag{3.35}$$

Le terme de l'énergie échange-corrélation n'est pas connu de manière exacte et quelques approximations seront développées plus loin.
L'équation aux valeurs propres correspondante est de la forme :

$$\left[-\frac{1}{2}\nabla_i^2 - \sum_\alpha \frac{Z_\alpha}{r_{1\alpha}} + \int \frac{\rho(r_2)}{r_{12}} d_{r2} + v_{xc}(1)\right]\Phi_i^{KS}(1) = \varepsilon_i^{KS}\Phi_i^{KS}(1) \tag{3.36}$$

En appliquant le principe variationnel a l'equation (3.29) et en tenant compte de la contrainte de l'equation (3.23), et grâce au multiplicateur de *Lagrange*, les orbitales $\phi_i(\overrightarrow{r})i = 1,..n$ de l'equation (3.32) sont des solutions du système d'équations à un

électron de *Kohn-Sham*,
Où le potentiel d'échange et de corrélation v_{xc} est défini comme la dérivée fonctionnelle de E_{xc} en fonction de la densité électronique :

$$v_{xc}(r) = \frac{\partial E_{xc}[\rho(r)]}{\partial \rho(r)} \tag{3.37}$$

Il existe plusieurs approximations de ce potentiel d'échange-corrélation, l'équation aux valeurs propres peut également s'écrire sou la forme :

$$\left[-\frac{1}{2}\nabla_i^2 + V_{eff}\right]\Phi_i^{KS}(1) = \varepsilon_i^{KS}\Phi_i^{KS}(1) \tag{3.38}$$

V_{eff} est appelé potentiel effectif.
Les orbitales de *Kohn-Sham* Φ_i n'ont pas de signification physique mais permettent de calculer la densité électronique $\rho(\overrightarrow{r})$

$$\rho(\overrightarrow{r}) = \sum_{i=1}^{n} |\Phi_i(\overrightarrow{r})|^2 \tag{3.39}$$

Ainsi des applications pratiques de la DFT deviennent possibles avec les travaux de *Kohn-Sham* (KS) [20] qui donnent un ensemble d'équations monoélectroniques l'equation.(3.32) à partir desquelles on peut, obtenir la densité électronique, et ensuite l'énergie totale.
En résumé, le problème pour trouver $\rho(\overrightarrow{r})$ est toujours présent avec les équations de *Kohn-Sham*, mais, la fonctionnelle exacte n'est pas connue, notamment la partie dite d'échange et de corrélation. Cela signifie qu'une fonctionnelle approchée doit être utilisée dans les calculs moléculaires, comme celle décrite par *Dirac* [23] pour un gaz homogène d'électrons.
La dépendance explicite de la forme analytique de la fonctionnelle d'échange et de corrélation par rapport à la densité électronique equation (3.37) n'est pas connue. Toutes les expressions analytiques de la littérature sont des approximations plus ou moins sophistiquées.

3.5.3.1 Approximation de la densité locale (LDA)

L'expression la plus simple de l'énergie d'échange et de corrélation E_{xc} est celle provenant de l'approximation LDA (approximation de la densité locale) dans laquelle un gaz homogène d'électrons est pris en compte. La fonctionnelle d'échange et de corrélation peut se scinder en une somme sur l'énergie d'échange et sur l'énergie de

corrélation, pour une densité électronique constante, l'énergie d'échange est définie de manière exacte par la fonctionnelle de *Dirac* [23]
Cependant des versions simplifiées de la LDA étaient connues longtemps avant le développement formel de la théorie de la fonctionnelle de la densité, la méthode de *Hartree-Fock-Slater*, ou $X\alpha$ avec $\alpha = 2/3$, retient seulement la partie d'échange de l'expression ε_{xc}, les équations relatives à cette méthode du calcul de l'énergie d'échange peuvent être trouvées dans les travaux de *Dirac* [23] et *Slater* [24].
Hohenberg et Kohn ont montré que si ρ varie extrêmement lentement avec la position, l'énergie d'échange-corrélation $E_{xc}[\rho]$ peut s'écrire comme suit,

$$E_{xc}^{LDA}[\rho] = \int \rho(r)\varepsilon_{xc}\rho(r)dr \tag{3.40}$$

ε_{xc} étant l'énergie d'échange-corrélation par électron, cette quantité est exprimée comme la somme des deux contributions, énergie d'échange ε_x et énergie de corrélation ε_c :

$$\varepsilon_{xc}(\rho) = \varepsilon_x(\rho) + \varepsilon_c(\rho) \tag{3.41}$$

Avec

$$\varepsilon_x(\rho) = -\frac{3}{4}\left(\frac{3}{\pi}\right)^{1/3}(\rho(r))^{1/3} \tag{3.42}$$

Donc

$$E_x^{LDA} = \int \rho\varepsilon_x dr = -\frac{3}{4}\left(\frac{3}{\pi}\right)^{1/3}\int [\rho(r)]^{4/3}\,dr \tag{3.43}$$

Le terme de corrélation $\varepsilon_x(\rho)$ a été déterminé dans l'expression de la fonctionnelle (VWN) développée par *Vosko* [25].

3.5.3.2 Approximation de la densité de spin locale LSDA

Pour les molécules à couches ouvertes et les géométries des molécules prés de leur état de dissociation, l'approximation LSDA donne des résultats meilleurs que l'approximation LDA. Dans LDA, les électrons ayant des spins opposés ont les mêmes orbitales KS spatiales. En revanche, LSDA distingue entre les orbitales des électrons de spins opposés $\Phi_{i\alpha}^{KS}$ pour les e^- de spin α et $\Phi_{i\beta}^{KS}$ pour les e^- de spin pour les e^- de spin β. Par conséquent, on

$$E_{xc}^{LSDA} = E_{xc}\left[\rho^{\alpha}\rho^{\beta}\right] \tag{3.44}$$

3.5.3.3 Approximation du gradient généralisé (GGA)

Les approximations LDA et LSDA sont basées sur le modèle du gaz électronique uniforme dans lequel la densité électronique r varie très lentement avec la position. La correction de cette approximation, plus au moins grossière, nécessite l'inclusion des gradients des densités des spin ρ^{α} *et* ρ^{β}. L'énergie d'échange-corrélation, dans le cadre de l'approximation du gradient généralisé GGA (Generalized gradient approximation), s'écrit alors :

$$E_{xc}^{GGA}\left[\rho^{\alpha} \ , \ \rho^{\beta}\right] = \int f\left(\rho^{\alpha}(r)\rho^{\beta}(r)\right)\nabla\left(\rho^{\alpha}(r)\rho^{\beta}(r)\right)dr \tag{3.45}$$

Où f est une fonction des densités de spin et de leurs gradients. et E_{xc}^{GGA} est la densité d'énergie d'échange-corrélation, la difficulté réside dés lors dans la recherche d'expressions analytiques de E_{xc}^{GGA}
E_{xc}^{GGA} est divisée en deux contributions : échange et corrélation

$$E_{xc}^{GGA} = E_{x}^{GGA} + E_{c}^{GGA} \tag{3.46}$$

Terme d'échange En **1988**, *Becke* [26] a utilisé le terme d'échange pour apporter une correction de l'approximation LSDA :

$$E_{x}^{B88} = E_{x}^{LSDA} - b\sum_{\rho=\alpha,\beta}\int \frac{(\rho^{\alpha})^{4/3}\chi_{\rho}^{2}}{1+6b\chi_{\rho}\Upsilon(\chi_{\rho})}dr \tag{3.47}$$

Avec :

$$\chi_{\rho} = \frac{|\nabla\rho^{\alpha}|}{(\rho^{\alpha})^{4/3}}$$

$$\Upsilon(x) = ln\left[x+(x^{2}+1)^{1/2}\right]$$

Et

$$E_{x}^{LSDA} = -\frac{3}{4}\left(\frac{6}{\pi}\right)^{1/3}\int\left[(\rho^{\alpha})^{4/3}+(\rho^{\beta})^{4/3}\right]dr \tag{3.48}$$

Terme de corrélation La fonctionnelle de l'énergie de corrélation $E_{c}[\rho]$, corrigée à l'aide de l'approximation GGA, est exprimée à l'aide de la formule de *Lee-Yang-Parr* [29] :

$$E_{c}^{GGA} = E_{c}^{LYP} \tag{3.49}$$

3.5.4 Nomenclature des fonctionnelles

Les fonctionnelles d'échange et de corrélation peuvent adopter des formes mathématiques souvent complexes. De manière à simplifier les notations, la convention est de noter les fonctionnelles du nom de leur(s) auteur(s) suivi de la date de publication dans le cas ou un même groupe a publié plusieurs fonctionnelles différentes. La fonctionnelle d'échange électronique développée par *Axel Becke* en **1988** est ainsi notée B88 [26] et la fonctionnelle de corrélation publiée par le même auteur en **1995** est notée B95. la fonctionnelles de *Perdew* en **1986** (P86) [27] Dans le cas où plusieurs auteurs sont impliqués dans le développement, les initiales de ceux-ci sont utilisées pour symboliser la fonctionnelle. la fonctionnelle de *Perdew et Wang* en **1991** (PW91) [28], la fonctionnelle de corrélation (LYP) [29–31] est ainsi nommée du nom de ses trois auteurs *Lee, Yang et Parr*. les fonctionnelles hybrides telles B3LYP [29, 30] ou G96LYP [29–33].

3.5.5 Fonctionnelle hybride B3LYP

La fonctionnelle hybride B3LYP (Becke 3-parameters *Lee-Yang-Parr*) est une fonctionnelle à trois paramètres combinant les fonctionnelles d'échange local, d'échange de *Becke* et d'échange HF, avec les fonctionnelles de corrélation locale (VWN) et corrigée du gradient de *Lee, Yang et Parr*.

$$E_{xc}^{B3LYP} = (a)E_x^{HF} + (1-a)E_x^{LSDA} + (b)E_x^{GGA} + (c)E_c^{GGA} + (1-c)E_c^{LSDA} \tag{3.50}$$

Dans cette expression, E_x^{HF} l'énergie d'échange exact.
E_x^{LSDA} est l'approximation la densité de spin locale(LSDA) dont la partie de corrélation est celle proposée par *Perdew et Wang* [28].
E_x^{GGA} la correction de gradient pour l'échange à la LSDA proposée par *Becke* en **1988**.
E_x^{B88} [26],
E_c^{GGA} est exprimée en E_c^{LYP} équation (3.49).
E_c^{LSDA} est la correction de gradient pour la corrélation de *Perdew et Wang* de **1991** [29]
E_c^{VWN}.
Les paramètres a, b et c sont ajustés par la méthode des moindres carrés, leurs valeurs sont, respectivement de 0,20, 0,72 et 0,81 [30].

$$E_{xc}^{B3LYP} = (0.20)E_x^{HF} + (0.80)E_x^{LSDA} + (0.72)E_x^{B88} + (0.81)E_x^{LYP} + (0.19)E_c^{VWN} \tag{3.51}$$

3.6 Les fonctions pour la description des orbitales atomiques

N'importe quel ensemble complet de fonctions peut servir à la formation de l'ensemble de fonctions de base, dans la pratique, le choix de la fonction repose sur deux critères : le taux de convergence des orbitales paramétriseés et la facilité d'évaluation des intégrales moléculaires impliquant ces fonctions. Chacune des fonctions proposées dans la littérature aura des caractéristiques particulières qui deviendront des avantages, ou des désavantages, selon l'application. Nous décrirons les deux type de fonctions les plus couramment employées soient les fonctions de type exponentiel et les fonctions de type gaussien.

3.6.1 Les fonctions de type exponentiel

Les fonctions de type exponentiel sont le plus souvent employées pour les calculs atomiques et des systèmes diatomiques et polyatomiques linéaires. Leur forme, déterminée par le nombre quantique principal n, angulaire *l* et de spin électronique m, est

$$\phi_{n,l,m,\zeta}(r,\theta,\phi) = N_n Y_l^m(\theta,\phi) r^{n-1} e^{-\zeta r} \tag{3.52}$$

Avec n, l et m des nombres quantiques associés à l'orbitale atomique, N_n est le facteur de normalisation, ζ est l'exposant orbitalaire ou l'exposant de Slater déterminant la taille de l'orbitale et appelée aussi la constante de la charge effective du noyau, et Y_{lm} sont les harmoniques sphériques décrivant la partie angulaire de la fonction,
les combinaisons formées à partir de ces fonctions convergent rapidement, c'est-à-dire qu'un petit nombre de fonctions est requis pour décrire convenablement les orbitales atomiques, elles donnent une représentation juste de la fonction d'onde dans la région près du noyau et dans celle la plus éloignée.
Cependant le calcul d'intégrales moléculaires sur la base STO est difficile

3.6.2 Les fonctions de type gaussien

Les fonctions de type gaussien, proposées par *Boys* [34], sont présentement les fonctions les plus couramment employées pour les calculs électroniques moléculaires.

Les gaussiennes cartésiennes ont la forme,

$$\phi_{n,l,m}(x,y,z;\zeta) = (\frac{\pi}{2\zeta})^{3/2} \frac{(2n-1)!(2l-1)!(2m-1)!}{2^2(n+l+m)\zeta(n+l+m)} x^n y^l z^m exp(-\zeta r^2) \quad (3.53)$$

Où x, y et z sont les coordonnées de la position de la fonction.
La somme (n+l+m) définie le type de l'orbitale atomique.
n + l + m = 0 (OA de type s)
n + l + m = 1 (OA de type p)
n + l + m = 2 (OA de type d)
Les fonctions gaussiennes sont largement utilisées dans les calculs ab initio, leur avantage principal réside dans la facilité avec laquelle les intégrales moléculaires à plusieurs centres sont évaluées. Cette caractéristique découle du fait que le produit entre deux gaussiennes, localisées sur des centres différents (A et B), est équivalent à une gaussienne centrée qui lorsque normalisée, est située sur la droite AB en un point C. Cette propriété mathématique permet de faciliter considérablement le calcul d'intégrales moléculaires multicentriques. est aussi une fonction gaussienne.

3.6.2.1 Les fonctions de type gaussien contractées

Actuellement, la grande majorité des calculs atomiques et moléculaires se fait à l'aide de contractions formées de fonctions de type gaussien. En général, les fonctions représentant les orbitales de cœur peuvent être contractées sans trop de conséquences alors qu'un certain nombre de fonctions représentant les orbitales de valence doivent rester libres afin de laisser une plus grande flexibilité à la fonction d'onde.
On appelle une fonction gaussienne contractée (CGTO) une combinaison linéaire de gaussiennes primitives (PGTOs) :

$$\phi^{CGTO} = N \sum_{\mu}^{k} d_\mu \phi_\mu^{PGTO} \quad (3.54)$$

Où ϕ^{PGTO} est l'ensemble de fonctions gaussienne primitives, de même symétrie et centrées à la même position, d_μ est l'ensemble des coefficients de contraction et N est la constante de normalisation.
La forme des contractions formées de fonctions de type gaussien de symétrie s, p et d, où l'orbitale atomique de type s est représentée par la contraction $s_i(r)$ formée de gaussiennes primitives g_{s_i}, est

$$s_i(r) = \sum_{k=1}^{k_{si}} d_{si}, kg_{si}, K(\alpha_{si}, K, r) \quad (3.55)$$

$$p_i(r) = \sum_{k=1}^{k_{pi}} d_{pi}, kg_{pi}, K(\alpha_{pi}, K, r) \tag{3.56}$$

$$d_i(r) = \sum_{k=1}^{k_{di}} d_{di}, kg_{di}, K(\alpha_{di}, K, r) \tag{3.57}$$

Où k_{gl_i}. est le nombre de primitives de symétrie *l* (le nombre quantique angulaire) et $d_{l_i,k}$ et $\alpha_{l_i,k}$, sont, respectivement, les coefficients de contraction (où d'expansion) et les exposants de la fonction de base pour l'orbitale de symétrie *l*.
En pratique les orbitales atomiques OA de *Slater* (STO) sont approchées par une combinaison de plusieurs OA gaussiennes (GTO).
Dans certains ensembles de base, tels les STO-nG et d'autres ensembles optimisés selon la philosophie développée par *Pople* [35, 36], les primitives des couches électroniques avec le même nombre quantique principal, auront les mêmes exposants. Par exemple, les exposants des primitives représentant les orbitales de type 2s et 2p seront les mêmes et formeront la couche-sp. Cette approximation donne de bons résultats et permet une évaluation plus efficace des intégrales moléculaires. Les premières contractions de fonctions gaussiennes ont été obtenues en faisant un lissage des moindres carrés des orbitales de *Slater*. Cette approche donna naissance aux représentations Simple Zêta (SZ), Double Zêta (DZ), Triple Zêta (TZ).
• 1 contraction $\longrightarrow$ base Single Zeta (SZ) ;
• 2 contractions $\longrightarrow$ base Double Zeta (DZ) ;
• 3 contractions $\longrightarrow$ base Triple Zeta (TZ).
La base de ce groupe la plus simple est la base STO-3G encore appelée base minimale. Ceci signifie que les orbitales de type Slater sont représentées par trois fonctions gaussiennes, pour approcher chacune des orbitales de type Slater.

3.6.3 Les ensembles de base du type ($n-ijG$) *et* ($n-ijkG$)

Les ensembles de base de *Pople*. [37] font certainement parti des ensembles de base les plus utilisés jusqu'à aujourd'hui. La notation adoptée pour ces ensembles est du type n-ijG et n-ijkG où n est le nombre de primitives pour les orbitales de coœr et les ensembles (i,j) et (i, j, k) servent à représenter le nombre de primitives par contraction pour les orbitales de valence. Les ensembles n-ijG et n-ijkG sont de type DZV et TZV (Split Valence-Double Zeta et Split Valence-Triple Zeta), respectivement, pour la plupart, ces ensembles ont été construits en utilisant le concept d'exposants

partagés pour les électrons de la même couche, et les coefficients des contractions seront différents.
Où
• n est le nombre de primitives pour les orbitales internes ;
• i, j, k sont les nombres de primitives pour les orbitales de valence.

3.6.4 Les fonctions de polarisation

L'amélioration consiste à inclure des orbitales virtuelles dans le calcul : des orbitales p pour l'atome d'hydrogène, des orbitales d pour les atomes usuels, des orbitales f pour les métaux de transition. Elles permettent des déformations dans un champ de ligand qui a une symétrie réduite par rapport à l'atome libre. Le rôle des orbitales polarisées est de permettre, lorsqu'un atome d'hydrogène est impliqué dans une liaison chimique, la symétrie sphérique est perdue. La présence d'une orbitale 2p confère au système la symétrie attendue.

3.6.5 Les fonctions diffuses

Ces fonctions sont principalement caractérisées par la valeur très petite que prenne les exposants des primitives. Ce sont souvent des fonctions de type s ou de type p. Ces fonctions sont nécessaires pour décrire correctement la partie externe de la densité (anions, liaisons faibles de type *van der Waals*) et pour le calcul de certaines propriétés telles le moment dipolaire et la polarisabilité électronique. Ou les anions sont volumineux à cause de la grande répulsion entre les électrons (Il y a plus de particules négatives que de particules positives ; le noyau est très écranté). Il faut donc des bases avec des orbitales particulièrement diffuses pour que les orbitales les plus hautes occupées restent liantes.
Dans la notation de *Pople*, ces fonctions sont dénotées par le symbole "+" ce qui conduit à la notation n-ij+G ou n-ijk+G, des ensembles avec des fonctions diffuses de type s et de type p, respectivement, pour les atomes lourds. Les ensembles n-ij++G ou n-ijk++G auront en plus une fonction diffuse de type s sur l'atome d'hydrogène. la base 6-311+G est une base proche de la base 6-31G dans laquelle on a ajouté des fonctions polarisées sur les atomes lourds (première étoile) et sur les hydrogènes (deuxième étoile), elle très utilisée,

Exemple : la base 6-311G et 6-31+G* pour l'atome d'oxygène :

	La base 6-311G		La base 6-31+G*	
Orbitale	Contractions	Primitives	Contractions	Primitives
1s	1	6 x 1 = 6	1	6 x 1 = 6
2s	3	(3 + 1 + 1) x 1 = 5	2	(3 + 1) x 1 = 4
2p	3	(3 + 1 + 1) x 3 = 15	2	(3 + 1) x 3 = 12
s diffuse	/	/	1	1x1=1
p diffuse	/	/	1	1x3=3
d polarisante	/	/	1	1x5=5

3.7 Le modèle de solvatation

3.7.1 Principe et description de la cavité

De nombreuses expériences étant réalisées en phase liquide, et donc, il est nécessaire de tenir compte des effets de solvant dans les calculs visant à reproduire ou à interpréter des résultats expérimentaux obtenus en phase solvatée. Il existe deux façons de modéliser un solvant : (*) La première est de traiter un grand nombre de molécules du solvant explicitement (par les méthodes quantiques ou dynamique moléculaire classique), l'intérêt de ce modèle est qu'il rend compte d'éventuelles interactions solvant-soluté, l'inconvénient est qu'il est nécessaire d'avoir un nombre important de molécules pour obtenir une description réaliste du système, ce qui limite le choix de la méthode utilisable. Pour cette raison un nouveau type d'approche dit hybride a été développé : elle mélange les approches quantique et classique. Le cœur du système est traité au niveau quantique, alors que la partie externe est traitée par l'intermédiaire d'un champ de force, ce qui permet de modéliser le solvant avec un nombre réaliste de molécules. (**) La seconde est d'utiliser un modèle de continuum [38, 39] : à la place de molécules discrètes, on a un milieu continu infini et sans structure qui est polarisé par le soluté placé dans une cavité de forme appropriée. L'idée de modéliser les interactions électrostatiques dues au solvant, date des travaux de *Kirkwood* [40] et *Onsager* concernant les effets de solvatation sur les molécules polaires [41–43]. Évidemment les modèles de continuum ne sont pas capables de reproduire les interactions chimiques entre le soluté et le solvant comme les liaisons hydrogènes. Pour tenir compte de ces interactions spécifiques, il faut au moins inclure explicitement les molécules du solvant de la première sphère de coordination. Les deux approches sont donc complémentaires. Les résultats des études en phase aqueuse présentés dans la deuxième partie ont été obtenus en utilise un continuum modélisant le solvant. Par conséquent la suite de cette section portera exclusivement sur l'approche du continuum.

3.7.2 Modèles de Born, d'Onsager et de Kirkwood

Le modèle de solvant le plus ancien est celui de *Born* (**1912**). Dans ce modèle, les interactions soluté-solvant dépendent uniquement de la charge du soluté et de la constante diélectrique du milieu diélectrique. *Onsager* (**1936**) [42] a affiné le modèle de *Born* en utilisant le moment dipolaire du soluté (et non pas sa charge) lors du calcul

de l'énergie de solvatation. Dans ces deux modèles, le soluté dans son ensemble est placé dans une cavité sphérique (Figure 3.1).
Le modèle de *Kirkwood* [40] est une extension du modèle *d'Onsager*. L'interaction du champ de réaction avec le soluté est exprimée sous forme d'un développement multipolaire et est ajoutée à l'hamiltonien mono-électronique. La différence avec le modèle *d'Onsager* est qu'il est ici possible d'y inclure un potentiel répulsif représentant l'interaction d'échange entre le soluté et le solvant, ce qui permet de confiner la densité électronique à l'intérieur de la cavité.

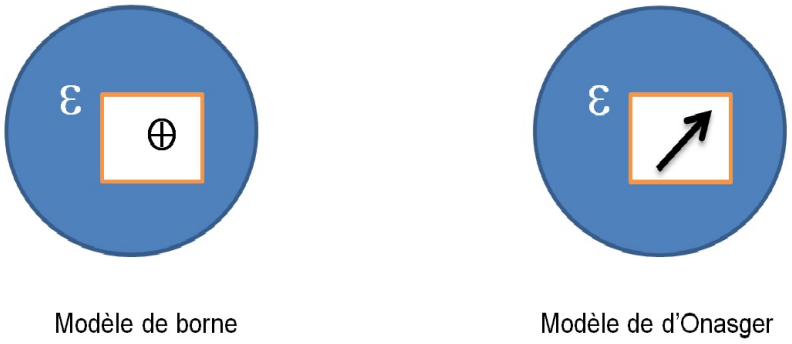


FIGURE *3.1: Modèle de born et d'Onasger*

3.7.3 Modèle SRCF

Dans un premier temps, nous allons décrire les différents termes énergétiques et le bilan énergétique du modèle de continuum tenant compte de différents types d'interactions. La première étape implique la création d'une cavité dans le continuum de solvant (Figure 3.2). En pratique, la cavité est construite à partir d'un ensemble de sphères centrées sur les noyaux et ayant un rayon de type *van der Waals*. La formation de cette cavité coûte une certaine quantité d'énergie (positive) : ΔG_{cav}, l'énergie libre de cavitation. Cette quantité dépend de la nature du solvant ainsi que de la topologie de la cavité. En second temps, le soluté est placé dans la cavité et celui-ci interagit avec le continuum. On distingue trois types d'interactions soluté-continuum : électrostatiques ΔG_{ele}, répulsives et dispersives. Les deux dernières sont calculées grâce à des relations empiriques.

Le processus d'interaction SCRF (*Self-Consistent Reaction Field*) d'une molécule dans sa cavité est autocohérent résolu : la distribution de charge de ce soluté pola-

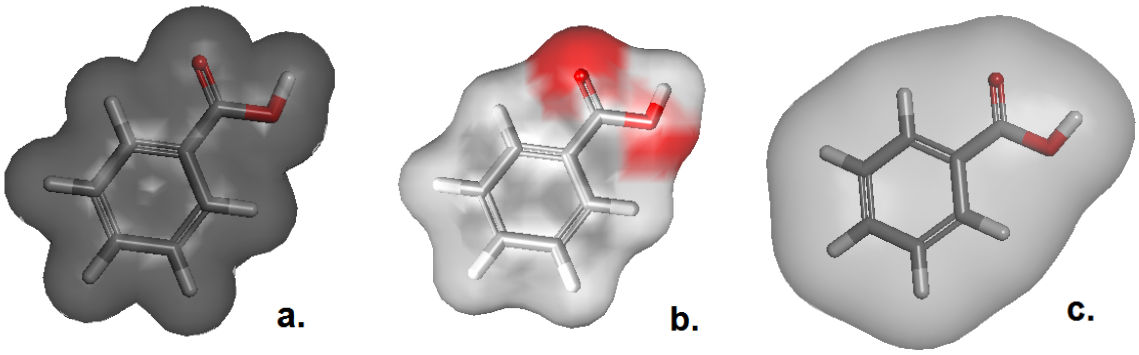


FIGURE *3.2: Modèle de la surface moléculaire (a) la surface de Van der Waals (b) la surface accessible au solvant, et (c) Surface excluant le solvant*

rise la surface de la cavité du continuum, des charges apparaissent donc à l'interface soluté-continuum, ces dernières induisent un potentiel de réaction qui, à son tour, modifie la densité électronique du soluté, ensuite, le continuum doit s'adapter et il y a une nouvelle modification de la distribution de charges à l'interface soluté-continuum, et ainsi de suite jusqu'à l'obtention d'une convergence électrostatique entre la distribution de charges propre au soluté et celle de la surface de la cavité, ce terme énergétique, toujours négatif, est la contribution électrostatique (ΔG_{ele}).
Les termes de répulsion et de dispersion sont associés. le terme de dispersion/répulsion (ΔG_{dis} et ΔG_{rep}) implique l'interaction du soluté avec le solvant à l'interface de la cavité, et donne une contribution négative/positive à la variation d'énergie.
Au final, l'énergie totale d'interactions s'exprime en un terme électrostatique et en trois termes non-électrostatiques :

$$\Delta G_{sol} = \Delta G_{cav} + \Delta G_{elec} + \Delta G_{dis} + \Delta G_{rep} \tag{3.58}$$

Pour résoudre le problème les modèles de continuum utilisent l'équation de $Poisson$ de l'électrostatique classique :

$$-\nabla|\varepsilon_r(r)\nabla V(r) = 4\pi\rho_m(r) \tag{3.59}$$

ε_r est une fonction diélectrique dans le milieu et le potentiel électrostatique total V(r) est la somme du potentiel électrostatique Vp(r) généré par la distribution de charges du soluté p et le potentiel de réaction $V\rho(r)$ créé par la polarisation du milieu diélectrique :

$$V(r) = V_\rho(r) + V_\sigma(r) \tag{3.60}$$

La constante diélectrique ($\varepsilon_r(r) = \varepsilon_s/\varepsilon_0$) peut prendre deux valeurs :
$\varepsilon_r(r) = 1, sir \in V_{int}$
$\varepsilon_r(r) = \varepsilon, sir \in V_{ext}$ Où V_{int}, V_{ext} est le volume à l'intérieur et à l'extérieur de la cavité, pour les deux régions, l'équation (3.59) devient :

$$-\nabla^2 V(r) = 4\pi\rho_m(r), r \in V_{int} \tag{3.61}$$

$$-\varepsilon\nabla^2 V(r) = 0, r \in V_{ext} \tag{3.62}$$

Les équations (3.61, 3.62) doivent être accompagnées de conditions frontières, à l'infini mais aussi et surtout à la surface de la cavité, $|v_{int} - V_{ext}| = 0$ afin de garantir la continuité du potentiel.
La distribution de charge surfacique $\sigma(r_s)$ peut être exprimée ainsi en termes de quantités aisément calculables avec des processus actuels :

$$\sigma(r_s) = \frac{\varepsilon - 1}{4\pi\varepsilon} E(r_s) \tag{3.63}$$

Où $E(r) = \left(\frac{\partial V(r)}{\partial n}\right)_k$ est le champ électrique perpendiculaire à la cavité.

3.7.4 Modèle PCM

Le modèle précédent repose sur l'hypothèse que toute la charge électronique du soluté est comprise dans la cavité, le modèle PCM [44–46] (Polarizable Continuum Model)[i] permet de tenir compte de la fraction de la charge électronique du soluté se trouvant à l'extérieur de la cavité. Celle-ci est constituée de sphères qui se pénètrent partiellement et qui sont centrées sur des atomes ou groupes d'atomes du soluté. Cette procédure garantit que la cavité suit la forme réelle du complexe. Chaque sphère est calculée en fonction du rayon de *van der Waals* de chaque atome et la densité de charge est répartie sur la surface de la cavité. Les triangles à l'intersection de deux ou plusieurs sphères sont modifiées avec un algorithme qui conserve les caractéristiques de la surface de la cavité et la distribution de charge $\sigma(s)$ [44]. Les rayons des sphères sont des paramètres importants car les énergies calculées dépendent de la taille de la cavité. (figure 3.3) Dans le cas de solvants très polaires, il est possible de faire une approximation supplémentaire en considérant que le solvant est conducteur (la

i. Bien que le modèle de Kirkwood soit aussi un modèle de continuum polarisable, il est d'usage de ne pas l'inclure dans les méthodes PCM.)

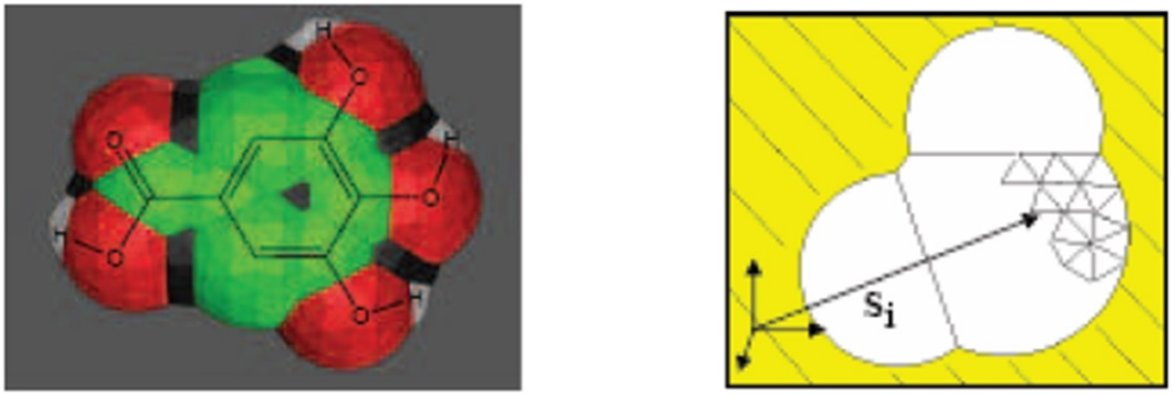


FIGURE *3.3: découpage de la surface d'une cavité en un ensemble de tesserae*

permittivité relative tend vers l'infini). En principe moins précises, les méthodes dites C-PCM (Conductor- Polarizable Continuum Model) permettent de simplifier le calcul des densités de charge à la surface de la cavité. Cette approximation est valable tant que la constante diélectrique est supérieure à 5.
Dans toutes les approches PCM, les charges de solvatation sont obtenues par la résolution d'un système linéaire de N équations couplées (N étant le nombre de tessères) :

$$Dq = -b \tag{3.64}$$

D est une matrice qui dépend de la constante diélectrique du solvant et des paramètres géométriques de la cavité,
q est un vecteur colonne constitué des charges de solvatation de chaque tessère,
b est un vecteur colonne constitué des valeurs du champ électrique ou du potentiel électrostatique du soluté pour respectivement le modèle PCM ou le modèle C-PCM.

3.7.4.1 Algorithme PCM

Dans cette méthode, la fonction d'onde électronique du système est déterminée à partir d'un hamiltonien électronique modifie par addition d'un terme électrostatique dû à la présence de charges surfaciques $\rho_s(r_s)$ à l'interface entre le soluté et le solvant.
L'hamiltonien de solvatation s'écrit alors :

$$H = H_o + V_\sigma \tag{3.65}$$

avec

$$V_\sigma(r_s) = \int \frac{\sigma_s(r_s)}{|r - r_s|}$$

Où H_0 est l'hamiltonien électronique dans le vide et V_σ est le potentiel électrostatique dû à la solvatation (c'est une perturbation), l'hamiltonien dépend donc de ses fonctions propres, d'où une résolution par itération.
Le but du processus itératif est l'évaluation de $\sigma(r_k)$ de chaque tesserae, avec une surface ΔS_k et une charge q_k au point r_k suivant la relation :

$$q_k = \sigma(r_k)\Delta S_k \tag{3.66}$$

Ce processus itératif se fait selon les étapes suivantes :

Étape 1 : On part d'une valeur d"essai de $\sigma(r_k)$ correspondant à un potentiel dû à la distribution de charges du soluté uniquement $V_\sigma(r) = 0$
D'ou

$$V(r) = V_\rho(r) \tag{3.67}$$

On appelle σ_k^{00} les charges surfaciques correspondant à cette approximation. Le premier indice 0 correspond aux ponctuelles de soluté seul. Le second indice 0 correspond au départ, donc au fait que l'on suppose : $V_\sigma(r) = 0$
On a alors :

$$\sigma_k^{00} = -\left[\frac{\varepsilon - 1}{4\pi}\right]\left(\frac{\partial V(r)}{\partial n}\right)_k \tag{3.68}$$

On obtient alors :

$$q_k^{00} = \sigma_k^{00}(S)\Delta S_k \tag{3.69}$$

Ces charges produisent au centre des éléments de surface une contribution supplémentaire au potentiel électrostatique et au champ électrique d'où :

$$V^{00}(r) = V_\rho^{00}(r) + V_\sigma^{00}(r) \tag{3.70}$$

On calcul alors une nouvelle distribution de charge surfacique σ_k^{01}

$$\sigma_k^{01} = -\left[\frac{\varepsilon - 1}{4\pi}\right]\left(\frac{\partial V(r_s)}{\partial n}\right)_k \tag{3.71}$$

Alors :

$$q_k^{01} = \sigma_k^{01}(S)\Delta S_k \tag{3.72}$$

L'itération est effectuée jusqu'à atteindre la convergence pour laquelle $V_\sigma^{01}(r)$ est obtenu à partir des q_k^{0r}
On ajoute alors le potentiel $V_\sigma(r)$ à l'hamiltonien du soluté Ho ;$H = Ho + H_\sigma$
On résout alors les équations IIF ou Kohn Sham (dans le cas de la DFT) avec cet hamiltonien.

Étape 2 : On obtient alors une nouvelle distribution de charge pour le soluté à partir de laquelle on déduit un nouveau jeu de départ de charges surfaciques q_k^{10} et on itère jusqu'à obtenir q_k^{if} et donc $V_\sigma^{if}(r)$. On répète la même procédure jusqu'à atteindre la convergence globale et donc l'obtention de ψ^f
On peut alors définir une énergie libre électrostatique G_{le} comme :

$$H\Psi = \left[H^0 + V^R\right]\Psi = E\Psi \tag{3.73}$$

En résolvant cette équation on obtient :

$$G_{el} = E - \frac{1}{2}\langle\Psi'|V^R|\Psi'\rangle = \langle\Psi'|H^0|\Psi'\rangle + \langle\Psi'|V^R|\Psi'\rangle - \frac{1}{2}\langle\Psi'|V^R|\Psi'\rangle \tag{3.74}$$

$$= \langle\Psi'|H^0|\Psi'\rangle + \frac{1}{2}\langle\Psi'|V^R|\Psi'\rangle \tag{3.75}$$

L'énergie libre de solvatation s'écrit alors

$$\Delta G_{el} = G_{el} - \Psi^0|H^0|\Psi^0\rangle$$

Soit :

$$\Delta G_{el} = \langle\Psi'|H^0|\Psi'\rangle + \frac{1}{2}\langle\Psi'|V'|\Psi'\rangle - \langle\Psi^0|H^0|\Psi^0\rangle \tag{3.76}$$

Avec

$$\langle\Psi'|V^R|\Psi'\rangle = \int V(R)\rho(r)dr \tag{3.77}$$

3.7.5 Modèle SMD

Le modèle de solvatation SMD [47] est un modèle continuum de solvatation basé sur des calculs théoriques de la densité de charge d'une molécule de soluté en interaction avec le modèle continuum du solvant.
Le modèle est appelé SMD, où "D" indiqué que la densité électroniques totale de soluté est utilisé sans définir les charges atomiques partielles. "Continuum" signifie que le solvant n'est pas traité de manière explicite, et on considère le solvant comme un

milieu continu de constant diélectrique élevé, SMD est un modèle de solvatation universel, où «universel» désigne son application à tous les types de solutés chargés ou non chargés dans n'importe quel solvant ou milieu.
Ce modèle propose la décomposition de l'énergie libre de solvatation en deux termes. La première composante est la contribution électrostatique résulte du processus SCRF champ de réaction auto-cohérent (Self-Consistent Reaction Field) et calculée avec l'équation de *Poisson-Boltzinann* (PB), qu'il est permet de travailler avec des cavités de forme plus réaliste, Les cavités pour le calcul électrostatique sont définis par des superpositions de sphères centrées.
La deuxième composante est appelée le terme cavité de dispersion solvant-structure et sa contribution résultant a des interactions à courte portée entre le soluté et les molécules de solvant dans la première couche de solvatation. Cette contribution est une somme de termes qui sont proportionnels au surface accessible au solvant (solvent accessible surface area), et les atomes de soluté.
Le modèle SMD a été paramétré avec une série d'apprentissage de 2821 données d'énergies libres de solvatation.

3.7.5.1 Description de Modèle SMD

L'énergie libre de solvatation a l'état standard, est l'énergie libre correspondant au transfert du système en phase gazeuse vers une phase liquide, qui s'exprimé en ;

$$\Delta G_s = \Delta G_{ENP} + \Delta G_{CDS} + \Delta G_{Con} \tag{3.78}$$

Avec
La composante ENP dans l'équation (3.78) représente les termes de l'énergie libre (l'électronique (E), nucléaire (N), et La polarisation du continuum (P)). La composante nucléaire est la différence entre l'énergie totale calculée en phase gazeuse pour la géométrie optimisée en phase gazeuse et l'énergie totale calculée en phase gazeuse pour la géométrie optimisée en phase aqueuse.
La deuxième composante CDS correspondant aux termes de l'énergie libre de cavitation (C), les variations de l'énergie de dispersion (D), et les changements éventuels sur la structure du solvant (S).
Le dernier terme de l'équation (3.78) représente le terme de correction de ΔG l'énergie libre de Gibbs entre l'état standard en phase gazeuse (1 atm) et l'état standard en phase liquide (1 M).

3.7.6 Les termes non-électrostatiques

Les termes non électrostatiques sont de trois types : cavitation, dispersion et répulsion [48].

3.7.6.1 Le terme de cavitation :

Le transfert d'une molécule d'une phase gazeuse à une phase liquide se produit en deux étapes. La première est la formation de la cavité dans la solution et la seconde est l'introduction de la molécule de soluté à l'intérieur de la cavité. Ces deux étapes résultent en un excès d'énergie libre. Le terme de cavitation ΔG_{cav} correspond à la première étape. Selon $Uhlig$ [49], le travail pour former une cavité macroscopique est fonction de la tension de surface γ du liquide et de la surface de la cavité S_m :

$$\Delta G_{cav} = \gamma S_m \tag{3.79}$$

3.7.6.2 Le terme de dispersion

Les interactions de $Van\ der\ Waals$ sont la combinaison de deux termes : **un terme répulsif** traduisant le recouvrement des nuages électroniques de même polarité à courtes distances et un **terme attractif** dû aux forces dispersives qui apparaissent instantanément durant les fluctuations des nuages électroniques. Les forces de dispersion [50] sont des forces faibles intermoléculaires créées par des dipôles instantanés : le dipôle instantané d'une molécule A induit un dipôle instantané sur une molécule B et interagit avec lui. Ces forces représentent en général la plus importante composante des forces de Van der Waals (entre 0.5 et 40 kcal/mol). Elles apparaissent lorsque la densité électronique d'une molécule n'est pas équitablement répartie autour de celle-ci, ce qui crée ainsi un léger moment dipolaire. Ces moments dipolaires instantanés varient très rapidement au cours du temps. Ainsi 'a chaque distribution inhomogène se crée un moment dipolaire instantané qui peut interagir avec les moments dipolaires qu'il induit sur les molécules voisines. . Ce qui permet d'exprimer l'énergie ou potentiel de dispersion par relation :

$$E_{disp}^{vdm}(r) = -C^{st}\left(\frac{\sigma}{r}\right)^6 \tag{3.80}$$

3.7.6.3 Le terme de répulsion

Les forces répulsives sont les plus locales et les plus intenses. Le potentiel ou l'énergie de répulsion est une fonction rapidement croissante lorsque la distance r séparant les deux atomes, diminue

$$E_{disp}^{vdm}(r) = \left(\frac{\sigma}{r}\right)^n$$

Avec une valeur pouvant n allant de 9 à 12. Dans le PCM, le terme de répulsion est évalué par l'approche *d'Amovilli et Mennucci* [51].

$$G_{rep} = \rho_s \int dr U_{ms}^{rep}(r) g_{ms}(r) \tag{3.81}$$

Avec m le soluté, s le solvant, r est un ensemble de coordonnées qui définissent la géométrie du complexe ms, ρ_s est la densité du solvant et g_{ms} est une fonction de corrélation qui vaut 0 à l'intérieur de la cavité et 1 à l'extérieur. En pratique, on estime le potentiel de répulsion (U_{rep}) suivant une approche de type *Lennard-Jones*.
La plus connue des fonctions de potentiel de type *Van der Waals* est la fonction de *Lennard-Jones*, dans laquelle les énergies de dispersion et de répulsion sont définie par une seule et même expression qui s'écrit pour les deux molécules, neutres et non polaires, i et j sous la forme

$$V^{VDM-LJ}(r) = E_{rep} + E_{dis} = \sum_i \sum_j 4\varepsilon_{ij} \left[\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6} \right] \tag{3.82}$$

Cette équation contient deux paramètres ajustables : le diamètre de collision (σ_{ij}) qui est la distance minimale d'approche entre les deux atomes i et j pour laquelle l'énergie entre deux atomes est nulle et la profondeur du puits (ε_{ij}) qui représente le minimum de l'énergie potentielle, ce qui correspond à l'interaction la plus stable.
Enfin, notons que bien que le potentiel de *Lennard-Jones* soit le plus utiliser pour décrire les interactions de *Van der Waals*, vu sa simplicité et le peu de paramètres qu'il fait intervenir.

Bibliographie

[1] McWeeny R., B. T. Sutcliffe, Methods of Molecular Quantum Mechanics, (**1969**), Academic Press, London and New York.

[2] Atkins P. W., Molecular Quantum Mechanics, (**1983**), Oxford University Press, Oxford.

[3] Szabo N., S Osllund, Modern Quantum Chemistry, (**1982**), Macmillan, New York.

[4] Rivail J. L., Eléments de chimie Ouantique a l'Usage des Chimistes, (**1994**), InterEditions, Paris

[5] Jensen F., Introduction to computational Chemistry, (**1999**), John Wiley, Sons Ltd. Chichester.

[6] Schrödinger E., Quantization as an Eigenvalue Problem, Ann. Phys. Leipzig., 79, (**1926**), 361.

[7] Born M. R., Oppenheimer, Zur Ouantentheorie der Molekeln, Ann. Phys., (Leipzig), 84, (**1927**), 457.

[8] Pauli W., Ûber den Zusammenhang des Abschlusses der Elektronengruppen im Atom mit der Komplexstruktur der Spektren, Z. Phys., 31, (**1925**), 765.

[9] Slater J. C., The Theory of Complex Spectra, Phys. Rev., 34, (**1929**), 1293.

[10] MinkineV., B. Simkine, R. Minaev, Théorie de la structure moléculaire, (**1982**), Edition Mir, Moscou.

[11] Fock V., Z. Näherungsmethode zur Losung des quanten-mechanischen Mehrkörperprobleme, Physik., 61, (**1930**), 126.

[12] Hartree D. R., The wave mechanics of an atom with a non-coulomb central field. I. Theory and methods, Proc, Cambridge Philos. Soc., 24, (**1928**), 89.

[13] Roothaan C. C. J., New Developments in Molecular Orbital Theory, Rev. Mod. Phys., 23, (**1951**), 69.

[14] Berthier G., j. Chem. Phys., 51, (**1954**), 363, J. A. Pople., R. K. Nesbet., Self Consistent Orbitals for Radicals, J Chem Phys., 22, (**1954**), 571.

[15] Shavitt I., Methods of Electronic Structure Theory, H. F. Shaefer, Ed., (**1977**), Plenum Press, New York.

[16] Jugl A,. Chimie Quantique Structurale et Eléments de Spectroscopie Théorique, (**1978**), Alger : O.P.U.

[17] Parr R. G., R. A. Donnelly, M. Levy, W. E. Palke, Electronegativity- the density functional viewpoint. J. Chem. Phys., 68, (**1978**), 3801.

[18] Hohenberg P., Kohn, W. Inhomogeneous Electron Gas, Phys. Rev., 136, (**1964**), 864.

[19] Parr R. G., W. Yang, Density Functional Theory of Atoms and Molecules, (**1989**), Oxford university press New-York.

[20] Kohn W., Sham, L. J. Self-Consistent Equations Including Exchange and Correlation Effects, Phys. Rev. A., 140, (**1965**), 1133.

[21] Ziegler T., Approximate density functional theory as a practical tool in molecular energetics and dynamics, Chem. Rev., 91, (**1991**), 651.

[22] Becke A. D., Correlation energy of an inhomogeneous electron gas : A coordinate-space model, J. Chem. Phys., 88, (**1988**), 1053

[23] Dirac P. A. M., Note on exchange phenomena in the Thomas-Fermi atom, Proc. Camb. Phil. Soc., 26, (**1930**), 376

[24] Slater J. C., A simplification of the Hartree-Fock method, Phys. Rev., (**1951**), 385.

[25] Vosko S. H., L. Wilk, M. Nusair, Accurate spin-dependent electron liquid correlation energies for local spin density calculations : A critical analysis, Can. J. Phys., 58, (**1980**), 1200.

[26] Becke A. D., Density-Functional Exchange-Energy Approximation with Correct Asymptotic Behavior, Phys. Rev. A., 38, (**1988**), 3098.

[27] Perdew J. P., Density-functional approximation for the correlation energy of the inhomogeneous electron gas, Phys. Rev. B., 33, (**1986**), 8822.

[28] Perdew J. P., Y. Wang, in Electronic Structure of Solids '91, ed. P. Ziesche, H. Eschrig, (**1991**), Academie Verlag, Berlin, p. 11

[29] Lee C., W. Yang, R. G. Parr, Development of the Colle-Salvetti correlation-energy formula into a functional of the electron density, Phys. Rev. B., 37, (**1988**), 785.

[30] Becke A. D. Density-Functional Thermochemistry. III. The Role of Exact Exchange, J. Chem. Phys., 98, (**1993**), 5648.

[31] Miehlich B., A. Savin, H. Stoll, H. Preuss, Results obtained with the correlation-energy density functionals of Becke and Lee, Yang and Parr, Chem. Phys. Lett., 157, (**1989**), 200.

[32] Adamo C., V. Barone, Implementation and validation of the Lacks-Gordon exchange functional in conventional density functional and adiabatic connection methods, J. Comp. Chem., 19, (**1998**), 418.

[33] Gill P. M. W., A new gradient-corrected exchange functional, Mol. Phys., 89, (**1996**), 433.

[34] Boys S. F., Electronic wave functions. I. A general method of calculation for stationary states of any molecular system. Proc. R. Soc. London Ser. A., 200, (**1950**), 542.

[35] Ditchfield R., W. J. Hehre, J. A. Pople, Self-Consistent Molecular Orbital Methods. 9. Extended Gaussian-type basis for molecular-orbital studies of organic molecules, J. Chem. Phys., (**1971**), 54, 724.

[36] Hehre W. J., R. Ditchfield, J. A. Pople, Self-Consistent Molecular Orbital Methods. 12. Further extensions of Gaussian-type basis sets for use in molecular-orbital studies of organic-molecules, J. Chem. Phys., 56, (**1972**), 2257.

[37] **(a)** McLean A. D., G. S. Chandler, Contracted Gaussian-basis sets for molecular calculations. 1. 2nd row atoms, Z=11-18, J. Chem. Phys. 72, (**1980**), 5639 , **(b)** Krishnan R., J. S. Binkley, R. Seeger, J. A. Pople, A basis set for correlated wave functions. J. Chem. Phys. 72, (**1980**), 650, **(c)** Wachters A. J. H., Gaussian Basis Set for Molecular Wavefunctions Containing Third Row Atoms, J. Chem. Phys. 52, (1970), 1033, **(d)** P. J. Hay, Gaussian basis sets for molecular calculations. The representation of 3d orbitals in transition metal atoms, J. Chem. Phys., 66, (**1977**), 4377.

[38] Cramer C. J., D. G. Truhlar, Implicit Solvation Models : Equilibria, Structure, Spectra, and Dynamics, Chem. Rev., 99(8), (**1999**), 2161,

[39] Tomasi J., B. Mennucci, R. Cammi, Quantum Mechanical Continuum Solvation Models, Chem. Rev., 105, (**2005**), 2999

[40] Kirkwood J. G., Theory of solutions of molecules containing widely separated charges with special application to zwitterions, j. Chem. Phys., 2, (**1934**), 351.

[41] Baldridge K., A Klamt, GAMESS/COSMO : Eirst principle implementation of solvent effects without outlying charge error, j. Chem. Phys., 106, (**1997**), 6622.

[42] Onsager L., Electric Moments of Molecules in Liquids, j. Am. Chem. Soc., 58, (**1936**), 1486.

[43] Maurizio C., Mennucci B., Pitarch J. Tomasi J., Correction of cavity-induced errors in polarization charges of continuum solvation models, j. Comp. Chem., 19, (**1998**), 833.

[44] Miertus, S. ; Scrocco, E. ; Tomasi, J., Electrostatic Interaction of a Solute with a Continuum. A Direct Utilization of ab Initio Molecular Potentials for the Prevision of Solvent Effects, Chem. Phys., 55, (**1981**), 117.

[45] Tomasi j., M. Persico, Molecular Interactions in Solution : An Overview of Methods Based on Continuous Distributions of the Solvent, Chem, Rev., 94, (**1994**), 2027.

[46] Cammi R., j. Tomasi, Remarks on the use of the apparent surface charges (ASC) methods in solvation problems : Iterative versus matrix-inversion procedures and the renormalization of the apparent charges, j. Comp. Chem., 16, (**1995**), 1449.

[47] Marenich, A. V., C. J. Cramer, D. G. Truhlar, Universal solvation model based on solute electron density and on a continuum model of the solvent defined by the bulk dielectric constant and atomic surface tensions, J. Phys. Chem. B., 113, (**2009**), 6378.

[48] Laurant A. D., Thèse de doctorat, Université de Henri Poicare, (**2010**), Nancy

[49] **(a)** Uhlig H. H. The solubilities of gases and surface tension. J., Phys. Chem., 41, (**1937**), 1215. **(b)** I. Tunon, E. Silla, J. L. Pascual-Ahuir, Continuum Uniform Approach Calculations of the Solubility of Hydrocarbons in Water, Chem. Phys. Lett. 203, (**1993**), 289.

[50] London F., Zur Theorie und Systematik der Molekularkräfte, Z. Phys., 60, (**1930**), 245.

[51] Amovilli C., B. Mennucci, Self Consistent Field Calculation of Pauli Repulsion and Dispersion Contributions to the Solvation Free Energy in the Polarizable Continuum Model, J. Phys. Chem., 101, (**1997**), 1051.

APPLICATIONS

CHAPITRE 4

Modélisation QSPR des températures de fusion des acides gras

L'objectif de cette application est le développement et l'évaluation des modèles (QSPR) pour la prédiction des points de fusion d'une série constituée de 62 acides gras. La méthode Best Multi-Linear Regression (BMLR) implémentée dans le logiciel CODESSA est utilisée pour développer ces modèles afin de mettre en place le modèle le plus performant possible. Le meilleur modèle QSPR est obtenu avec cinq descripteurs moléculaires est caractérisé par les paramètres statistiques ($R^2 = 0,945, R^2_{adj} = 0,942$, F= 190,90).

4.1 Introduction

Dans cette application, des modèles QSPR sont développés pour prédire des températures de fusion des acides gras. La détermination de la température de fusion a fait l'objet de plusieurs travaux [1–4] car il est nécessaire de définir l'état physique des lipides (forme solide ou liquide) pour comprendre certains phénomènes importants présents dans les organismes vivants. Au laboratoire, le point de fusion d'une substance est mesuré à l'aide de la table chauffante de *Kofler*, appelée en pratique *banc Kofler*, il s'agit d'un appareil de mesure ou d'une plaque chauffante présentant un gradient de température permettant d'estimer la température de fusion d'une matière. Cet appareil permet une mesure rapide et précise mais ne convient que pour des composés ayant une température de fusion comprise entre 50 et 250 °C. Alors que dans cette application, nous avons élaboré des modèles QSPR pour prédire la température de fusion des acides gras ayant des valeurs au-dessous de 50 °C c'est-à-dire jusqu'à -65 °C. Nous avons utilisé la méthode dite Best Multi-Linear Regression (BMLR) implémentée dans le programme CODESSA [5] pour développer des modèles avec plusieurs descripteurs moléculaires [6].

4.2 Généralités sur les acides Gras

Les acides gras font partie de la grande famille des lipides [7]. I ls sont formés d'une chaîne hydrocarbonée avec un groupement carboxyle ($COOH$) à une extrémité et un groupement méthyle (CH_3) à l'autre extrémité [8]. Ils sont insaturés (AGI) ou saturés (AGS) selon qu'ils contiennent ou pas des doubles liaisons. Ils peuvent également lorsqu'ils possèdent des doubles liaisons présenter des configurations en cis ou trans. Á l'état naturel, la majorité des acides gras ont une configuration cis. Différentes nomenclatures existent. Les nutritionnistes utilisent la nomenclature qui indique la longueur de leur chaîne carbonée l'absence ou la présence de doubles liaisons qui reflète la réactivité biochimique, et l'appartenance à la famille (position de la première double liaison par rapport au groupement méthyle terminal). Parmi les AGI, on trouve les acides gras monoinsaturés (AGMI) qui ne présentent qu'une seule double liaison et les acides gras polyinsaturés (AGPI) qui présentent plusieurs doubles liaisons. Les acides gras font une partie intégrante de notre alimentation quotidienne. Selon les recommandations nutritionnelles récentes, les lipides alimentaires devraient représenter

35 % à 40 % de l'apport énergétique total quotidien, soit environ 20 % en masse de l'ensemble des macronutriments[i]. L'utilisation et l'assimilation des acides gras par l'organisme ne peuvent se produire qu'en présence et en synergie avec des nutriments tels que minéraux, oligoéléments, vitamines, enzymes... D'où découle la nécessité de remédier aux éventuelles carences constatées lors d'examens spécifiques (par prise de sang total), par l'alimentation quotidienne biologique ou par une supplémentation appropriée.

Des études ont démontrés qu'il y a une corrélation entre la quantité de gras saturé dans la diète et un taux anormalement élevé de cholestérol sanguin. Un taux élevé de cholestérol augmente les risques de maladies cardiovasculaires. En résumé, les acides gras saturés ont tendance à faire monter le cholestérol et le cholestérol favorise les maladies cardiovasculaires.

La température de fusion et la solidification des acides gras a une grande importance économique [9]. Pour un acide gras, la transformation de l'état liquide à l'état solide est une réaction exothermique(chaleur latente de cristallisation). Ce phénomène est la base pour la détermination de l'indice de matière grasse solide. La fusion d'un acide gras est une réaction instantanée, tandis que la cristallisation est généralement un processus lent [10–14]. *Svenstrup et al.* [15] ont montré qu'il y a une corrélation entre la température de fusion et la vitesse de refroidissement des acides gras, et qu'un refroidissement rapide conduit à la diminution du point de fusion qui produit des cristaux instables.

Les méthodes QSPR (Relations Quantitatives Structure-Propriété) proposent un protocole qui estime les valeurs expérimentales de point de fusion des acides gras basé sur des descripteurs dérivés de la structure moléculaire. L'avantage de cette approche consiste seulement la connaissance de la structure de la molécule et ne dépend pas des propriétés expérimentales. Une fois le modèle établi, il est applicable pour la prédiction de la propriété étudiée de nouveaux composés qui n'ont pas été encore synthétisés ou trouvés. Ainsi, les méthodes QSPR peuvent accélérer le processus de développement des nouvelles molécules ayant des propriétés souhaitées.

Dans cette étude, nous présentons une base de données des 62 valeurs expérimentales de points de fusion des acides gras (tableau 4.1), disponibles dans la littérature [17]. Nous décrivons également une nouvelle approche pour la sélection rapide des descripteurs dans l'analyse (QSPR). Et nous proposons un modèle qui corrèle T_f

i. Les macronutriments sont protéines, glucides et lipides.

températures de fusion des 62 acides gras AGs .

4.3 Méthodologie

4.3.1 Optimisation de la géométrie et calculs de chimie quantique

Les structures 2-D des 62 acides gras sont générées automatiquement à l'aide *ChemBioDraw Ultra* [18]. Les géomètres sont optimisées au niveau semi-empirique PM6 [19] implémenté dans le logiciel *MOPAC* [20]. Des calculs de fréquences vibrationnelles ont été réalisés au même niveau de théorie afin de s'assurer qu'aucune des structures optimisées ne présentent de fréquence imaginaire. Les structures optimisées ont été suivies par un autre calcul single point (single-point calculation), de fréquences et de NBO au niveau la Théorie de la Fonctionnelle de la Densité (DFT), à l'aide de la fonctionnelle hybride B3LYP [21, 22] et la base 6-31+G*, réalisé avec le logiciel *Gaussian 2003* [23]. Les structures obtenues ont été transférées au programme CODESSA [5] pour le calcul des descripteurs moléculaires.

4.3.2 Analyse QSPR

Depuis le premier modèle QSPR/QSAR de *Hansch* [24], différentes techniques permettent de réduire le nombre de descripteurs. L'une d'entre elles est la technique Best Multi-Linear Regression (BMLR) pour la mise en place de modèles multi-linéaires (Voir Section 2.3 (page 36)) utilisées par le programme CODESSA, la BMLR est une méthode « *stepwise* » méthode qui procèdent pas à pas par élimination successive. Cette méthode permet premièrement d'exclure tous les descripteurs présentant une variance non significative et deuxièmement d'éliminer les deux descripteurs qui sont fortement corrélés entre eux. Seul celui présentant la meilleure corrélation avec la propriété est conservé. En effet par des tests du type F (*Fisher*), elle ajoute ou élimine des variables dans l'équation. L'algorithme s'arrête quand on ne peut plus ajouter ou retrancher des variables sans dépasser les seuils statistiques choisis. Ces deux étapes permettent de s'assurer que de tels descripteurs ne sont pas inclus par chance dans le modèle final. Non seulement cette méthode évite l'introduction de descripteurs inappropriés dans le modèle mais elle rend l'analyse moins coûteuse en terme de temps de calcul, puisqu'elle réduit le nombre de variables restant à traiter. Par la suite, en par-

tant de paires de descripteurs orthogonaux (i.e. non corrélés entre eux), des modèles de rangs supérieurs sont développés en incluant de nouveaux descripteurs orthogonaux de manières successives tant qu'une augmentation de corrélation est observée.

Au final, l'analyse BMLR sélectionne les meilleurs modèles à chaque rang et le modèle final doit être choisi parmi eux. Celui-ci doit être suffisamment corrélé sans être sur-paramétré, ce qui amènerait alors à une dégradation du pouvoir prédictif pour des molécules hors du jeu d'entraînement. c'est à dire, la méthode « *stepwise* » semblant être la meilleure.

La méthodologie QSAR/QSPR a été développée et codée dans le programme CODESSA qui permet le calcul d'une centaine de descripteurs moléculaires basés sur des informations structurelles de type $Hansch$. Ces descripteurs moléculaires ont ensuite été utilisés dans l'analyse statistique par la méthode « Best MLR».

Dans cette étude, nous avons utilisé le programme CODESSA pour prédire le point de fusion des AGs. La Modélisation QSPR implémentée dans CODESSA permet de traiter et calculer jusqu'à 600 différents descripteurs de type électrostatiques, constitutionnels, géométriques, topologiques, quantiques, et thermodynamiques. Dans ce travail tous les descripteurs utilisés sont dérivés de la structure moléculaire et ne nécessitent pas de données expérimentales. Par la suite, la méthode de régression multilinéaire permet de traiter de nombreux descripteurs moléculaires, corrèle des variables indépendantes (descripteurs) avec des variables dépendantes T_f.

4.3.3 Analyse de régression linéaire multiple

Cette méthode a été présentée dans la section 2.3 (page 36). La MLR donne donc une équation du type

$$Y = a_1X_1 + a_2X_2 + ... + a_nX_n + C \tag{4.1}$$

Où $a_1, a_2, .., a_n$ et C sont des constantes.

4.3.4 Validation des modèles

Une fois le modèle mis en œuvre, le modèle doit être alors validé en termes de corrélation. Une démarche de validation interne et externe est réalisée pour améliorer la robustesse de l'analyse. On commence par les paramètres statistiques qui déterminent la qualité d'un modèle. Le plus répandu est le coefficient de corrélation R^2 qui

évalue la part de la variance expliquée par le modèle (Voir Section 2.4.3 (page 40)). Si R^2 = 0 : la variable indépendante n'explique rien, si R^2 = 1 : la variable explique complètement Y, c'est à dire les valeurs prédites et observées sont totalement corrélées et si R^2 = 0,93 : 93 $\%$ des variations de Y sont expliquées par le modèle. Bien entendu, un autre indicateur est l'erreur absolue moyenne (MAE, pour mean absolute error). D'autres critères sont utilisés pour valider un modèle consiste tel que le test de Fisher F (Voir Section 2.4.7 (page 43)). Ce critère, justifié dans le cas explicatif car basé sur une qualité d'ajustement, est aussi utilisé à titre indicatif pour comparer des séquences de modèles emboîtés.

Les techniques de validation croisée ont été appliquées pour l'évaluation de la prédiction interne. La validation croisée définie par «leave-one-out» ou «leave-many-out» (LOO) selon qu'une ou plusieurs molécules est (sont) retirée(s) consiste à recalculer le modèle sur (n-1) objets et à utiliser le modèle ainsi obtenu pour prédire la valeur de la variable dépendante du composé écarté. Le procédé est répété pour chacun des n objets de l'ensemble d'essai (Voir Section 1.4.2 (page 23)),

4.4 Résultats et discussions

Les données expérimentales de la température de fusion utilisées dans cette étude ont été extraites de la littérature [17]. Il s'agit de 62 composés d'acides gras présentés en (figure 4.1) le tableau (4.1) donne le nom et le point de fusion des AGs, qui sont des constituants importants des huiles naturelles, les valeurs de T_f sont classées dans l'ordre croissant.

Dans cette application, nous avons utilisé le programme CODESSA pour calculer plus de 320 descripteurs moléculaires. La procédure BMLR a été ensuite appliquée pour réduire le nombre des descripteurs à 278, afin d'obtenir les meilleurs modèles possibles.

Pour déterminer le nombre optimal de descripteurs décrivant le meilleur modèle parmi les 14 équations données par CODESSA, nous utilisons la méthode dite "over-parameterization". Sur tous les modèles obtenus, on trace la courbe de variation des coefficients de corrélations R^2 et les coefficients de corrélations R^2_{adj} ajustés en fonction du nombre de descripteurs. figure (4.2). Et c'est le meilleur modèle qui correspond à n descripteurs qui est choisi, lorsque la différence (R^2_{i+1} - R^2_i) est inférieure à 0,02. (0,02 a été choisie comme critère de point d'arrêt "$breakpoint\ criterion$") [16],

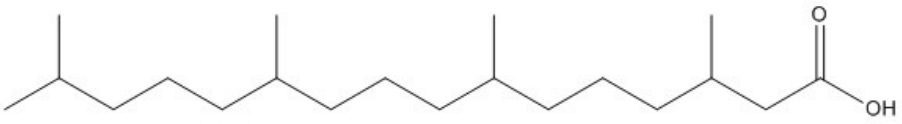
TABLE *4.1: Les valeurs expérimentales et prédites de point de fusion des AGs ; A, B et C sont les sous-ensembles utilisés dans la procédure de validation croisée*

		Structure	**Exp.** T_f	**Cal.** T_f	**Résiduel**
1	A	3-7-11-15-Tetramethylhexadecanoic acid	-65.00	-58.48	06.52
2	B	cis-cis-cis-cis-6-9-12-15-Octadecatetraenoic acid	-57.00	-30.60	26.40
3	C	cis-cis-cis-cis-5-8-11-14-Eicosatetraenoic acid	-49.00	-46.83	02.17
4	A	$(cis)_6$-4-7-10-13-16-19-Docosahexaenoic acid	-45.00	-48.57	-3.57
5	B	Pentanoic acid	-33.00	-23.17	9.83
6	C	3-Methylbutanoic acid	-29.00	-32.87	-3.87
7	A	cis-cis-cis-9-12-15-Octadecatrienoic acid	-11.00	-27.57	-16.57
8	B	cis-cis-9-12-Octadecadienoic acid	-7.00	-15.55	-8.55
9	C	Heptanoic acid	-7.00	2.78	9.78
10	A	Butanoic acid	-5.00	-5.86	-0.86
11	B	cis-9-Tetradecenoic acid	-4.00	7.32	11.32
12	C	cis-cis-5-13-Docosadienoic acid	-4.00	23.09	27.09
13	A	Hexanoic acid	-3.00	-9.30	-6.30
14	B	cis-9-Hexadecenoic acid	00.00	13.23	13.23
15	C	12-Hydroxy-cis-9-octadecenoic acid	05.00	17.75	12.75
16	A	Nonanoic acid	12.00	19.58	7.58
17	B	cis-9-Octadecenoic acid	13.00	16.96	3.96
18	C	cis-11-Octadecenoic acid	15.00	18.63	3.63
19	A	Octanoic acid	16.00	9.20	-6.80
20	B	cis-trans-9-11-Octadecadienoic acid	20.00	1.97	-18.03
21	C	trans-cis-10-12-Octadecadienoic acid	23.00	11.44	-11.56
22	A	cis-11-Eicosenoic acid	24.00	22.99	-1.01
23	B	cis-9-Eicosenoic acid	24.00	25.98	1.98
24	C	9-Decenoic acid	26.00	-1.29	-27.29
25	A	cis-5-Eicosenoic acid	27.00	28.72	1.72
26	B	Undecanoic acid	28.00	32.25	4.25
27	C	cis-6-Octadecenoic acid	29.00	18.29	-10.71
28	A	Decanoic acid	31.00	27.79	-3.21
29	B	cis-12-13-Epoxy-cis-9-octadecenoic acid	32.00	25.54	-6.46
30	C	trans-trans-cis-9-11-13-Octadecatrienoic acid	32.00	53.27	21.27
31	A	cis-11-Docosenoic acid	33.00	26.18	-6.82
32	B	cis-13-Docosenoic acid	34.00	28.14	-5.86

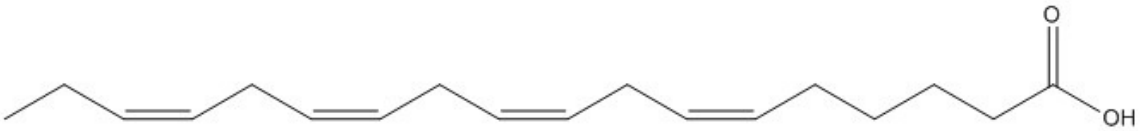
		Structure	Exp. T_f	Cal. T_f	Résiduel
33	C	trans-trans-cis-8-10-12-Octadecatrienoic acid	40.00	51.09	11.09
34	A	Tridecanoic acid	41.00	45.56	4.56
35	B	cis-15-Tetracosenoic acid	43.00	34.37	-8.63
36	C	Dodecanoic acid	43.00	39.59	-3.41
37	A	trans-11-Octadecenoic acid	44.00	43.13	-0.87
38	B	cis-trans-cis-9-11-13-Octadecatrienoic acid	45.00	51.33	6.33
39	C	trans-9-Octadecenoic acid	45.00	43.09	-1.91
40	A	cis-trans-trans-9-11-13-Octadecatrienoic acid	49.00	41.18	-7.82
41	B	Pentadecanoic acid	52.00	56.52	4.52
42	C	Tetradecanoic acid	54.00	51.09	-2.91
43	A	Heptadecanoic acid	61.00	63.70	2.70
44	B	trans-13-Docosenoic acid	61.00	58.86	-2.14
45	C	Hexadecanoic acid	62.00	57.66	-4.34
46	A	Nonadecanoic acid	69.00	67.84	-1.16
47	B	Octadecanoic acid	69.00	64.26	-4.74
48	C	trans-trans-trans-9-11-13-Octadecatrienoic acid	71.00	69.86	-1.14
49	A	Eicosanoic acid	76.00	70.73	-5.27
50	B	Pentacosanoic acid	77.00	83.88	6.88
51	C	Tricosanoic acid	79.00	79.26	0.26
52	A	Docosanoic acid	81.00	77.12	-3.88
53	B	Heneicosanoic acid	82.00	73.79	-8.21
54	C	cis-trans-trans-cis-9-11-13-15-Octadecatetraenoic acid	86.00	74.09	-11.91
55	A	Heptacosanoic acid	87.00	88.46	1.46
56	B	Tetracosanoic acid	87.00	80.95	-6.05
57	C	Hexacosanoic acid	88.00	85.69	-2.31
58	A	Nonacosanoic acid	90.00	92.97	2.97
59	B	Octacosanoic acid	90.00	90.45	0.45
60	C	Hentriacontanoic acid	93.00	97.57	4.57
61	A	Triacontanoic acid	93.00	94.84	1.84
62	B	Dotriacontanoic acid	96.00	102.76	3.08
Minimum			-65.00	-58.48	–
Maximum			96.00	102.76	–

Avec cette méthodologie, l'analyse statistique des modèles avec 2 jusqu'à 14 descripteurs, confirme que le meilleur modèle est celui donné par l'équation (4.2) avec 5 descripteurs :
Modèle (5 descripteurs) : $R^2 = 0.945; R^2_{adj} = 0.942; F = 197.93; SD = 0.099$

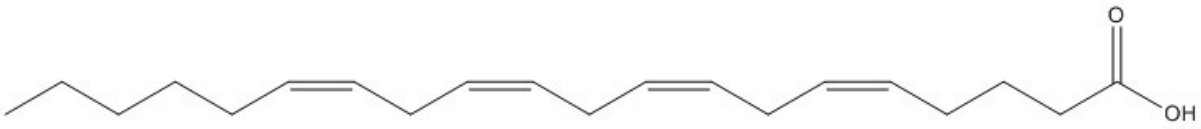
$$T_f = -70.1 + 17.6xD3 + 37.2xD4 - 31.4xD5 - 1.70xD6 + 267657xD7 \qquad (4.2)$$



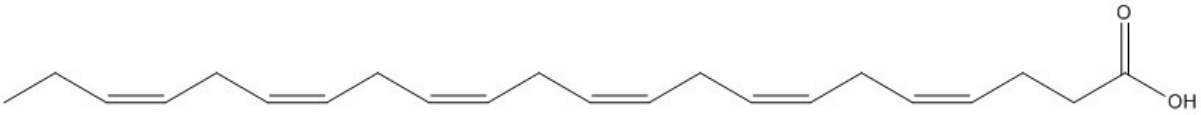
(01) 3-7-11-15-Tetramethylhexadecanoic acid



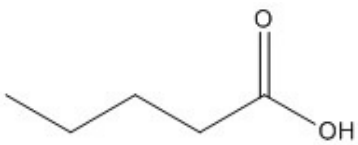
(02) cis-cis-cis-cis-6-9-12-15-Octadecatetraenoic acid



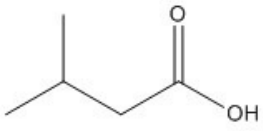
(03) cis-cis-cis-cis-5-8-11-14-Eicosatetraenoic acid



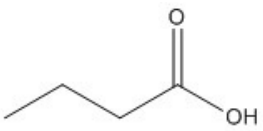
(04) cis-cis-cis-cis-cis-cis-4-7-10-13-16-19-Docosahexaenoic acid



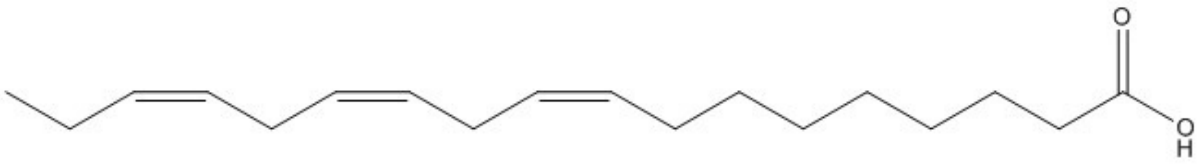
(05) Pentanoic acid



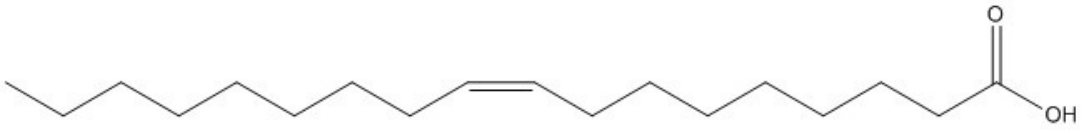
3-Methylbutanoic acid
(06)



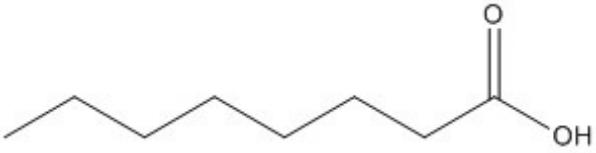
(07) Butanoic acid



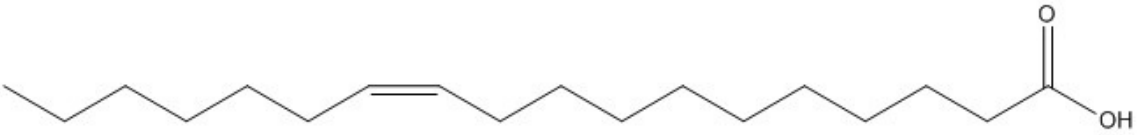
(08) cis-cis-cis-9-12-15-Octadecatrienoic acid



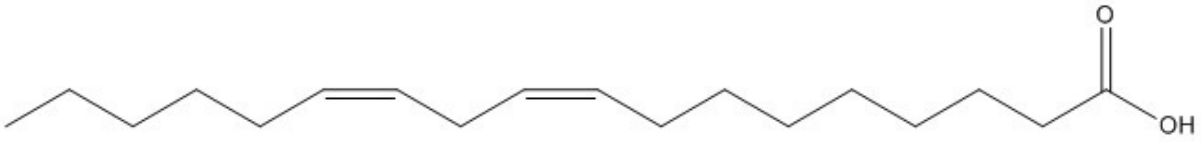
(17) cis-9-Octadecenoic acid



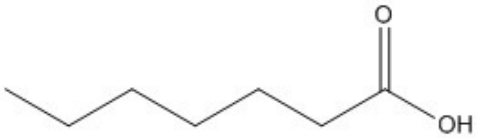
(18) Octanoic acid



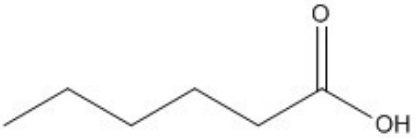
(19) cis-11-Octadecenoic acid



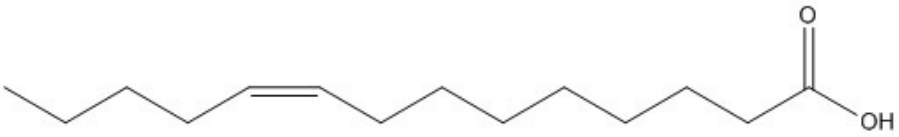
(09) cis-cis-9-12-Octadecadienoic acid



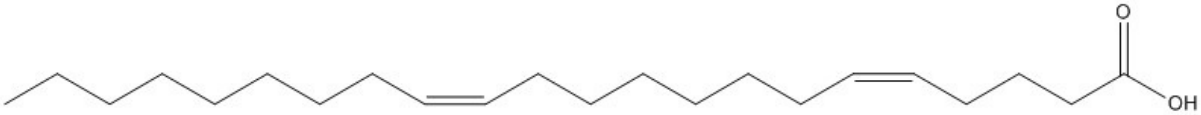
(10) Heptanoic acid



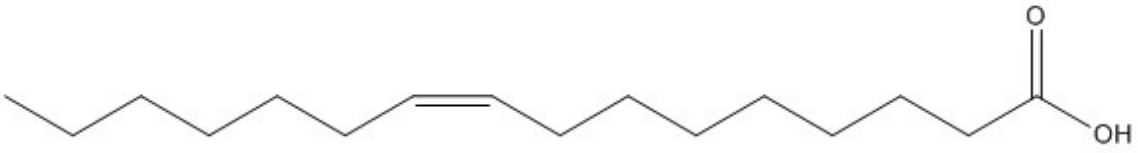
(11) Hexanoic acid



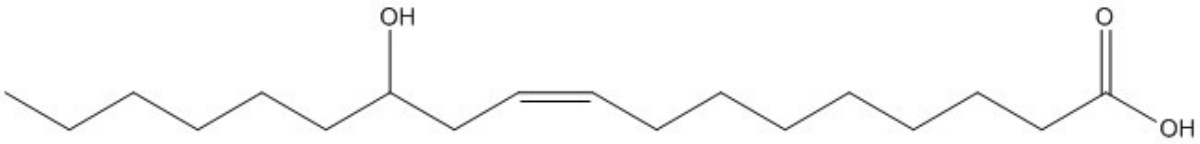
(12) cis-9-Tetradecenoic acid



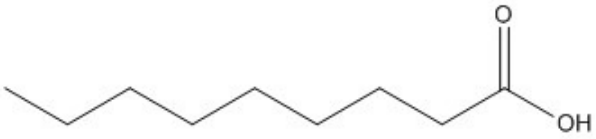
(13) cis-cis-5-13-Docosadienoic acid



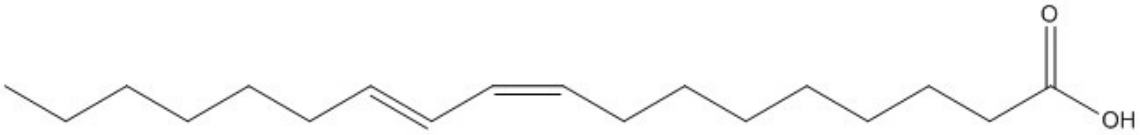
(14) cis-9-Hexadecenoic acid



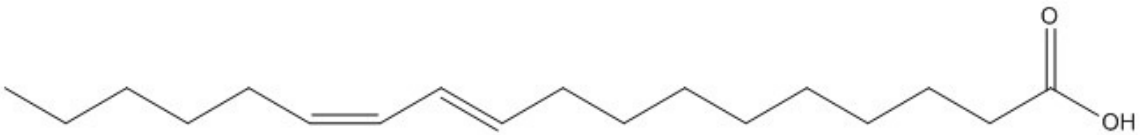
(15) 12-Hydroxy-cis-9-octadecenoic acid



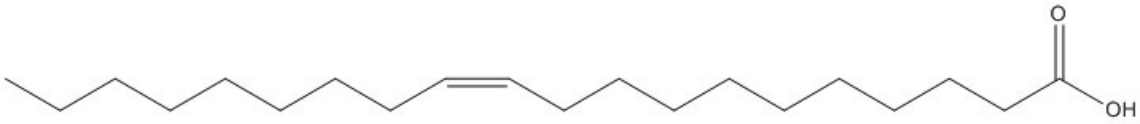
(16) Nonanoic acid



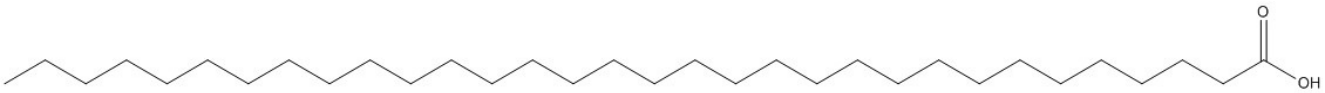
(20) cis-trans-9-11-Octadecadienoic acid



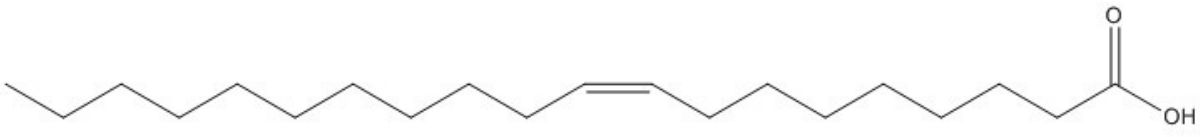
(21) trans-cis-10-12-Octadecadienoic acid



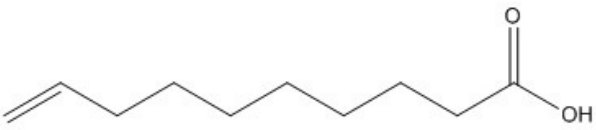
(22) cis-11-Eicosenoic acid



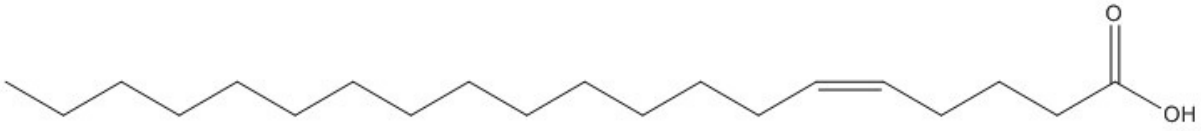
(23) Dotriacontanoic acid



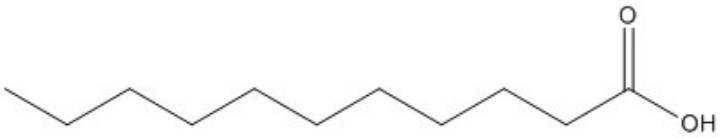
(24) cis-9-Eicosenoic acid



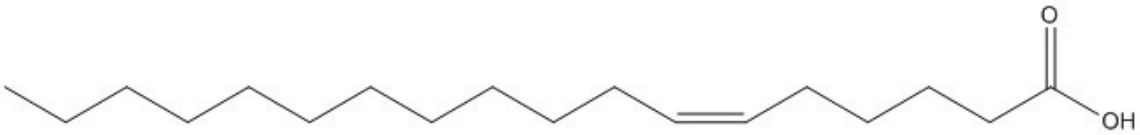
(25) 9-Decenoic acid



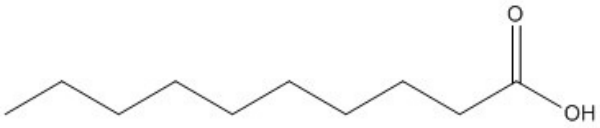
(26) cis-5-Eicosenoic acid



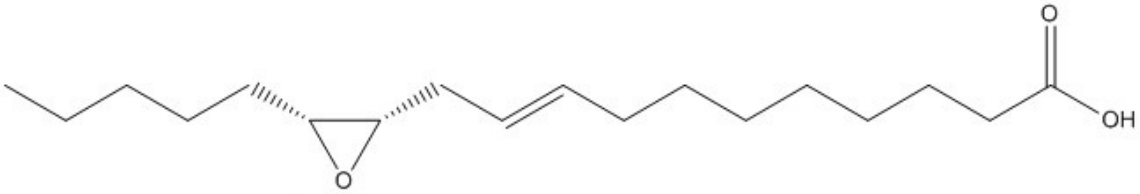
(27) Undecanoic acid



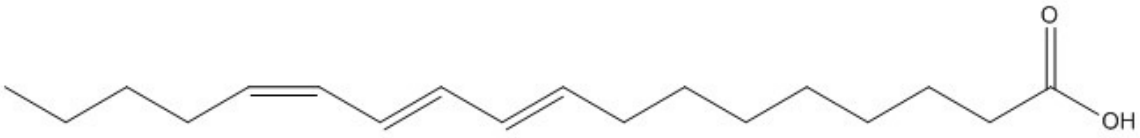
(28) cis-6-Octadecenoic acid



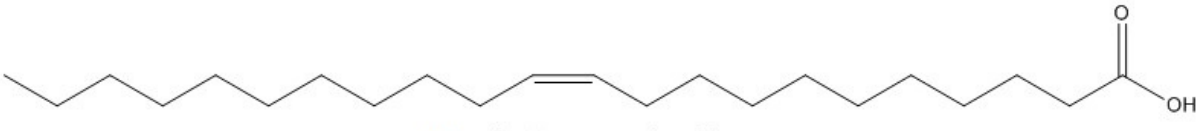
(29) Decanoic acid



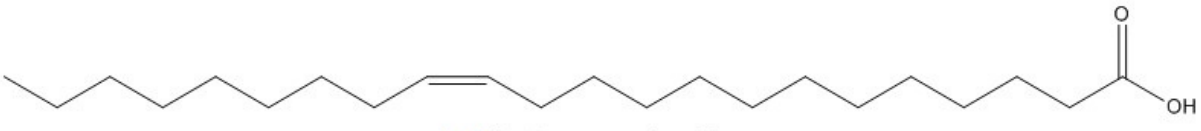
(30) cis-12-13-Epoxy-cis-9-octadecenoic acid



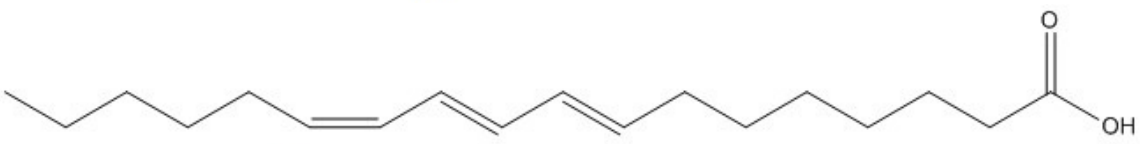
(31) trans-trans-cis-9-11-13-Octadecatrienoic acid



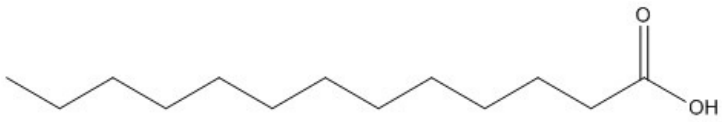
(32) cis-11-Docosenoic acid



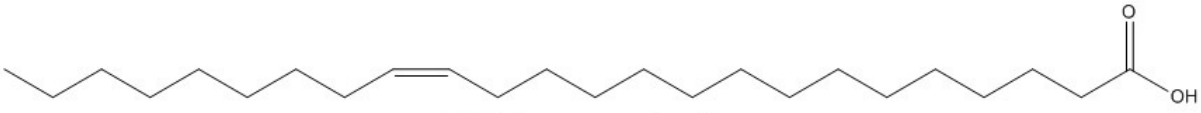
(33) cis-13-Docosenoic acid



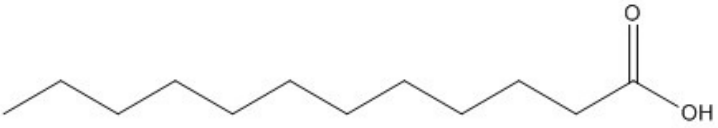
(34) trans-trans-cis-8-10-12-Octadecatrienoic acid



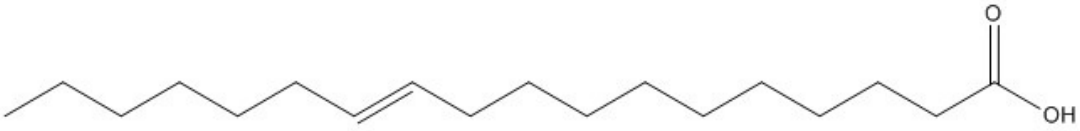
(35) Tridecanoic acid



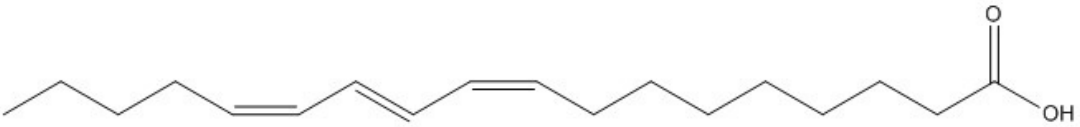
(36) cis-15-Tetracosenoic acid



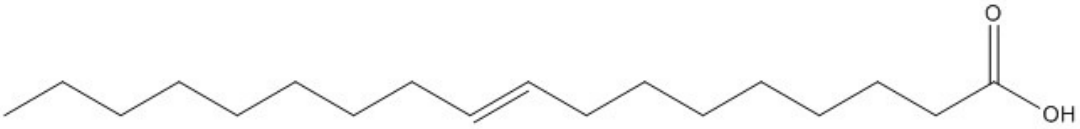
(37) Dodecanoic acid



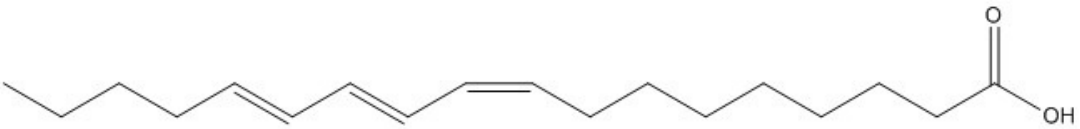
(38) trans-11-Octadecenoic acid



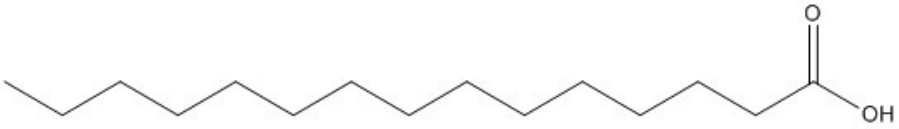
(39) cis-trans-cis-9-11-13-Octadecatrienoic acid



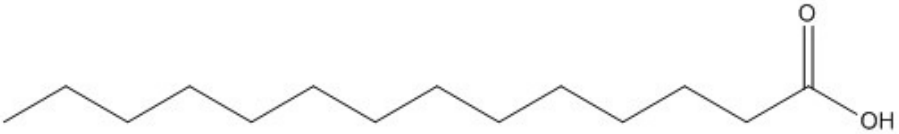
(40) trans-9-Octadecenoic acid



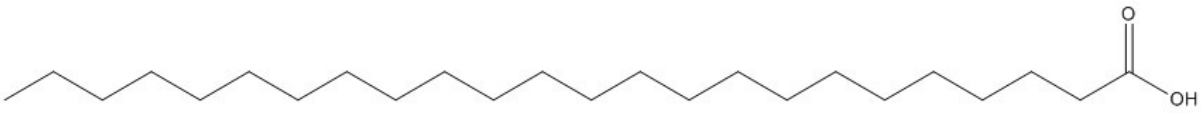
(41) cis-trans-trans-9-11-13-Octadecatrienoic acid



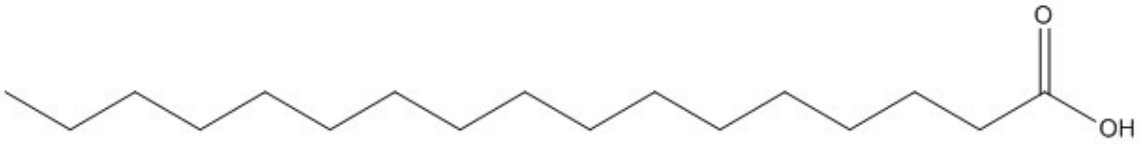
(42) Pentadecanoic acid



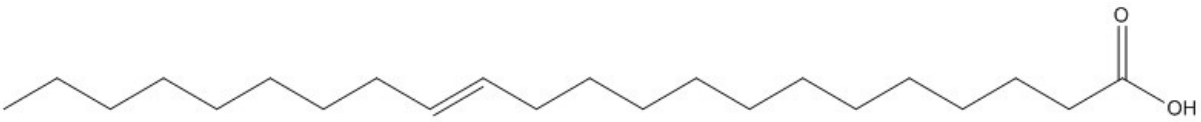
(43) Tetradecanoic acid



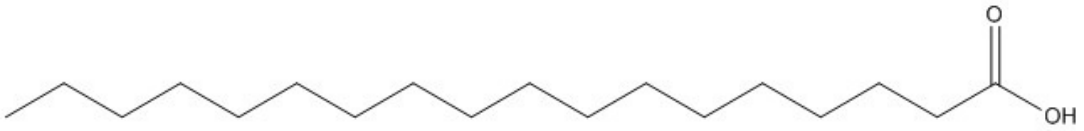
(44) Tetracosanoic acid



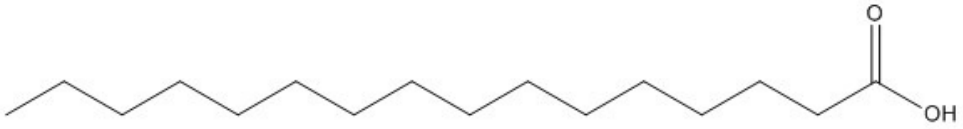
(45) Heptadecanoic acid



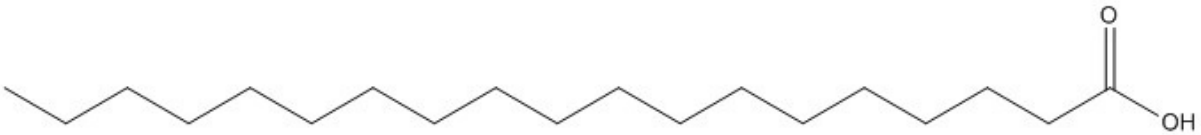
(46) trans-13-Docosenoic acid



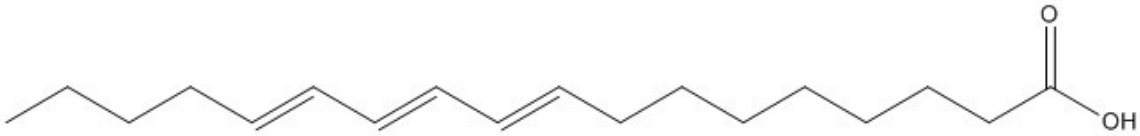
(47) Octadecanoic acid



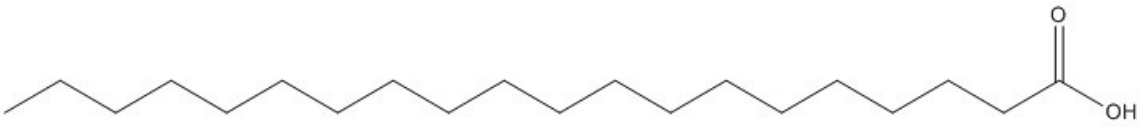
(48) Hexadecanoic acid



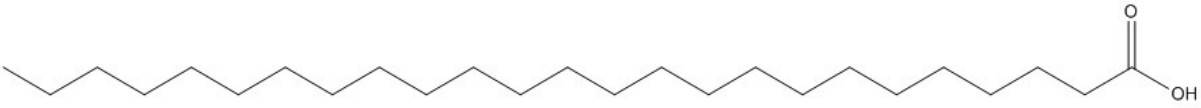
(49) Nonadecanoic acid



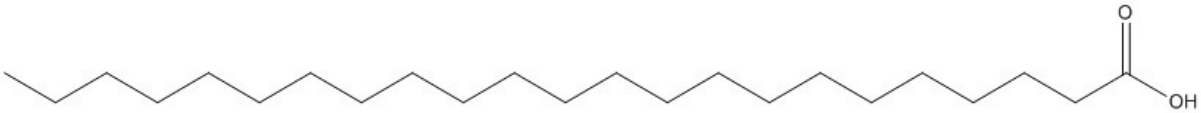
(50) trans-trans-trans-9-11-13-Octadecatrienoic acid



(51) Eicosanoic acid



(52) Pentacosanoic acid



(53) Tricosanoic acid

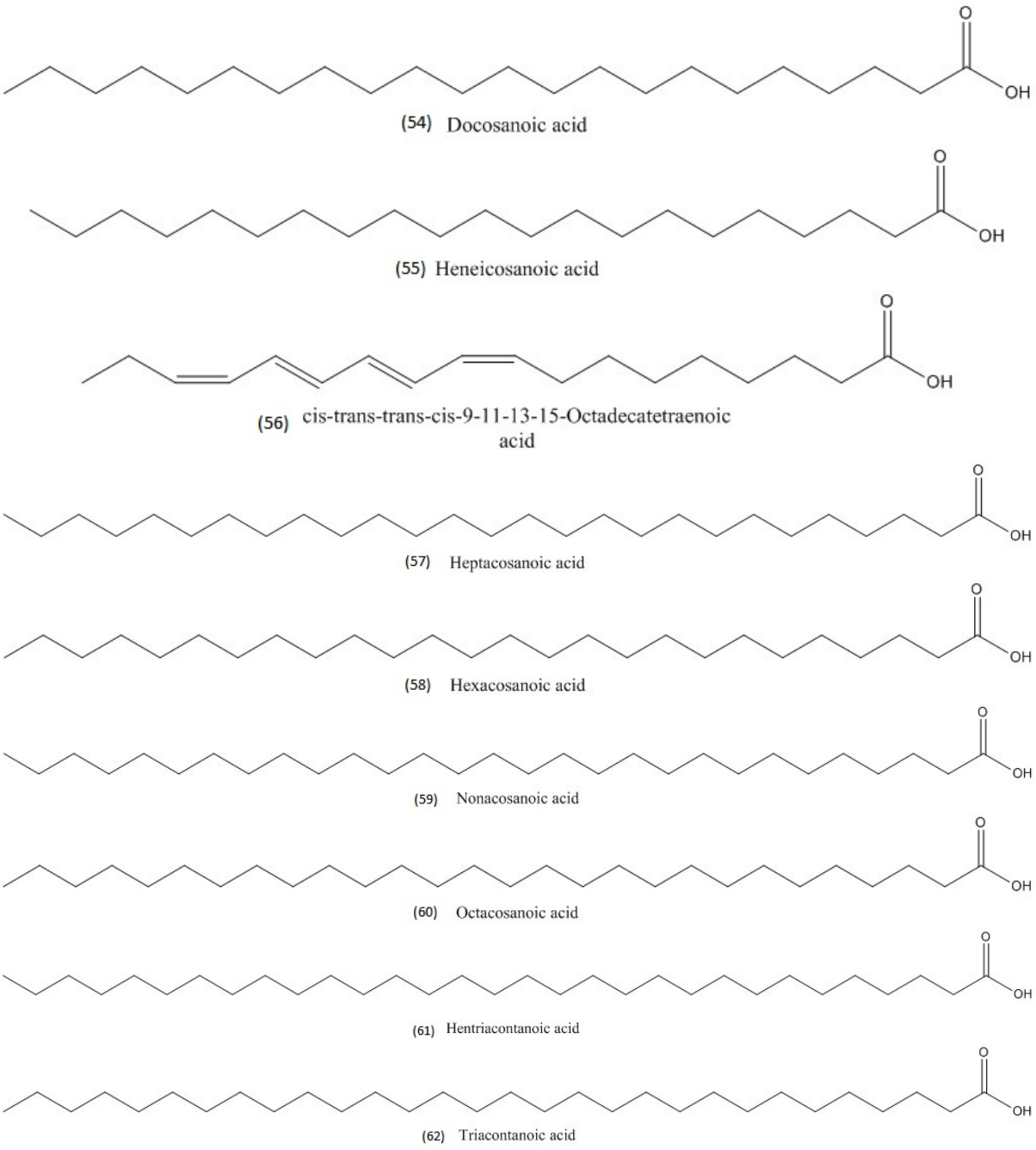


FIGURE *4.1: Acides Gras 1-62*

En outre, dans la figure (4.2), les très faibles écarts entre les valeurs de (R^2 et R^2_{adj}) montrent à chaque fois la qualité de l'ajustement et la capacité de prédiction interne.

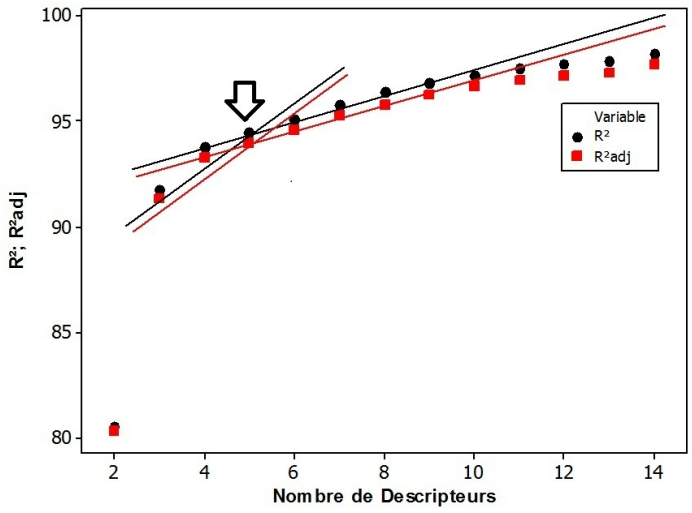


FIGURE *4.2: Evolution des $R^2 et R^2_{adj}$ en fonction du nombre de descripteurs*

Le tableau (4.2) représente les modèles obtenus avec deux, trois, quatre et cinq descripteurs. Tous ces modèles ont été validés par la méthode "cross validation" selon la procédure de Leave-One-Out (LOO).

Les descripteurs D1, D7 sont définis dans le tableau (4.3),

Les descripteurs moléculaires sélectionnés dans le modèle avec 5 descripteurs sont :
D3 (the Number of aromatic bonds).
D4 (Average Complementary Information content (order 1)).
D5 (Balaban index).
D6 (YZ Shadow).
D7 (min atomic orbital electronic population).

Selon les valeurs absolues du t-test (|t-test|), l'importance des descripteurs impliqués dans le modèle diminue dans l'ordre suivant : D5 > D3 > D4 > D6 > D7.

TABLE *4.2: Les modèles QSPR obtenus avec 2, 3, 4 et 5 descripteurs*

Model #1 : $R^2 = 0.81; R^2_{adj} = 0.80, F = 281.67$; Sd = 17.88				
Descripteur	X	Δ X	t-test	p-value
Intercept	-51.98	9.96	-5.22	
D1	-21.21	1.91	-11.08	<10-4
D2	8.83	0.78	11.20	<10-4

Model #3 : $R^2 = 0.92; R^2_{adj} = 0.91, F = 215.46$; Sd = 12.03				
Descripteur	X	Δ X	t-test	p-value
Intercept	-38.59	8.38	-4.604	
D3	0.44	0.03	11.62	<10-4
D4	0.73	0.03	19.34	<10-4
D6	-0.57	0.03	-14.96	<10-4

Model #4 : $R^2 = 0.93; R^2_{adj} = 0.93, F = 213.73$; Sd = 19.52				
Descripteur	X	X	t-test	p-value
Intercept	75.67	19.52	3.87	
D3	0.48	0.03	14.14	<10-4
D4	0.75	0.03	22.26	<10-4
D5	-0.17	0.03	-4.42	<10-4
D6	-0.41	0.03	-10.70	<10-4

Model #2 : $R^2 = 0.94; R^2_{adj} = 0.94, F = 197.93$; Sd = 9.94				
Descripteur	X	Δ X	t-test	p-value
Intercept	227.18	20.84	10.90	
D3	-1.71	0.12	-13.95	<10-4
D4	17.00	1.25	13.57	<10-4
D5	-164.12	10.82	-15.17	<10-4
D6	3.16	0.45	6.97	<10-4
D7	15601	4745	3.29	0.002

TABLE *4.3: Les descripteurs impliqués dans les modèles #1-4*

Variable	Signification
D1	Number of double bonds " Constitutional "
D2	PPSA-3 Atomic charge weighted PPSA " Electrostatic "
D3	Number of aromatic bonds " Electrostatic "
D4	Average Complementary Information content (order 1)"Topological "
D5	Balaban index "Topological "
D6	YZ Shadow "Geometrical"
D7	Min atomic orbital electronic population " Quantum Chemical "

PPSA-3 : Atomic charge weighted $\mathbb{P}artial\ \mathbb{P}ositively\ charged\ \mathbb{S}urface\ \mathbb{A}rea$ [25]

$$PPSA-3 = \sum_{A} q_A.S_A \qquad A \in \{\delta_A > 0\} \tag{4.3}$$

S_A - La charge positive de la surface atomique accessible au solvant.

q_A - la charge partielle atomique.

Shadow areas of a molecule : Les zones d'ombre de la molécule [26]

$$S_K = \frac{1}{2}\oint(vdp - pdv) \tag{4.4}$$

C - La projection de la molécule sur le plan défini par deux axes principaux de la molécule (k = XY, XZ et YZ).

v - x ou y

p - y ou z

La matrice de corrélation est présentée dans le tableau (4.4). Les corrélations respectives de la température de fusion et les cinq descripteurs sélectionnés de modèle #2 sont inférieur à 0,5. Ce qui est acceptable, et le modèle est suffisamment stable pour supposer que les descripteurs sont indépendants dans le modèle QSPR élaboré. Nous notons que le modèle à 5 paramètres présente des données statistiques satisfaisantes, coefficient de corrélation $R^2 = 0.945$, coefficient de corrélation ajusté $R^2_{adj} = 0.94$ et test de Fisher = 197.

TABLE *4.4: Matrice de corrélation des cinq descripteurs impliqués dans le modèle* #2

D3	D4	D5	D6	D7	Mp	
D3	1.00					
D4	-0.14	1.00				
D5	0.20	0.01	1.00			
D6	0.16	-0.03	0.49	1.00		
D7	-0.14	-0.44	0.02	-0.18	1.00	
Mp	0.25	0.68	-0.28	-0.52	-0.22	1.00

Pour valider le modèle #2, deux approches sont souvent utilisées *(i)* : validation interne et *(ii)* validation externe. Nous avons utilisé la première approche en appliquant les étapes suivantes : *(1)* Toutes les températures de fusion sont classées dans l'ordre croissant *(2)*. Ces 62 points ont été subdivisés en trois sous-ensembles (A,B et C) : le premier, quatrième, septième, etc points forment le premier sous-ensemble (A), le deuxième, cinquième, huitième, etc points forment le second sous-ensemble (B), et le troisième, sixième, neuvième, etc points forment le troisième sous-ensemble (C). *(3)* Trois nouveaux ensembles "training set" ont été construits en utilisant les combinaisons des sommes binaires : (A + B), (A + C) et (B + C). *(4)*, les procédures QSPR, y compris la méthode de régression multilinéaire (B-MLR) ont été appliquées aux trois sous-ensembles de données obtenus à l'étape 3, c'est à dire pour chaque "training set" les équations de corrélations ont été obtenues avec les mêmes descripteurs correspondant au modèle #2. *(5)* Le modèle #2 a été validé une deuxième fois à l'aide des procédures de validation croisée interne "leave many-out" ; suivant la procédure décrite ci-dessus qui a été appliquée à l'ensemble de 62 points. Trois sous-ensembles d'entraînement sont construits avec des 42 composés et les 20 composés restants ont été utilisés comme données de validation externe tableau (4.5). L'efficacité des modèles QSPR pour prédire les valeurs des températures de fusion en utilisant le R^2_{adj}. Les valeurs moyennes de $R^2_{adj(Fit)}$ et $R^2_{adj(Pred)}$ sont très proches (0,94 et 0,91, respectivement), tableau (4.5) qui suggère la prédictivité relativement stable du modèle de QSPR proposé.

Les températures de fusions expérimentales et prédites sont présentées dans le tableau (4.1) obtenues avec l'équation (4.2) (Model #2), la figure (4.3), confirme l'ac-

cord entre les valeurs calculées et expérimentales pour chaque molécule.

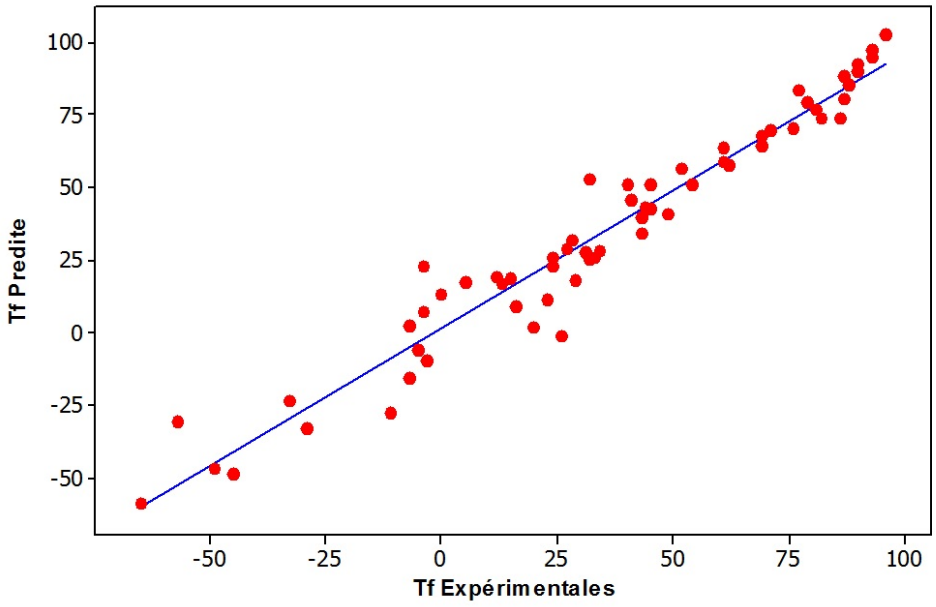


FIGURE *4.3: Corrélation entre les températures de fusions expérimentales et prédites*

TABLE *4.5: Validation interne du modèle #2 l'équation (4.2)*

Série d'apprentissage	N	$R^2_{(fit)}$	$R^2_{adj}(fit)$	F	Série de test	N	$R^2_{(pre)}$	$R^2_{adj}(pre)$	F
A + B	42	0.947	0.920	130.62	C	20	0.961	0.919	69.85
A + C	42	0.951	0.900	137.00	B	21	0.972	0.650	106.34
B + C	42	0.950	0.935	173.56	A	21	0.954	0.814	62.25
average		0.949	0.918	147.06			0.962	0.794	79.48

4.5 Conclusion

En résumé, nous avons utilisé la méthode de la régression multilinéaire BMLR implémentée dans le logiciel CODESSA pour élaborer des modèles QSPR fiables capables de prédire le point de fusion de 62 acides gras dont les valeurs expérimentales comprises entre -65.00 et 96.00 $^{\circ}$C afin de mettre en place le modèle le plus performant possible. Le meilleur modèle QSPR est caractérisé par les paramètres statistiques ($R^2 = 0,945, R^2_{adj} = 0,942$, F= 190,90). La présente étude montre que ce

modèle est défini avec cinq descripteurs moléculaires non corrélés entre eux et qui sont essentiellement de type électrostatiques et topologiques. Par conséquent, le modèle QSPR élaboré pourrait être utilisé pour prédire le point de fusion de nouveaux acides gras ou des acides gras pour lesquels la température de fusion expérimentale est indisponible dans la littérature.

Bibliographie

[1] Claus K. Zéberg-Mikkelsen., Erling H. Stenby., Predicting the melting points and the enthalpies of fusion of saturated triglycerides by a group contribution method, Fluid Phase Equilibria, 162, (**1999**), 1, 7.

[2] Yi Liu, Andrew J. Holder., A quantum mechanical quantitative structure–property relationship study of the melting point of a variety of organosilicons, J Mol Graph Model, 31, (**2011**), 57.

[3] Dorin Dadarlat, Dane Bicanic, Jurgen Gibkes, William Kloek, Ivon van den Dries, Edo Gerkema, Study of melting processes in fatty acids and oils mixtures. A comparison of photopyroelectric (PPE) and differential scanning calorimetry (DSC), Chemistry and Physics of Lipids, 82, (**1996**), 1, 15

[4] Nasrin Farahani, Farhad Gharagheizi, Seyyed Alireza Mirkhani, Kaniki Tumba, Ionic liquids : Prediction of melting point by molecular-based model, Thermochimica Acta, 549, (**2012**), 17.

[5] Katritzky A.R., Lobanov, V.S., Karelson, M., Codessa : Comprehensive Descriptors for Structural and Statistical Analysis, User Manual. University of Florida, Gainesville, (**1997**), Florida.

[6] Roberto T., V, Consonni., Handbook of Molecular Descriptors, (**2000**), WILEY-VCH.

[7] Fahy E,. Subramaniam, S,. Brown, A,. A comprehensive classification system for lipids. J Lipid Res., 46, (**2005**), 839.

[8] Guesnet P,. Alessandri J. M,. Astorg P,. Les rôles physiologiques majeurs exercés par les acides gras polyinsaturés (AGPI). Oléagineux, Corps gras et Lipides, 12, (**2005**), 333.

[9] Bailey A.E., Melting and Solidification of Fats. Interscience Publishers, (**1950**), Inc.,New York.

[10] Moziar C., deMan, J.M., deMan, L.,Effect of tempering on the physical properties of shortening. Can. Inst. Food Sci. Technol. J., 22, (**1989**), 238.

[11] Ghotra B.S., Dyal, S.D., Narine, S.S., Lipid shortenings : a review. Food Res. Int., 35, (**2002**), 1015.

[12] Humphrey K.L., Moquin, P., Narine, S.S., Phase behavior of a binary lipid shortening system : from molecules to rheology. J. Am. Oil Chem. Soc., 80, (**2004**), 1175.

[13] Narine S.S., Marangoni, A.G., Relating structure of fat networks to mechanical properties : a review. Food Res. Int., 32, (**1999**), 227.

[14] Narine S.S., Marangoni, A.G., Structure and mechanical properties of fat crystal networks. Adv. Food Nutr. Res., 44, (**2002**), 33.

[15] Svenstrup G., Bruggemann, D., Kristensen, L., Risbo, J., Skibsted, L.H., The influence of pretreatment on pork fat crystallization. Eur. J. Lipid Sci. Technol., 107, (**2005**), 607.

[16] Katritzky A.R., Taemm, K., Kuanar, M., Fara, D., Oliferenko, A., Oliferenko, P., et al., Aqueous biphasic systems. Partitioning of organic molecules : a QSPR treatment. J. Chem. Inf. Comput. Sci., 44, (**2004**), 136.

[17] David R. Lide, CRC Handbook of Chemistry and Physics 89th (**2009**), Edition CRC Press.

[18] Chemdraw is a molecule editor © (**2008**), http ://www.cambridgesoft.com/

[19] Stewart J.J.P., Optimization of parameters for semiempirical methods. V. Modification of NDDO approximations and application to 70 elements. J. Mol. Model., 13, (**2007**), 1173.

[20] Stewart J.J.P., MOPAC2009. Stewart Computational Chemistry. Colorado Springs, CO, USA, (**2008**), Available from : http ://openmopac.net/MOPAC2009.html.

[21] Becke A. D., Correlation energy of an inhomogeneous electron gas : A coordinate-space model, J. Chem. Phys., 88, (**1988**), 1053

[22] Becke A. D. Density-Functional Exchange-Energy Approximation with Correct Asymptotic Behavior, Phys. Rev. A., 38, (**1988**), 3098.

[23] Gaussian 03, Revision .E01, Frisch, M. J., Trucks, G. W., Schlegel, H. B., Scuseria, G. E., Robb, M. A., Cheeseman, J. R., Scalmani, G., Barone, V., Mennucci, B., Petersson, G. A., Nakatsuji, H., Caricato, M., Li, X., Hratchian, H. P., Izmaylov, A. F., Bloino, J., Zheng, G., Sonnenberg, J. L., Hada, M., Ehara, M., Toyota, K., Fukuda, R., Hasegawa, J., Ishida, M., Nakajima, T., Honda, Y., Kitao, O., Nakai, H., Vreven, T., Montgomery, J. A., Jr., Peralta, J. E., Ogliaro, F., Bearpark, M., Heyd, J. J., Brothers, E., Kudin, K. N., Staroverov, V. N., Kobayashi, R., Normand, J., Raghavachari, K., Rendell, A., Burant, J. C., Iyengar, S. S., Tomasi, J., Cossi, M., Rega, N., Millam, M. J., Klene, M., Knox, J. E., Cross, J. B., Bakken, V., Adamo, C., Jaramillo, J., Gomperts, R., Stratmann, R. E., Yazyev, O., Austin, A. J., Cammi, R., Pomelli, C., Ochterski, J. W., Martin, R. L., Morokuma, K., Zakrzewski, V. G., Voth, G. A., Salvador, P., Dannenberg, J. J., Dapprich, S., Daniels, A. D., Farkas, Ö., Foresman, J. B., Ortiz, J. V., Cioslowski, J., Fox, D. J. Gaussian, Inc., (**2004**), Wallingford CT,

[24] Hansch C., Leo, A., et Hoekmann, D. Exploring QSAR : hydrophobic, electronic and steric constants. Washington, DC., (**1995**), American Chemical Society,

[25] **(a)** Stanton D.T., P.C. Jurs, Anal. Chem., 62, (**1990**), 2323 , **(b)** Stanton D.T., L.M. Egolf, P.C. Jurs, M.G. Hicks, J. Chem. Inf. Comput. Sci., 32, (**1992**), 306.

[26] Rohrbaugh R. H., P. C. Jurs, Anal. Chim. Acta., 199, (**1987**), 99.

[27] **(a)** Balaban A. T., Chem. Phys. Lett., 89, (**1981**), 399. **(b)** Balaban A. T., Pure and Appl. Chem., 55, (**1983**), 199.

CHAPITRE 5

Modélisation QSPR des constantes d'acidité des acides Benzoïques

Résumé :

Notre objectif dans cette application est la prédiction théorique des valeurs de pKa des acides benzoïques, en solution aqueuse. Les calculs des énergies libres de Gibbs de déprotonation des acides et de leurs anions correspondants en phase gazeuse et phase aqueuse, ont été effectués en utilisant la théorie de la fonctionnelle de densité (DFT) avec la fonctionnelle hybride B3LYP et la base étendue 6-311++G (d, p). Les effets de solvant ont été pris en compte dans les calculs en phase aqueuse avec le modèle de solvatation SMD. Les cycles thermodynamiques ont été étudiés pour décrire la déprotonation dans les deux phases. Des modèles QSPR fiables ont été obtenus avec les techniques de régression linaire simple entre les valeurs expérimentales aqueuses de pKa de 51 acides benzoïques et ΔG_{aq}.

Le pouvoir prédictif de ces modèles a été testé avec succès sur une série de test constituée de 25 acides non inclus dans la série d'apprentissage. L'écart absolu moyen des valeurs de pKa calculé SD (déviation standard) est inférieur à 0,36 en unités de pKa et le coefficient de corrélation $R^2 = 0,93$.

5.1 Introduction

L'acidité d'une solution est une propriété physico-chimique importante est essentielle pour la compréhension de nombreuses réactions fondamentales de la chimie et de la biochimie, généralement exprimée en pKa. Il s'agit d'un facteur primordial dans la pharmacocinétique [i] des médicaments et dans les interactions des protéines avec d'autres molécules. La détermination des valeurs de pKa est d'un intérêt particulier pour de nombreux chimistes dans toutes les branches de la chimie et sciences de la vie [1, 2]. La valeur de pKa d'une molécule détermine la quantité de formes protonées et non-protonées à un pH spécifique, une discussion détaillée de l'importance de pKa peut être trouvée dans les travaux de *Schüürmann* [3] et *Stewart* [4].

Au cours des dernières années, il y a eu de nombreuses études portant sur les méthodes pour améliorer l'efficacité de la recherche de nouveaux médicaments dans le but de réduire le temps de développement. En outre, la propriété pKa d'une molécule de médicament est un paramètre clé pour le développement de celui-ci, car il régit l'absorption, la distribution, le métabolisme et l'élimination, en particulier pour le développement de nouvelles API (Active pharmaceutical ingredient). Le pKa est devenu d'une grande importance car le transport de médicaments dans les cellules et entre les membranes dépend des propriétés physico-chimiques et de la pKa des médicaments [5]. Il existe plusieurs méthodes pour la détermination des constantes de dissociation, la potentiométrie [6, 7] et UV-VIS spectrométrie d'absorption [8] sont les techniques les plus utilisées, du fait de leur précision et de reproductibilité. Récemment, de nouvelles approches ont été appliquées pour prédire les constantes d'acidité, en se appliquant des méthodes de chimie quantique, qui permettent une meilleure compréhension des facteurs structurels et environnementaux qui influent sur les valeurs de pKa dans différents systèmes [9]. Plusieurs travaux décrivant les méthodes de calcul et le protocole à suivre pour la détermination des constantes d'acidité peuvent être retrouvés dans la littérature [10–17].

L'acide benzoïque est un agent antimicrobien et antifongique. Il est utilisé comme conservateur ou additif alimentaire et il est également présent dans certaines plantes. L'acide benzoïque est parfois combiné avec l'alcool et l'eau pour utilisation comme

i. La pharmacocinétique est l'étude des actions d'une substance active contenue dans un médicament sur l'organisme après son ingestion. La pharmacocinétique désignée sous le signe ADME : Absorption, Distribution, Métabolisme et Excrétion

agent de nettoyage en médecine. La faible acidité de l'acide benzoïque est utilisée pour convertir les autres composés (typiquement les esters) en composants (médicaments) digestibles plus facilement. L'acide benzoïque est également utilisé comme un lubrifiant pour l'ingestion des comprimés et les pilules [18].

L'objectif principal de cette application est l'élaboration de modèles QSPR pour la prédiction des valeurs des constantes d'acidité des acides benzoïques en utilisant une base de données expérimentale. Dans cette application, nous avons utilisé la fonctionnelle B3LYP qui est l'une des méthodes qui a le mieux réussi de la chimie quantique.

Le choix de cette méthode se trouve justifié par le fait qu'elle tient compte de la corrélation électronique, d'une part, et par le fait qu'elle est moins coûteuse en temps de calcul d'autre part en comparant avec les autres méthodes comme G2 G3, G2MP2, QCISD(T), CBS-4, CBS-Q, CBS-QB3, and CBS-APNO qui sont utilisées seulement pour les molécules de petites tailles.

5.2 Méthodologie

5.2.1 Calcul d'acidité

La détermination des valeurs de constantes d'acidité est essentielle pour comprendre le fonctionnement de nombreuses réactions chimiques et biochimiques. Selon *Brønsted-Lowry*. L'acide est défini comme toute substance capable de céder des protons H^+. Inversement, une base est toute substance capable de capter des protons.

$$AH \leftrightharpoons A^- + H^+ \tag{5.1}$$

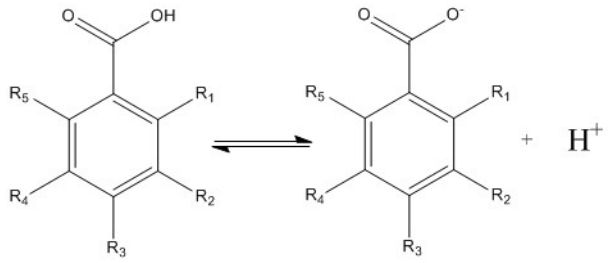
$$AH + H_2O \leftrightharpoons A^- + H_3O^+ \tag{5.2}$$

pKa est défini comme $pKa = -log_{10}\, ka$

Les deux méthodes les plus utilisées sont représentées dans les schémas (5.1) et (5.2), basées sur les réactions (5.1) et (5.2) et les travaux de (*Liptak et al.*) [15, 19, 20].

Dans les deux schémas (5.1) et (5.2), pour l'acide AH et sa base conjuguée A^-,

• ΔG_{aq} représente la variation de l'enthalpie libre de Gibbs de déprotonation en solution.



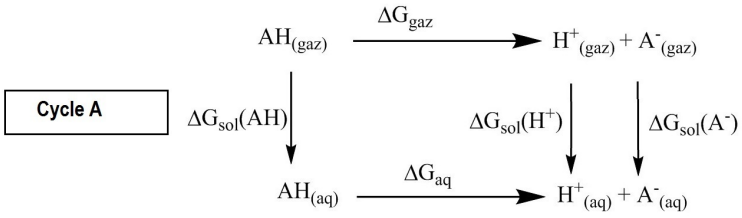


Schéma (5.1)

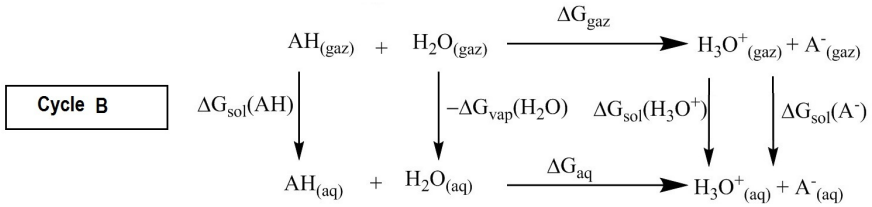


Schéma (5.2)

• ΔG_{gaz} représente la variation de l'enthalpie libre de Gibbs de déprotonation en phase gazeuse.

• $\delta\Delta G_{solv}$ représente la variation de l'enthalpie libre de solvatation.

Pour calculer les constantes d'acidité en phase aqueuse, on utilise le cycle thermodynamique 1, (Schéma 1., modèle A, et l'équation (5.3)), selon le protocole suivant.

$$pKa = \frac{\Delta G_{aq}}{RTln(10)} \tag{5.3}$$

Pour le cycle 1

$$\Delta G_{aq} = \Delta G_{gaz} + \delta\Delta G_{solv}$$

Où

$$\Delta G_{gaz} = G^{\circ}_{gaz}(H^+) + G^{\circ}_{gaz}(A^-) - G^{\circ}_{gaz}(AH)$$

Et

$$\delta\Delta G_{solv} = \Delta G_{solv}(H^+) + \Delta G_{solv}(A^-) - \Delta G_{solv}(AH)$$

$$\Delta G_{aq} = G^{\circ}_{gaz}(H^+) + G^{\circ}_{gaz}(A^-) - G^{\circ}_{gaz}(AH) + \Delta G_{solv}(H^+) + \Delta G_{solv}(A^-) - \Delta G_{solv}(AH) \quad (5.4)$$

Toutes ces enthalpies libres sont calculées par les méthodes de chimie quantique sauf celles $G^{\circ}_{gaz}(H^+)$ et $\Delta G_{solv}(H^+)$, pour lesquelles les valeurs expérimentales sont utilisées [21–23].

$G^{\circ}_{gaz}(H^+)$ = -6.28 kcal/mol.

$\Delta G_{solv}(H^+)$ = -265.9 kcal/mol.

Un terme de correction doit être ajouté, afin de combiner la phase gazeuse avec la phase aqueuse. L'état standard en phase gazeuse (1 atm) doit être converti en solution (1 M), et ceci par l'utilisation de la formule suivante :

$$G(1M) = G(1atm) + RTln[Vm] \quad (5.5)$$

$$= G(1atm) + RTln(24.46) \quad (5.6)$$

$$= G(1atm) + 1.8929(kcal/mol) \quad (5.7)$$

D'après les travaux de *Cramer et Liptak* [11, 15, 19, 24], la source de l'erreur dans le calcul de pKa est la variation d'énergie libre de solvatation ΔG_{solv}, qui est basée sur le type de modèle de solvatation utilisé et le niveau de calcul quantique.

Le cycle thermodynamique 1 (Schéma 1.) et les équations (5.1, 5.3 et 5.5) représentent un outil fiable pour prédire les constantes d'acidité, selon l'expression suivante,

$$pKa = [G^{\circ}_{gaz}(A^-) - G^{\circ}_{gaz}(AH) + \Delta G_{solv}(A^-) - \Delta G_{solv}(AH) - 270.283]/1.364 \quad (5.8)$$

Les quatre énergies sont calculées en kcal/mol.

Pour le cycle 2 [22–26]

$$\Delta G_{aq} = \Delta G_{gaz} + \delta \Delta G_{solv}$$

Où

$$\Delta G_{gaz} = G^{\circ}_{gaz}(H_3O^+) + G^{\circ}_{gaz}(A^-) - G^{\circ}_{gaz}(AH) - G^{\circ}_{gaz}(H_2O)$$

Et

$$\delta \Delta G_{solv} = \Delta G_{solv}(H_3O^+) + \Delta G_{solv}(A^-) - \Delta G_{solv}(AH) - \Delta G_{vap}(H_2O)$$

$$\Delta G_{aq} = G^{\circ}_{gaz}(H_3O^+) + G^{\circ}_{gaz}(A^-) - G^{\circ}_{gaz}(AH) - G^{\circ}_{gaz}(H_2O)$$
$$+ \Delta G_{solv}(H_3O^+) + \Delta G_{solv}(A^-) - \Delta G_{solv}(AH) - \Delta G_{vap}(H_2O) \quad (5.9)$$

$$pKa = \frac{\Delta G_{aq}}{RTln(10)} - log[H_2O] \quad (5.10)$$

Les valeurs de $G_{solv}(H_3O^+)$ et $G_{solv}(H_2O)$ sont une des limites de ce cycle, en raison de la charge élevée de l'ion hydronium et la grande polarité de H2O. Donc, il est difficile de calculer l'énergie libre de solvatation pour ces deux espèces. Pour cette raison, nous avons utilisé les valeurs expérimentales à l'état standard 1M [24, 25].

$\Delta G_{solv}(H_3O^+)$ = -110,3 kcal/mol.

$\Delta G_{solv}(H_2O)$ = -6,32 kcal/mol.

Une valeur de référence 55.5 mol/l pour la concentration de l'eau à l'état liquide :

$$c = \frac{n}{v} = \frac{m}{M.V}\frac{1000}{18.1} = 55.55mol/l$$

Les valeurs de l'énergie libre de Gibbs en phase gazeuse pour les deux molécules H_3O^+ et H_2O sont calculés avec le même niveau B3LYP/6-311++G(d,p) et les valeurs respectives sont :

$G^{\circ}H_3O^+$ = -76.715053 hartree.

$G^{\circ}H_2O$ = -76.454884 hartree.

Une fois les valeurs de pKa pour toute la série des acides benzoïques (76 composés) basant sur les deux cycles thermodynamiques sont calculées, on passe au traitement statistique et la modélisation QSPR.

5.2.2 Analyse de régression linéaire simple

La modélisation statistique consiste généralement à rechercher une relation approximative entre les variables dépendantes et indépendantes. La méthode la plus simple consiste à utiliser une régression linéaire simple de la forme :

$$Y = a * X + b \quad (5.11)$$

Où a et b sont des constantes.

Cette méthode a été présentée dans la section (2.2) (page 30).

5.2.3 Optimisation de la géométrie et calcul de chimie quantique

Nous avons utilisé le logiciel *ChemBioDraw* [35] pour dessiner la structure moléculaire 2D pour l'ensemble des 76 composés, neutres et anioniques (AH, A^-) respectivement, et le *Chem3D* pour créer les structures 3D. Ces structures ont été ensuite optimisées avec la méthode B3LYP/6-311++G (d, p) [36, 37], implémentée dans le programme *GAUSSIAN 09* [38] sous LINUX. Les calculs de fréquence ont été effectués au même niveau de théorie pour calculer l'énergie libre de Gibbs de déprotonation de l'acide en phase gazeuse (ΔG_{gaz}) qui est basée sur l'équation (5.1, 5.2) Les effets de solvant ont été considérés au moyen du modèle SMD [39] en effectuant des calculs d'optimisation de géométrie pour les structures neutres et ses correspondants anions. Les calculs SMD utilisent le rayon atomique UAHF[i] pour calculer l'énergie libre de Gibbs solvatation ΔG_{solv}.
La régression linéaire simple a été réalisée avec le logiciel MINITAB [40]..

5.3 Résultats et discussions

Nous avons choisi un nombre important d'acides benzoïque dont les valeurs de pKa expérimental étaient disponibles dans le *CRC Handbook of Chemistry and Physics* [41] et la série des travaux de *Pytela et al.* [42–47], *Bosch et al.* [48, 49], *Izutsu* [50], *Kolthoff et al.* [51, 52], et *Kulhánek et al.* [53]. Ces données ont été divisées en série d'apprentissage contient 51 composés (tableau (5.1)) et une série de test contient 25 composés (tableau (5.5)), les valeurs de pKa ont été mesurées à $25\,^{\circ}\mathrm{C}$ et en solution aqueuse. Le tableau (5.1) rassemble la nomenclature des acides benzoïques étudiés avec ces valeurs expérimentales de pKa [41–53],
La figure (5.1) représente les structures 2D des acides ainsi que leurs substitutions (*H, Cl, F, Br, CH3, NH2, OH, Ph, ...*),

La liste se compose des acides avec des valeurs de pKa comprises entre 1.22 à 5.10 pKa unités, les acides ont été classés dans l'ordre croissant des valeurs de pKa, c'est-à-dire une augmentation de pKa correspond donc bien à une diminution du caractère acide de l'acide et donc, à une augmentation du caractère basique de la

i. United Atom Topological Model applied on radii optimized for the HF/6-31G(d) level of theory, Dans ce modèle les atomes d'Hydrogène sont inclus dans la sphère de l'atome auquel ils sont liés. Le rayon de la sphère dépend du numéro atomique, de la charge et de l'hybridation de l'atome

base conjuguée.

51 acides benzoïques ont été inclus dans ce travail, pour étudier l'influence des groupes aliphatiques et la position du radical sur l'acidité des acides.

Le tableau (5.1) présente les énergies libres de Gibbs de déprotonation ΔG_{aq} en solution calculées pour les 51 acides avec les deux cycles thermodynamiques (Modèle A, Modèle B).
Les équations de régressions linéaires simples obtenues avec MINITAB sont comme suit :

Modèle A : Représente le modèle de corrélation linéaire simple obtenu entre les valeurs de pKa-pred (calculées à partir de l'équation (5.8) du cycle A) et les valeurs de pKa expérimentales.

$$pKa_{pre} = 0,049 + 1,026 \;\; x \;\; pKa_{exp} \tag{5.12}$$

R = 0,967, SD = 0,034, $R^2 = 0.936$, R^2_{CV} = 0,934 F = 717,530, N = 51 composés.

Modèle B : Représente le modèle de corrélation linéaire simple obtenu entre les valeurs de pKa-pred (calculées à partir de l'équation (5.10) du cycle B) et les valeurs de pKa expérimentales.

$$pKa_{pre} = -0,08215 + 1,02501 \;\; x \;\; pKa_{exp} \tag{5.13}$$

R = 0,967, SD =0.035, $R^2 = 0.936$, R^2_{CV} = 0,934 F = 717,459, N = 51 composés

Modèle C : Représente le modèle de corrélation linéaire simple obtenu entre les valeurs de pKa expérimentales et les valeurs de ΔG_{aq} (calculées à partir de l'équation (5.4) du cycle A). Cette corrélation nous permet de prédire les valeurs de pKa en utilisant l'équation suivante :

$$pKa_{pre} = -3,08997 * 10^{-5} + 1 \;\; x \;\; \Delta G_{aq} \tag{5.14}$$

R = 0,967, SD =0.035, $R^2 = 0.936$, R^2_{CV} = 0,934 F = 719,932, N = 51 composés

Modèle D : Représente le modèle de corrélation linéaire simple obtenu entre les valeurs de pKa expérimentales et les valeurs de ΔG_{aq} (calculées à partir de l'équation

(5.9) du cycle B). Cette corrélation nous permet de prédire les valeurs de pKa en utilisant l'équation suivante :

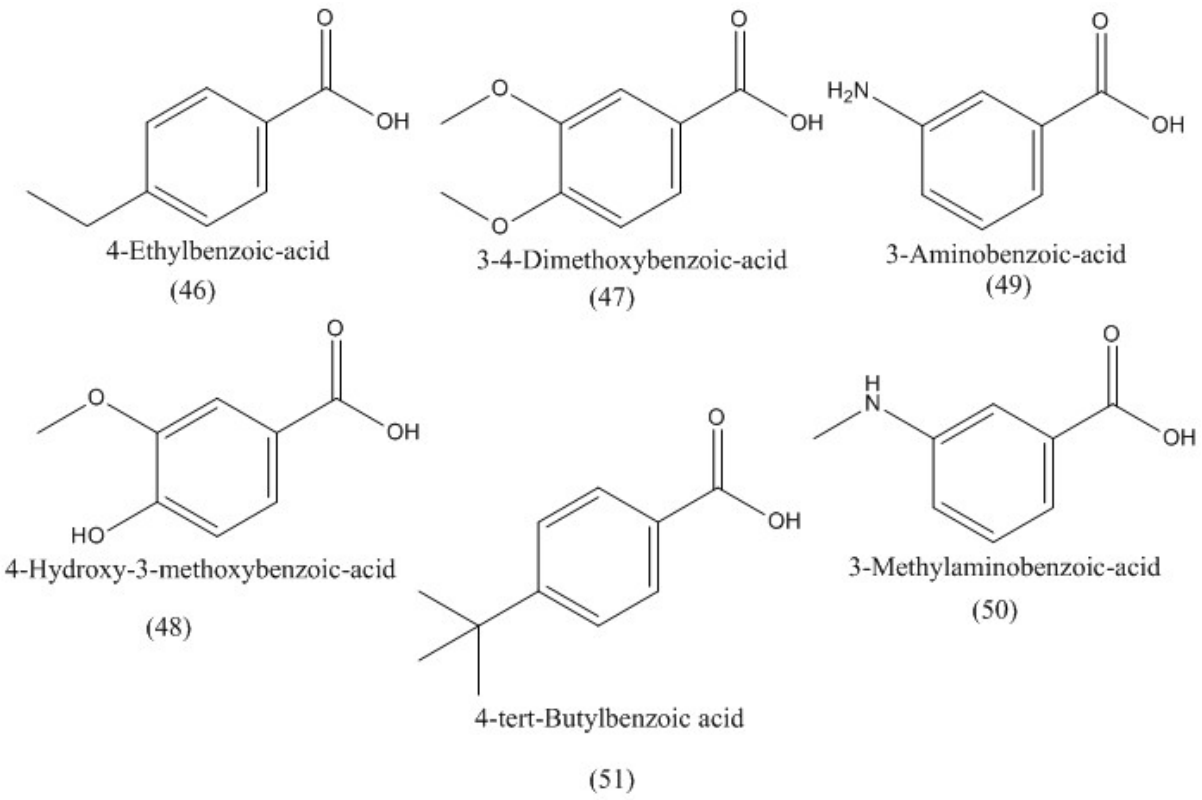
$$pKa_{pre} = -1,21469 * 10^{-6} + 0.668 \;\; x \;\; \Delta G_{aq} \tag{5.15}$$

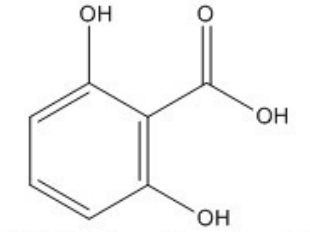
R = 0,967, SD =0.037, $R^2 = 0.936$, R^2_{CV}= 0,934 F = , N = 51 composés

Remarque

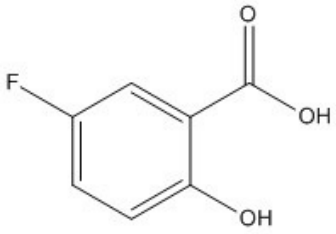
Les paramètres statistiques sont très proches pour les 4 modèles ; ce qui montre une relation réciproque entre les deux cycles thermodynamiques A et B. Pour cette raison, nous nous sommes limités à la discussion et l'analyse des résultats obtenus avec le modèle C.

FIGURE *5.1: Structures 2D des 51 composés (série d'apprentissage)*

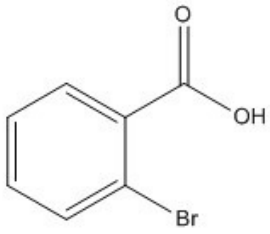




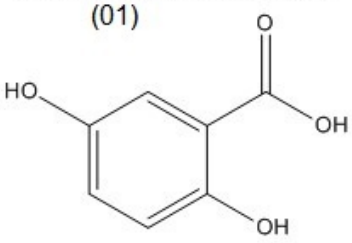
2-6-Dihydroxybenzoic-acid
(01)



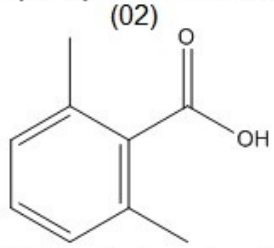
2-Hydroxy-5-fluorobenzoic-acid
(02)



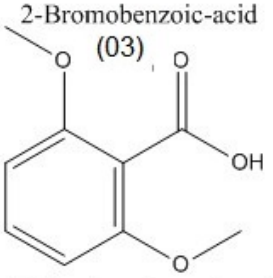
2-Bromobenzoic-acid
(03)



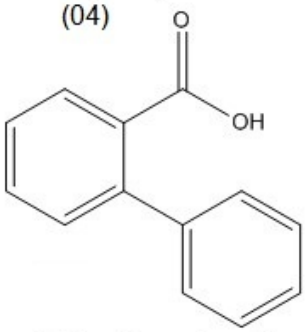
2-5-Dihydroxybenzoic-acid
(04)



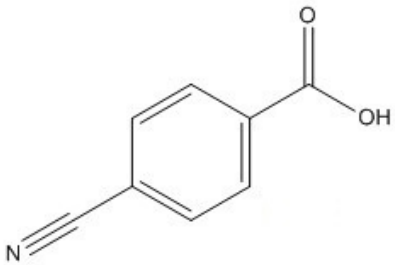
2-6-Dimethylbenzoic-acid
(05)



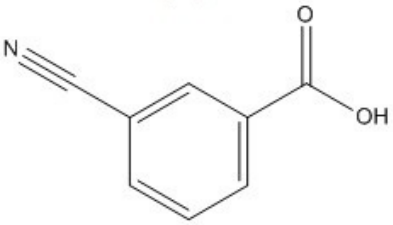
2-6-Dimethoxybenzoic-acid
(06)



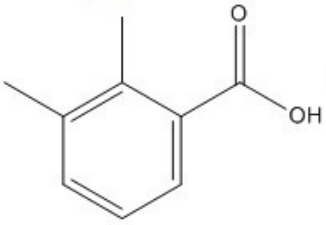
2-Phenylbenzoic-acid
(07)



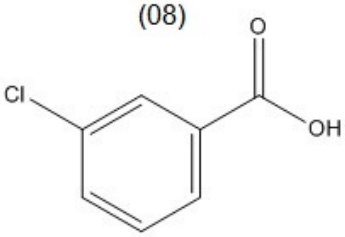
4-Cyanobenzoic-acid
(08)



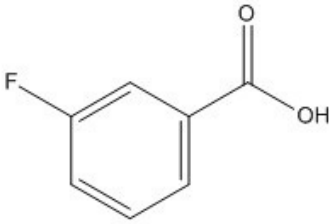
3-Cyanobenzoic-acid
(09)



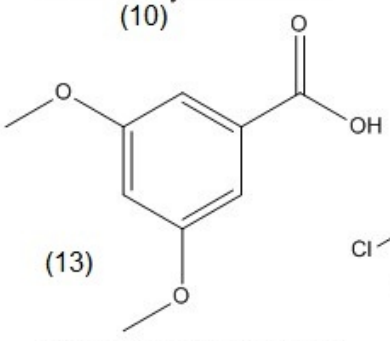
2-3-Dimethylbenzoic-acid
(10)



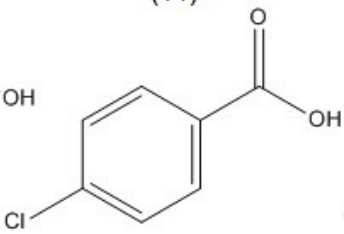
3-Chlorobenzoic-acid
(11)



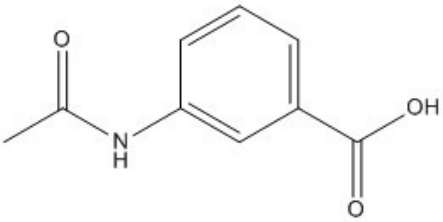
3-Fluorobenzoic-acid
(12)



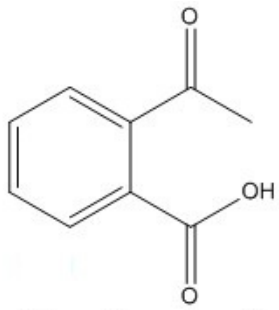
(13)
3-5-Dimethoxybenzoic-acid



4-Chlorobenzoic-acid
(14)

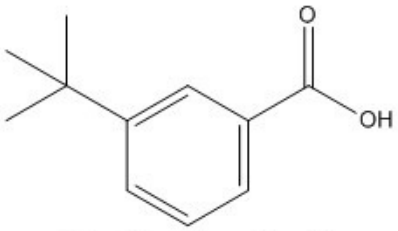


(15) 3-Acetamidobenzoic-acid



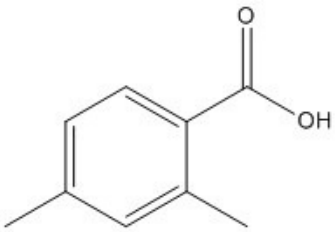
2-Acetylbenzoic-acid

(16)



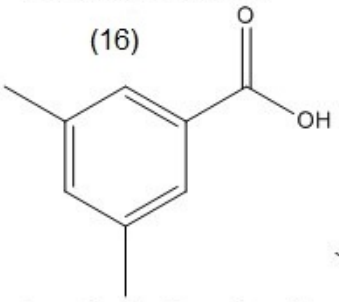
3-tert-Butylbenzoic-acid

(17)



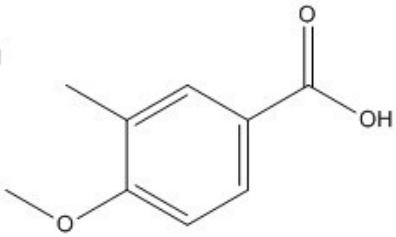
2-4-Dimethylbenzoic-acid

(18)



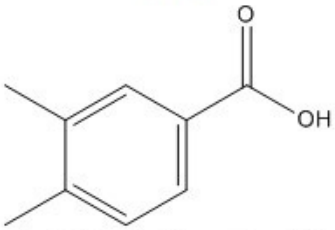
3-5-Dimethylbenzoic-acid

(19)



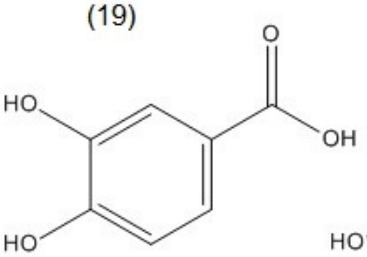
3-Methyl-4-methoxybenzoic-acid

(20)



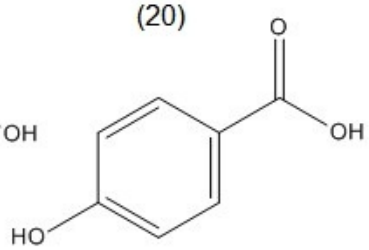
3-4-Dimethylbenzoic-acid

(21)



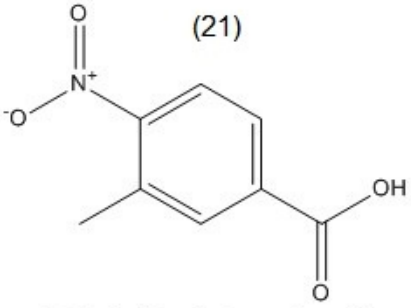
3-4-Dihydroxybenzoic-acid

(22)



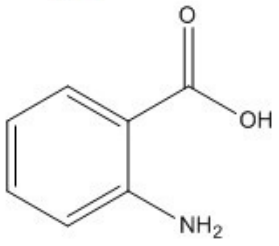
4-Hydroxybenzoic-acid

(23)



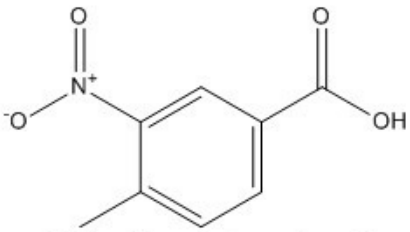
3-Methyl-4-nitrobenzoic acid

(24)



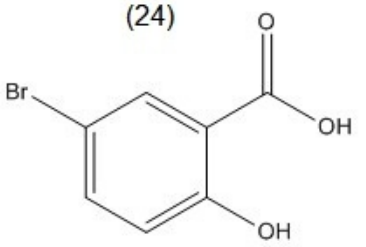
2-Aminobenzoic-acid

(25)



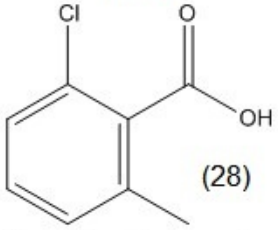
3-Nitro-4-methylbenzoic acid

(26)



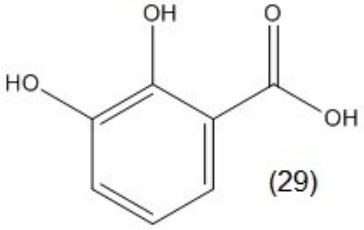
2-Hydroxy-5-bromobenzoic-acid

(27)



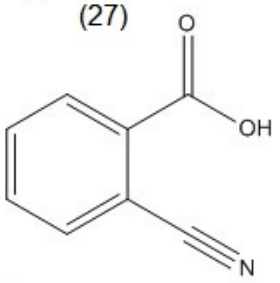
(28)

2-Chloro-6-methylbenzoic-acid

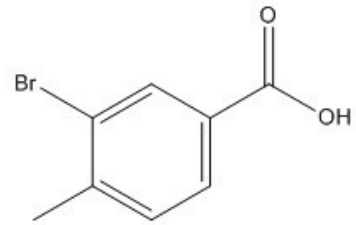


(29)

2-3-Dihydroxybenzoic-acid

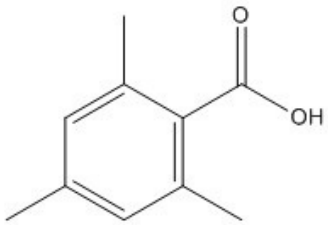


(30) 2-Cyanobenzoic-acid



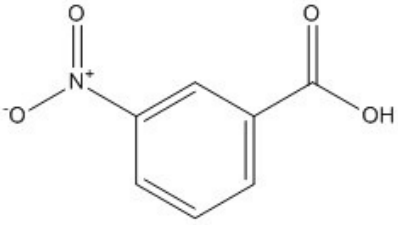
3-Bromo-4-methylbenzoic-acid

(31)



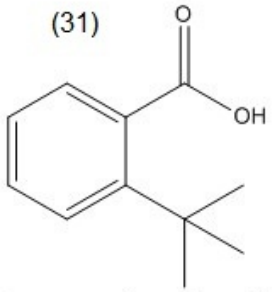
2-4-6-Trimethylbenzoic-acid

(32)



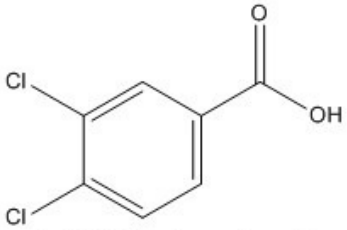
3-Nitrobenzoic-acid

(33)



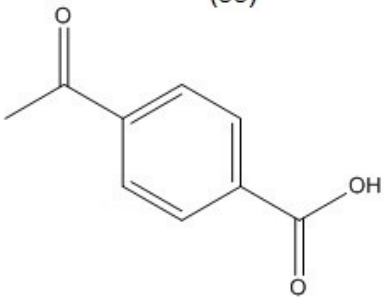
2-tert-Butylbenzoic-acid

(34)



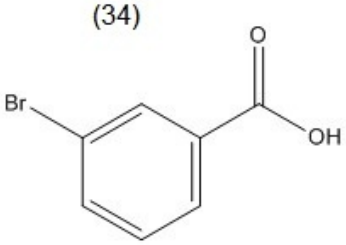
3-4-Dichlorobenzoic-acid

(35)



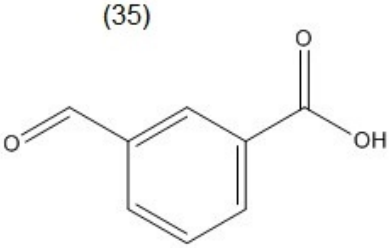
4-Acetylbenzoic-acid

(36)



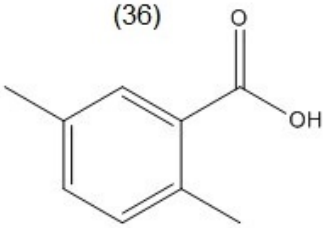
3-Bromobenzoic-acid

(37)



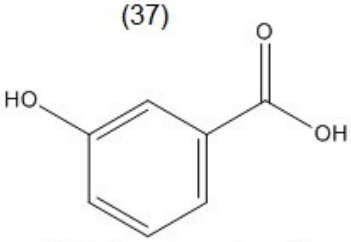
3-Formylbenzoic-acid

(38)



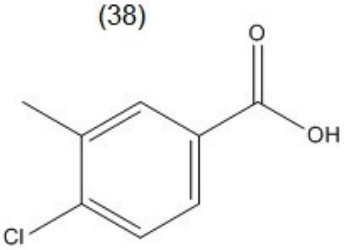
2-5-Dimethylbenzoic-acid

(39)



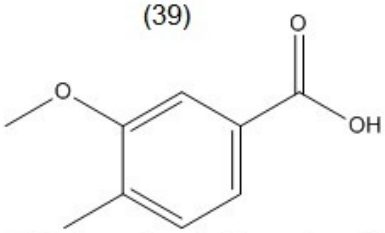
3-Hydroxybenzoic-acid

(40)



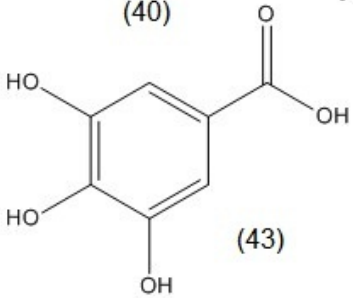
3-Methyl-4-chlorobenzoic-acid

(41)



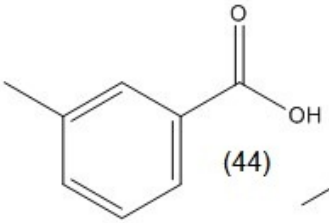
3-Methoxy-4-methylbenzoic-acid

(42)



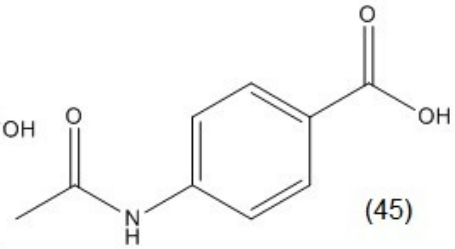
(43)

3-4-5-trihydroxybenzoic-acid



(44)

3-Methylbenzoic-acid



(45)

4-Acetamidobenzoic-acid

TABLE *5.1: Valeurs expérimentales et calculées pKa (modèles A, B, C et D)*

		L'ensemble d'apprentissage 51 composés	pKa-exp	ΔG_{aq}, Eq. (5.4), cycle A	pKa-pre Modèle A Eq. (5.12)	ΔG_{aq}, Eq. (5.9), cycle B	pKa-pre Modèle B Eq. (5.13)	pKa-pre Modele C Eq. (5.14)	pKa-pre Modèle D Eq. (5.15)
1	A	2-6-Dihydroxybenzoic-acid	1.22	1.81	1.32	4.85	1.19	1.408	1.406
2	B	2-Hydroxy-5-fluorobenzoic-acid	2.70	3.91	2.86	6.95	2.73	2.811	2.809
3	C	2-Bromobenzoic-acid	2.85	3.95	2.90	7.00	2.76	2.838	2.842
4	A	2-5-Dihydroxybenzoic-acid	2.95	4.50	3.30	7.55	3.16	3.206	3.210
5	B	2-6-Dimethylbenzoic-acid	3.25	4.75	3.48	7.80	3.35	3.373	3.377
6	C	2-6-Dimethoxybenzoic-acid	3.44	5.10	3.74	8.14	3.60	3.607	3.604
7	A	2-Phenylbenzoic-acid	3.46	4.78	3.50	7.83	3.37	3.393	3.397
8	B	4-Cyanobenzoic-acid	3.53	4.55	3.33	7.59	3.20	3.239	3.236
9	C	3-Cyanobenzoic-acid	3.60	4.73	3.47	7.78	3.33	3.360	3.363
10	A	2-3-Dimethylbenzoic-acid	3.74	5.39	3.95	8.43	3.81	3.801	3.797
11	B	3-Chlorobenzoic-acid	3.80	5.19	3.80	8.24	3.67	3.667	3.671
12	C	3-Fluorobenzoic-acid	3.88	5.33	3.90	8.37	3.77	3.761	3.757
13	A	3-5-Dimethoxybenzoic-acid	3.97	5.56	4.08	8.61	3.94	3.914	3.918
14	B	4-Chlorobenzoic-acid	4.00	5.63	4.12	8.67	3.99	3.961	3.958
15	C	3-Acetamidobenzoic-acid	4.06	5.58	4.09	8.62	3.95	3.928	3.924
16	A	2-Acetylbenzoic-acid	4.13	5.83	4.27	8.87	4.13	4.095	4.091
17	B	3-tert-Butylbenzoic-acid	4.20	6.17	4.52	9.21	4.38	4.322	4.319
18	C	2-4-Dimethylbenzoic-acid	4.22	6.01	4.41	9.06	4.27	4.215	4.218
19	A	3-5-Dimethylbenzoic-acid	4.30	6.30	4.62	9.35	4.48	4.409	4.412
20	B	3-Methyl-4-methoxybenzoic-acid	4.35	6.52	4.78	9.57	4.64	4.556	4.559
21	C	3-4-Dimethylbenzoic-acid	4.40	6.55	4.80	9.59	4.66	4.576	4.572
22	A	3-4-Dihydroxybenzoic-acid	4.48	6.67	4.89	9.71	4.75	4.656	4.653

23	B	4-Hydroxybenzoic-acid	4.55	6.56	4.81	9.60	4.67	4.583	4.579
24	C	2-Aminobenzoic-acid	4.87	6.52	4.78	9.57	4.64	4.556	4.559
25	A	3-Methyl-4-nitrobenzoic acid	3.65	5.30	3.89	8.35	3.75	3.741	3.744
26	B	3-Nitro-4-methylbenzoic acid	3.62	4.70	3.44	7.74	3.31	3.340	3.337
27	C	2-Hydroxy-5-bromobenzoic-acid	2.54	3.89	2.85	6.93	2.72	2.798	2.795
28	A	2-Chloro-6-methylbenzoic-acid	2.75	3.61	2.65	6.65	2.51	2.611	2.608
29	B	2-3-Dihydroxybenzoic-acid	2.91	4.59	3.37	7.64	3.23	3.266	3.270
30	C	2-Cyanobenzoic-acid	3.08	4.09	3.00	7.13	2.86	2.932	2.929
31	A	3-Bromo-4-methylbenzoic-acid	3.29	5.02	3.68	8.07	3.54	3.553	3.557
32	B	2-4-6-Trimethylbenzoic-acid	3.45	4.81	3.53	7.86	3.39	3.413	3.417
33	C	3-Nitrobenzoic-acid	3.46	4.41	3.23	7.45	3.10	3.146	3.143
34	A	2-tert-Butylbenzoic-acid	3.54	4.91	3.60	7.96	3.47	3.480	3.484
35	B	3-4-Dichlorobenzoic-acid	3.64	5.56	4.08	8.61	3.94	3.914	3.918
36	C	4-Acetylbenzoic-acid	3.74	5.17	3.79	8.22	3.65	3.654	3.657
37	A	3-Bromobenzoic-acid	3.81	4.95	3.63	7.99	3.49	3.507	3.504
38	B	3-Formylbenzoic-acid	3.88	5.29	3.88	8.34	3.74	3.734	3.737
39	C	2-5-Dimethylbenzoic-acid	3.98	5.62	4.12	8.66	3.98	3.955	3.951
40	A	3-Hydroxybenzoic-acid	4.03	5.82	4.26	8.86	4.13	4.088	4.085
41	B	3-Methyl-4-chlorobenzoic-acid	4.07	5.83	4.27	8.87	4.13	4.095	4.091
42	C	3-Methoxy-4-methylbenzoic-acid	4.13	6.05	4.44	9.10	4.30	4.242	4.245
43	A	3-4-5-Trihydoxybenzoic-acid	4.21	6.04	4.42	9.08	4.29	4.235	4.232
44	B	3-Methylbenzoic-acid	4.24	6.05	4.43	9.10	4.30	4.242	4.245
45	C	4-Acetamidobenzoic-acid	4.30	6.17	4.52	9.22	4.39	4.322	4.325
46	A	4-Ethylbenzoic-acid	4.35	6.31	4.63	9.36	4.49	4.416	4.419
47	B	3-4-Dimethoxybenzoic-acid	4.44	6.50	4.76	9.54	4.62	4.543	4.539
48	C	4-Hydroxy-3-methoxybenzoic-acid	4.51	6.67	4.89	9.72	4.75	4.656	4.659
49	A	3-Aminobenzoic-acid	4.74	6.50	4.76	9.54	4.62	4.543	4.539
50	B	3-Methylaminobenzoic-acid	5.10	6.82	5.00	9.86	4.86	4.757	4.753
51	C	4-tert-Butylbenzoic acid	4.36	6.22	4.56	9.27	4.42	4.356	4.359

Discussion et analyse du modèle (C) Eq. (5.14)

Validation interne :

Nous avons utilisé deux méthodes pour la validation interne :

i) En utilisant la méthode de validation interne croisée (leave one out), on obtient $R^2_{CV} = 0.934$ cette valeur est proche de $R^2 = 0.934$, ce qui montre la stabilité interne du modèle obtenu.

ii) Afin de vérifier la fiabilité et la stabilité du meilleur modèle QSPR (Eq. (5.14)), nous avons utilisé également la validation interne « leave-1/3-of-set-out » de la manière suivante : les valeurs de données expérimentales parentes ont été divisées en fonction des valeurs expérimentales dans trois sous-ensembles (1er, 4ème, 7ème,..., etc points forment le premier sous-ensemble A, le 2ème, 5ème, 8ème,..., etc points forment le second sous-ensemble (B), et le 3ème, 6ème, 9ème,..., etc points forment le troisième sous-ensemble (C). En combinant deux des sous-ensembles A, B , C, on obtient trois combinaisons et l'équation de corrélation est dérivée avec les mêmes descripteurs. L'équation obtenue a été utilisée pour prédire les données pour le sous-ensemble restant. Il s'avère que les valeurs prédites en utilisant R^2 pour les sous-ensembles (A + B), (B + C), (A + C) sont très proches de celles correspondant à l'ensemble complet de la série d'apprentissage (A + B + C) et les valeurs moyennes de $R^2_{(Fit)}$ et $R^2_{(Predites)}$ (voir Tableau 5.2) sont également très proches. Notons que la valeur R^2_{CV} des modèles correspondant à des sous-ensembles A + B, A + C, et B + C sont beaucoup plus grandes que 0.90, ce qui indique que notre modèle est stable et peut être efficacement utilisée pour estimer les valeurs de pKa des autres dérivés des acides benzoïques pour lesquels les données expérimentales ne sont pas disponibles.

TABLE *5.2: Validation interne du modèle C (l'équation(5.14))*

Série d'apprentissage	N	$R^2_{(fit)}$	$R^2_{cv}(fit)$	SD(Fit)	Série de test	N	$R^2_{(pre)}$	$R^2_{cv}(pre)$
A + B	34	0.941	0.939	0.14	C	17	0.927	0.923
A + C	34	0.950	0.948	0.03	B	17	0.894	0.887
B + C	34	0.910	0.907	0.41	A	17	0.965	0.962
average		0.934	0.931	0.19			0.929	0.924

Par ailleurs, pour s'assurer que notre modèle C est fiable. La technique appelée Y-

randomisation a été utilisée. Pour évaluer la part de chance dans le modèle construit, la procédure consiste à mélanger aléatoirement les valeurs de pKa dans le modèle initial et à recréer de nouveaux modèles à calculer chaque fois les performances statistiques.

Dans ce travail, les valeurs de pKa ont été randomisées dans l'ensemble d'apprentissage dans 500 itérations successives (Tableau 5.3) et dans chaque nouveau jeu de données aléatoire on trouve que la valeur moyenne des coefficients R^2 et le coefficient de cross-validation pour les modèles aléatoires (notés $R^2_{(ran)}$ et $R^2_{(cv-ran)}$, respectivement) en dessous de 0,10 pour $R^2_{(ran)}$ et 0,05 de $R^2_{(cv-ran)}$(Tableau 5.4).

TABLE *5.3: Procédure de Y-randomisation dans le modèle C (15 itérations)*

Comp	ΔG_{aq}	Y-1	Y-2	Y-3	Y-4	Y-5	Y-6	Y-7	Y-8	Y-9	Y-10	Y-11	Y-12
1	1.81	4.13	3.97	4.30	3.25	1.22	4.13	3.97	4.30	3.25	5.10	4.74	3.64
2	3.91	3.54	2.75	3.88	4.87	4.30	3.64	3.25	4.51	4.40	1.22	4.13	3.97
3	3.95	4.24	4.35	3.29	3.60	5.10	4.74	3.64	4.48	3.74	3.64	4.40	3.29
4	4.50	5.10	4.20	3.88	3.74	4.36	4.51	2.95	3.62	4.21	4.07	3.88	4.74
5	4.75	3.45	3.98	4.36	2.95	3.74	4.30	5.10	1.22	3.97	2.70	3.54	2.75
6	5.10	3.46	3.45	3.64	3.29	4.20	3.65	3.46	3.81	3.44	2.54	3.98	4.24
7	4.78	3.81	4.22	4.44	4.13	3.65	3.46	3.65	3.65	3.65	4.74	4.13	3.53
8	4.55	4.03	1.22	4.22	3.81	4.03	3.74	4.30	4.13	4.51	3.81	4.22	3.44
9	4.73	4.21	2.91	3.45	3.88	2.95	5.10	4.20	3.88	3.74	3.46	3.81	4.22
10	5.39	4.30	5.10	1.22	3.97	4.06	4.30	4.48	4.40	4.00	2.95	5.10	4.20
11	5.19	2.91	2.85	4.13	4.07	2.54	3.98	4.24	4.30	3.08	4.22	2.54	3.81
12	5.33	2.70	4.87	2.70	2.75	3.64	4.40	3.29	4.87	3.62	4.87	4.44	3.62
13	5.56	3.60	4.06	3.25	3.98	3.74	4.36	4.03	4.24	2.70	3.80	2.91	2.85
14	5.63	3.44	4.13	3.74	5.10	3.53	4.03	1.22	4.22	3.81	3.44	3.46	3.45
15	5.58	4.30	4.48	4.40	4.00	4.74	4.13	3.53	3.80	2.85	3.98	2.95	3.54
16	5.83	3.29	4.07	3.60	2.91	4.48	4.55	3.74	4.21	3.88	4.03	3.74	4.30
17	6.17	3.65	3.46	3.81	3.44	4.21	4.06	3.88	4.20	3.46	3.25	3.45	3.98
18	6.01	2.54	3.81	3.53	1.22	3.46	3.88	4.36	4.06	4.48	4.00	3.44	4.13
19	6.30	4.00	3.08	3.97	4.06	2.75	4.48	3.80	4.35	4.55	4.51	3.53	4.40
20	6.52	3.74	4.30	4.00	4.13	4.30	4.00	3.08	3.97	4.06	4.21	4.06	3.88
21	6.55	4.87	4.00	5.10	3.64	4.35	3.74	4.30	4.00	4.13	3.74	4.36	4.03
22	6.67	4.55	3.74	4.21	3.88	4.51	3.53	4.40	4.03	4.30	3.65	3.46	3.65
23	6.56	3.25	2.54	4.07	4.74	4.13	3.29	4.07	3.60	2.91	4.30	4.00	3.08
24	6.52	4.44	3.62	2.85	4.35	3.88	2.70	4.87	2.70	2.75	4.06	4.30	4.48
25	5.30	3.46	3.65	3.65	3.65	2.91	4.07	4.35	3.08	3.46	4.48	4.55	3.74
26	4.70	1.22	4.21	2.95	4.20	4.35	4.35	4.55	4.35	4.30	3.62	1.22	4.21
27	3.89	3.98	4.24	4.30	3.08	3.98	2.95	3.54	2.54	4.24	4.13	3.97	4.51
28	3.61	4.48	3.80	4.35	4.55	4.55	3.25	2.54	4.07	4.74	3.74	4.30	5.10
29	4.59	4.07	4.35	3.08	3.46	4.87	4.44	3.62	2.85	4.35	3.08	2.85	3.46

Comp	ΔG_{aq}	Y-1	Y-2	Y-3	Y-4	Y-5	Y-6	Y-7	Y-8	Y-9	Y-10	Y-11	Y-12
30	4.09	2.85	3.46	3.44	3.45	2.70	3.54	2.75	3.88	4.87	4.13	3.29	4.07
31	5.02	4.20	3.60	3.74	4.03	3.08	2.85	3.46	3.44	3.45	3.88	2.70	4.87
32	4.81	3.80	3.88	3.46	4.22	3.46	3.81	4.22	4.44	4.13	3.46	3.88	4.36
33	4.41	3.88	4.36	4.06	4.48	3.81	4.22	3.44	2.75	3.80	3.53	4.03	1.22
34	4.91	3.08	4.44	3.54	4.44	4.07	3.88	4.74	4.55	2.54	4.24	2.75	2.70
35	5.56	4.40	3.29	4.87	3.62	3.25	3.45	3.98	4.36	2.95	2.91	4.07	4.35
36	5.17	4.36	4.03	4.24	2.70	2.85	4.24	4.35	3.29	3.60	2.75	4.48	3.80
37	4.95	4.22	3.44	2.75	3.80	3.44	3.46	3.45	3.64	3.29	4.55	3.25	2.54
38	5.29	3.62	3.74	2.91	4.35	4.22	2.54	3.81	3.53	1.22	3.88	3.62	3.74
39	5.62	2.95	3.54	2.54	4.24	4.44	4.35	4.13	3.98	3.54	3.97	3.60	4.06
40	5.82	3.74	4.30	4.13	4.51	4.40	4.87	4.00	5.10	3.64	4.20	3.65	3.46
41	5.83	3.88	4.74	4.55	2.54	3.80	2.91	2.85	4.13	4.07	4.40	4.87	4.00
42	6.05	3.97	4.51	4.74	3.53	3.29	4.20	3.60	3.74	4.03	3.60	4.21	2.91
43	6.04	4.06	3.88	4.20	3.46	4.13	3.97	4.51	4.74	3.53	4.30	3.64	3.25
44	6.05	2.75	2.70	3.46	4.36	3.97	3.60	4.06	3.25	3.98	4.35	4.35	4.55
45	6.17	3.64	3.25	4.51	4.40	3.54	3.08	4.44	3.54	4.44	4.36	4.51	2.95
46	6.31	4.35	4.55	4.35	4.30	4.00	3.44	4.13	3.74	5.10	3.54	3.08	4.44
47	6.50	4.35	4.13	3.98	3.54	3.45	3.80	3.88	3.46	4.22	3.45	3.80	3.88
48	6.67	3.53	4.40	4.03	4.30	4.24	2.75	2.70	3.46	4.36	4.44	4.35	4.13
49	6.50	4.13	3.53	3.80	2.85	3.62	1.22	4.21	2.95	4.20	4.35	3.74	4.30
50	6.82	4.74	3.64	4.48	3.74	3.60	4.21	2.91	3.45	3.88	3.29	4.20	3.60
51	6.22	4.51	2.95	3.62	4.21	3.88	3.62	3.74	2.91	4.35	2.85	4.24	4.35

Comp : Composée numéro x,

Y-x : Les valeurs de pKa randomisées dans l'ensemble d'apprentissage

TABLE *5.4: Les valeurs de R^2 et R^2_{cv} obtenues dans la procedure Y-Randomisation (332 itérations)*

	$R^2_{(ran)}$	$R^2_{(cv-ran)}$		$R^2_{(ran)}$	$R^2_{(cv-ran)}$		$R^2_{(ran)}$	$R^2_{(cv-ran)}$
Rand-1	0.0100	0.0401	Rand-31	0.0084	0.0266	Rand-61	0.0070	0.0199
Rand-2	0.0099	0.0395	Rand-32	0.0084	0.0260	Rand-62	0.0069	0.0194
Rand-3	0.0098	0.0386	Rand-33	0.0084	0.0260	Rand-63	0.0069	0.0191
Rand-4	0.0098	0.0382	Rand-34	0.0084	0.0260	Rand-64	0.0069	0.0190
Rand-5	0.0097	0.0378	Rand-35	0.0083	0.0256	Rand-65	0.0069	0.0188
Rand-6	0.0097	0.0365	Rand-36	0.0082	0.0256	Rand-66	0.0069	0.0186
Rand-7	0.0097	0.0364	Rand-37	0.0082	0.0255	Rand-67	0.0069	0.0185
Rand-8	0.0097	0.0363	Rand-38	0.0082	0.0250	Rand-68	0.0069	0.0181
Rand-9	0.0095	0.0362	Rand-39	0.0081	0.0247	Rand-69	0.0068	0.0179
Rand-10	0.0094	0.0360	Rand-40	0.0081	0.0246	Rand-70	0.0068	0.0178
Rand-11	0.0093	0.0359	Rand-41	0.0080	0.0241	Rand-71	0.0067	0.0177
Rand-12	0.0093	0.0348	Rand-42	0.0080	0.0237	Rand-72	0.0066	0.0171
Rand-13	0.0092	0.0343	Rand-43	0.0080	0.0234	Rand-73	0.0066	0.0171
Rand-14	0.0092	0.0333	Rand-44	0.0078	0.0233	Rand-74	0.0065	0.0167
Rand-15	0.0091	0.0333	Rand-45	0.0078	0.0230	Rand-75	0.0065	0.0167
Rand-16	0.0091	0.0333	Rand-46	0.0078	0.0228	Rand-76	0.0064	0.0166
Rand-17	0.0090	0.0331	Rand-47	0.0077	0.0224	Rand-77	0.0064	0.0165
Rand-18	0.0089	0.0326	Rand-48	0.0077	0.0221	Rand-78	0.0064	0.0163
Rand-19	0.0089	0.0325	Rand-49	0.0077	0.0217	Rand-79	0.0063	0.0160
Rand-20	0.0088	0.0319	Rand-50	0.0076	0.0214	Rand-80	0.0063	0.0159
Rand-21	0.0088	0.0311	Rand-51	0.0075	0.0214	Rand-81	0.0063	0.0156
Rand-22	0.0087	0.0305	Rand-52	0.0075	0.0214	Rand-82	0.0062	0.0151
Rand-23	0.0086	0.0304	Rand-53	0.0074	0.0211	Rand-83	0.0062	0.0150
Rand-24	0.0086	0.0298	Rand-54	0.0074	0.0210	Rand-84	0.0061	0.0147
Rand-25	0.0086	0.0297	Rand-55	0.0074	0.0205	Rand-85	0.0061	0.0146
Rand-26	0.0086	0.0293	Rand-56	0.0073	0.0204	Rand-86	0.0060	0.0146
Rand-27	0.0085	0.0276	Rand-57	0.0073	0.0203	Rand-87	0.0060	0.0144
Rand-28	0.0085	0.0275	Rand-58	0.0072	0.0202	Rand-88	0.0060	0.0144
Rand-29	0.0084	0.0274	Rand-59	0.0071	0.0201	Rand-89	0.0060	0.0141
Rand-30	0.0084	0.0269	Rand-60	0.0071	0.0200	Rand-90	0.0059	0.0141

Rand-x : Randomisation numéro x

Validation externe :

D'une manière générale, les techniques de validation interne permettent l'évaluation de la robustesse du modèle QSPR, bien qu'elles soient importantes et nécessaires mais pas suffisantes. Donc, la validation externe doit être effectuée pour une évaluation plus forte du modèle [30] la validité externe concerne la généralisation des résultats précédents.
L'équation de régression linéaire est de la forme :

$$pKa_{pre} = -3,08997 * 10^{-5} + 1 \;\; x \;\; \Delta G_{aq}$$

R = 0,967, SD =0.035, $R^2 = 0.936$, R^2_{CV} = 0,934 F = 719,932, N = 51 composés

La méthode de validation externe a été appliquée sur un jeu de test de 25 composés (figure 5.2 et tableau 5.5). En utilisant l'équation de régression linéaire simple (Eq. 5.14), nous avons calculé les valeurs de pKa de ces composés et un modèle linéaire a été obtenu

$$pKa_{pre} = 0.22 + 0.935 \;\; x \;\; \Delta G_{aq}$$

$$R = 0.958 \;\; ,sd = 0.027, R^2 = 0.917 \;\; R^2_{cv} = 0.915 \;\; F = 260,090 \;\; N = 25 composs.$$

Vérification des critères de Tropsha pour l'équation (5.14) (voir section (1.4.3) page 24). [29–32]

$$R^2 = 0.936 \qquad (Seuil \;\; R^2 > 0.7) \;\; R^2 pour \;\; la \;\; serie \;\; d'apprentissage \tag{5.16}$$

$$R^2_{cv} = 0.934 \qquad (Seuil \;\; R^2_{cv} > 0.6) R^2_{cv} \;\; pour \;\; la \;\; serie \;\; d'apprentissage \tag{5.17}$$

$$|R^2_{pre} - R^2_0| = 0.01 \qquad (Seuil \;\; |R^2_{pre} - R^2_0| < 0.3) \tag{5.18}$$

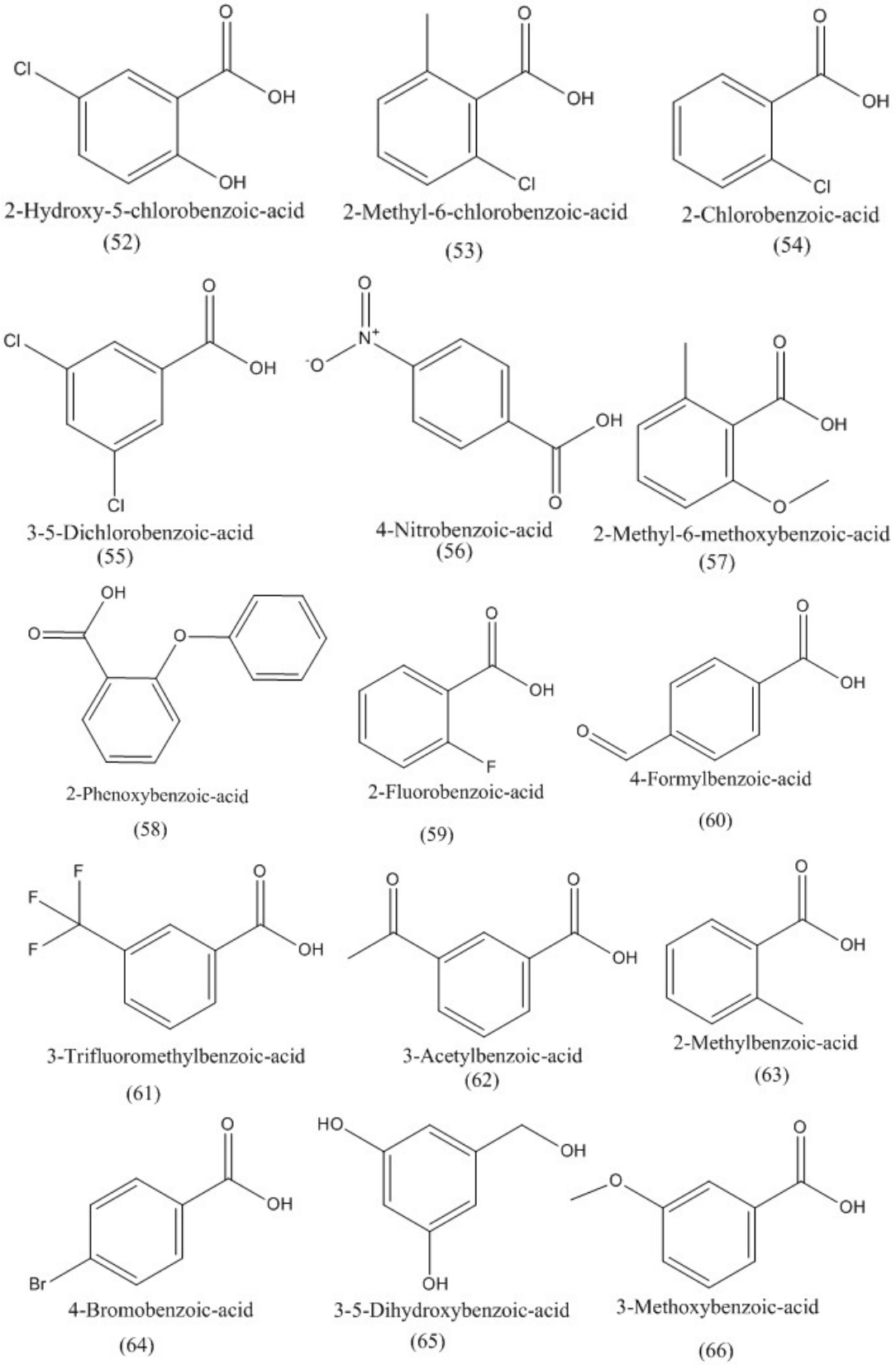
$$\frac{R^2_{pre} - R^2_0}{R^2_{pre}} = 0.02 \qquad (Seuil \;\; \frac{R^2_{pre} - R^2_0}{R^2_{pre}} < 0.1) \tag{5.19}$$

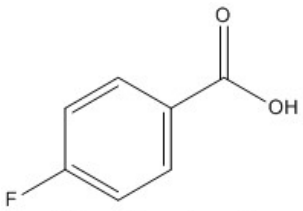
$$R^2_{pre} \;\; pour \;\; la \;\; serie \;\; d'apprentissage \;\; R^2_0 \;\; pour \;\; la \;\; serie \;\; de \;\; test. \tag{5.20}$$

$$k = 1, \;\; k' = 1 \qquad (Seuil \;\; 0.85 \leq k \leq 1.15 \;\; et \;\; 0.85 \leq k' \leq 1.15) \tag{5.21}$$

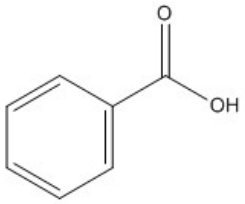
On remarque que tous les critères de Tropsha sont vérifiés.

FIGURE *5.2: Structures 2D des 25 composés (série de test)*

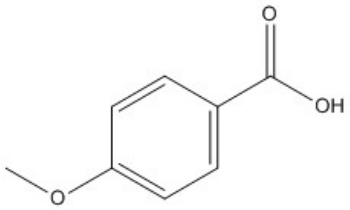




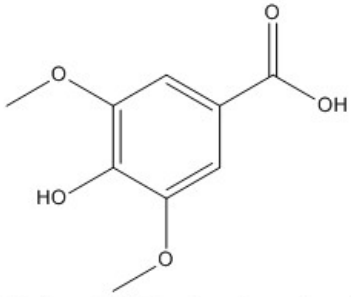
4-Fluorobenzoic-acid
(67)



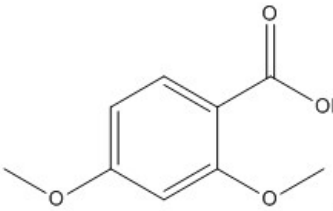
Benzoic-acid
(68)



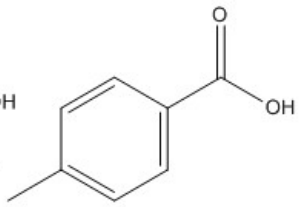
4-Methoxybenzoic-acid
(69)



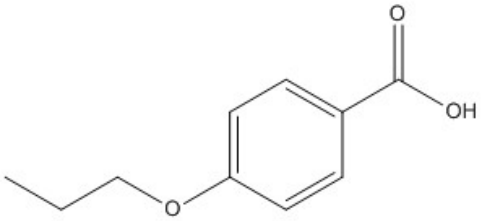
4-Hydroxy-3-5-dimethoxybenzoic-acid
(70)



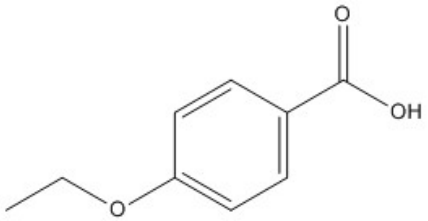
2-4-Dimethoxybenzoic-acid
(71)



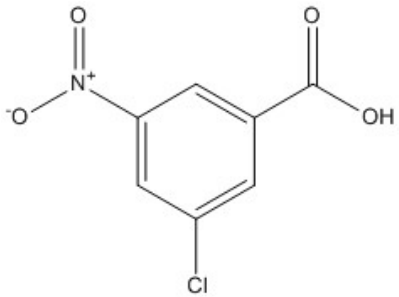
4-Methylbenzoic-acid
(72)



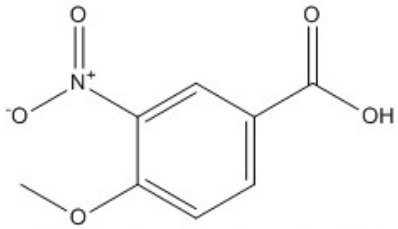
4-Propoxybenzoic-acid
(73)



4-Ethoxybenzoic-acid
(74)



3-Chloro-5-nitrobenzoic acid
(75)



3-Nitro-4-methoxybenzoic acid
(76)

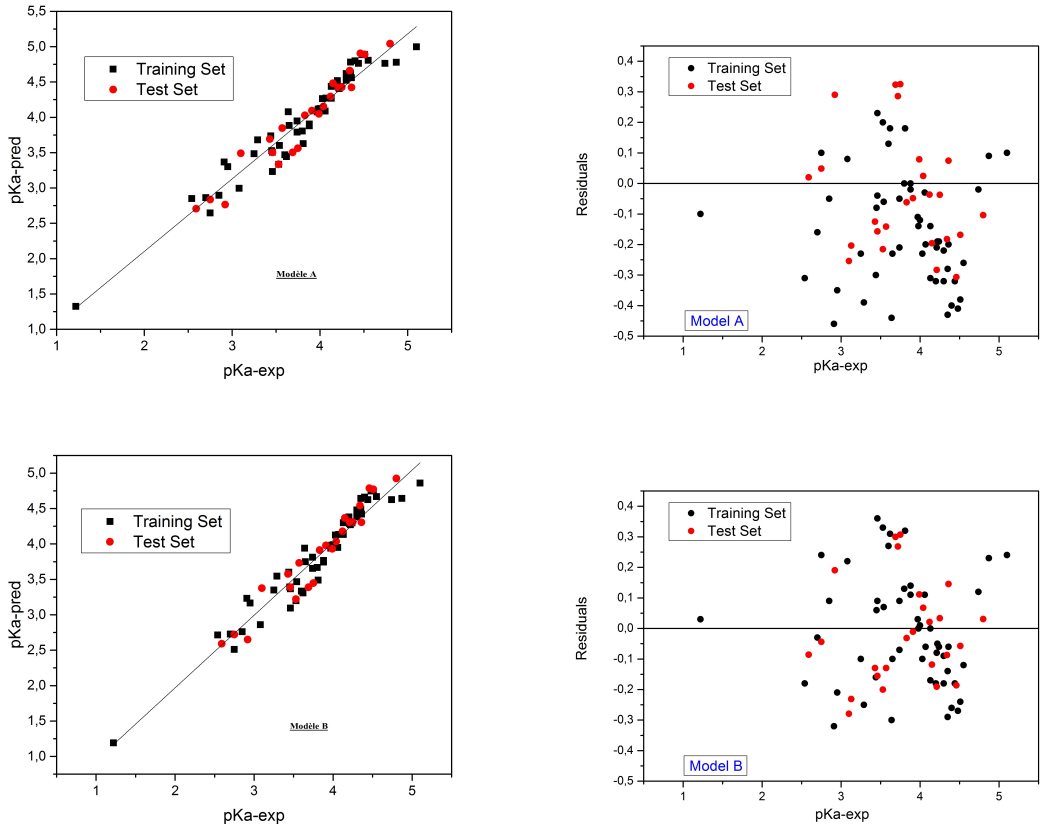
TABLE *5.5: Les valeurs expérimentales et calculées de pKa pour la série de test (25 composés) en utilisant le cycle A et le modèle C (Eq. (5.14)).*

	L'ensemble de test 25 composés	pKa-exp	ΔG_{aq}, Eq. (5.4) cycle A	pKa-pre Modele C Eq. (5.14)	residual
52	2-Hydroxy-5-chlorobenzoic-acid	2.59	3.69	2.67	0.08
53	2-Methyl-6-chlorobenzoic-acid	2.75	3.87	2.78	0.03
54	2-Chlorobenzoic-acid	2.92	3.77	2.72	0.20
55	3-5-Dichlorobenzoic-acid	3.10	4.76	3.38	0.28
56	4-Nitrobenzoic-acid	3.43	5.04	3.56	0.13
57	2-Methyl-6-methoxybenzoic-acid	3.46	5.12	3.62	0.16
58	2-Phenoxybenzoic-acid	3.53	5.30	3.74	0.21
59	2-Fluorobenzoic-acid	3.57	5.25	3.71	0.14
60	4-Formylbenzoic-acid	3.69	4.78	3.39	0.30
61	3-Trifluoromethylbenzoic-acid	3.75	4.86	3.45	0.30
62	3-Acetylbenzoic-acid	3.83	5.50	3.87	0.04
63	2-Methylbenzoic-acid	3.91	5.59	3.93	0.02
64	4-Bromobenzoic-acid	3.99	5.52	3.89	0.10
65	3-5-Dihydroxybenzoic-acid	4.04	5.67	3.98	0.06
66	3-Methoxybenzoic-acid	4.12	5.86	4.11	0.01
67	4-Fluorobenzoic-acid	4.15	6.12	4.29	0.14
67	Benzoic-acid	4.21	6.32	4.42	0.21
69	4-Methoxybenzoic-acid	4.25	6.04	4.23	0.02
70	4-Hydroxy-3-5-dimethoxybenzoic-acid	4.34	6.36	4.45	0.11
71	2-4-Dimethoxybenzoic-acid	4.36	6.03	4.23	0.13
72	4-Propoxybenzoic-acid	4.46	6.69	4.67	0.21
73	4-Methylbenzoic-acid	4.51	6.57	4.59	0.08
74	4-Ethoxybenzoic-acid	4.80	6.88	4.80	0.00
75	3-Chloro-5-nitrobenzoic acid	3.13	4.73	3.36	0.23
76	3-Nitro-4-methoxybenzoic acid	3.72	4.87	3.45	0.27

Les figures (5.3) représentent pka-pred en fonction de pka-exp pour les séries d'apprentissages (51 composés) et les séries de test (25 composés) pour les 4 modèles A, B, C et D.

Analyse de résidus

Le résidu est défini comme la différence entre la valeur prédite et la valeur expérimentale de pKa. Les figures (5.3) présentent les résidus en fonction des valeurs expérimentales de pKa pour les quatre modèles (A. B. C et D) pour les deux séries (apprentissage et test). On remarque que les résidus sont distribués d'une façon homogène de part et d'autre de zéro (résidu nul et le résidu maximal est inférieur à 0.5 unités de pKa).



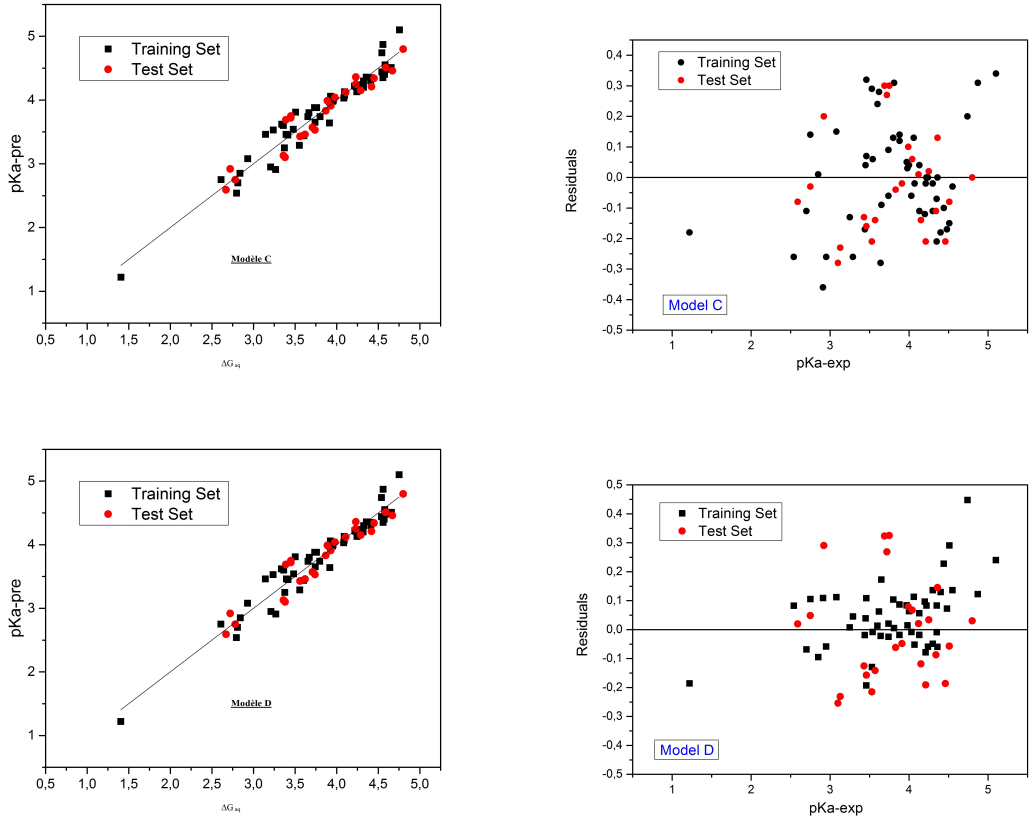


FIGURE *5.3: Pka-pred en fonction de pka-exp et résidus pour les 4 modèles A, B, C et D.*

5.4 Conclusion

Dans cette application, nous avons développé des modèles QSPR simples pour la prédiction des constantes d'acidité des acides benzoïques en solution aqueuse. Cette propriété a été modélisée pour une série de molécules dont les valeurs de pKa sont comprises entre 1.22 et 5.10 unité de pKa. Dans cette application,
i) Nous avons utilisé la théorie de la fonctionnelle de densité au niveau de calcul B3LYP/6-311++G**.
ii) L'effet du solvant a été pris en compte dans les calculs en l'utilisant d'un modèle de solvatation implicite de type SMD pour calculer les énergies libre de Gibbs en phase aqueuse.
iii) Deux cycles thermodynamiques ont été utilisés pour décrire le processus de déprotonation en phases gazeuse et aqueuse.
iv) La méthode de régression linéaire simple a été appliquée pour élaborer des modèles QSPR avec un seul descripteur ΔG_{aq}. Nous avons utilisé une série d'apprentissage constituée de 51 composés pour élaborer les modèles QSPR. La validation interne a été effectuée avec deux méthodes (la validation croisée et la division de la série d'apprentissage en 3 sous séries A, B et C). La Y-randomisation montre que les modèles QSPR obtenus ne sont pas dus à un coup de chance. La validation externe a été également effectuée en utilisant une série de test constituée de 25 composés hors-série et les 5 critères de Tropsha ont été également vérifiés confirment la stabilité, la fiabilité , la prédictivité des modèles QSPR obtenus.

Bibliographie

[1] Magill A. M., B.F. Yates, Basicity of Nucleophilic Carbenes in Aqueous and Nonaqueous Solvents-Theoretical Predictions, Aust. J. Chem., 57, (**2004**), 1205.

[2] Magill A. M., K.J. Cavell, B.F. Yates, An Assessment of Theoretical Protocols for Calculation of the pKa Values of the Prototype Imidazolium Cation J. Am. Chem. Soc., 126, (**2004**), 8717.

[3] Schüürmann, G. Modelling pKa of carboxylic acids and Chlorinated phenols, Quant. Struct. Act. Relat., 15, (**1996**), 121.

[4] Stewart R., The Proton : Applications to Organic Chemistry, Wasserman, H. H., Ed., Vol. 46 of Organic Chemistry, A series of Monographs., (**1985**), Academic Press : New York,

[5] Andrasi M., P. Buglyo, L. Zekany, A. Gaspar, A comparative study of capillary zone electrophoresis and pH-potentiometry for determination of dissociation constants. J. Pharm. Biomed. Anal., 44, (**2007**), 1040.

[6] Qiang Z., C. Adams, Potentiometric determination of acid dissociation constants (pKa) for human and veterinary antibiotics, Water Res., 38, (**2004**), 2874.

[7] Wröbel R., L. Chmurzynski, Potentiometric pKa determination of standard substances in binary solvent systems, Anal. Chim. Acta 405., (**2000**), 303.

[8] Beltrán J. L., N. Sanli, G. Fonrodona, D. Barrón, G. özkan, J. Barbosa, Spectrophotometric, potentiometric and chromatographic pKa values of polyphenolic acids in water and acetonitrile-water media, J. Barbosa, Anal. Chim. Acta., 484, (**2003**), 253.

[9] Elizabeth L., M. Miguel, Poliana L. Silva., Josefredo R. Pliego, Theoretical Prediction of pKa in Methanol : Testing SM8 and SMD Models for Carboxylic Acids, Phenols, and Amines, (**2014**), J. Phys. Chem. B., 118 (**2014**), (21), 5730.

[10] Lim C. D., Bashford, M. Karplus, Absolute pKa calculations with continuum dielectric methods, J. Phys. Chem., 95, (**1991**), 5610.

[11] Kelly C.P., C.J. Cramer, D.G. Truhlar, Adding explicit solvent molecules to continuum solvent calculations for the calculation of aqueous acid dissociation constants, J. Phys. Chem. A., 110, (**2006**), 2493.

[12] **(a)** Jasna J., Kliciń, Richard A. Friesner , Shi-Yi Liu , and Wayne C. Guida , Accurate prediction of acidity constants in aqueous solution via density functional theory and self-consistent reaction field methods, J. Phys. Chem. A., 106, (**2002**), 1327, **(b)** J. Ulander, A. Broo, Use of empirical correction terms in calculating ionization constants, Int. J. Quantum Chem., 105, (**2005**), 866.

[13] **(a)** Jang Y.H., L.C. Sowers, T. Cagin, W.A. III Goddard, First principles calculation of pKa values for 5-substituted uracils, J. Phys. Chem. A., 105, (**2001**), 274, **(b)** K.N. Rogstad, Y.H. Jang, L.C. Sowers, W.A. III Goddard, First principles calculations of the pKa values and tautomers of isoguanine and xanthine, Chem. Res. Toxicol., 16, (**2003**), 1455, **(c)** Hwang S., Y.H. Jang, D.S. Chung, Gas phase proton affinity, basicity, and pKa values for nitrogen containing heterocyclic aromatic compounds, Bull. Korean Chem. Soc., 26, (**2005**), 585.

[14] Charif I.E., S.M. Mekelleche, D. Villemin, N. Mora-Diez, Correlation of aqueous pKa values of carbon acids with theoretical descriptors : A DFT study, 818, (**2007**), 1, 1.

[15] Liptak M.D., G.C. Shields, Accurate pKa calculations for carboxylic acids using complete basis set and Gaussian-n models combined with CPCM continuum solvation methods, J. Am. Chem. Soc., 123, (**2001**), 7314.

[16] Ghalami C. B., Dezhampanah, H., Nikparsa, P. and Ghiami-Shomami, A. Theoretical calculation of the pKa values of some drugs in aqueous solution. Int. J. Quantum Chem., 112, (**2012**), 2275.

[17] Habibi-Yangjeh A., Danandeh-Jenagharad, M., and Nooshyar, M., Prediction Acidity Constant of Various Benzoic Acids and Phenols in Water Using Linear and Nonlinear QSPR Models, Bull. Korean Chem. Soc., 26, (**2005**), 12.

[18] Fiche de sécurité du Programme International sur la Sécurité des Substances Chimiques, mai **2009**.

[19] Liptak M.D., Gross, K.C., Seybold, P.G., Feldgus, S., Shields, G.C. Absolute pKa determinations for substituted phenols. J. Am. Chem. Soc., 124(22), (**2002**), 6421.

[20] Liptak M.D., Shields, G.C. Experimentation with different thermodynamic cycles used for pKa calculations on carboxylic acids using complete basis set and Gaussian-n models combined with CPCM continuum solvation methods. Int. J. Quantum Chem., 85, (**2001**), 727.

[21] Gõmez-Bombarelli R., M. Gonzàlez-Pérez, M. Teresa Pérez -Prior, E. Calle, J. Casado, Computational Study of the Acid Dissociation of Esters and Lactones. A Case Study of Diketene, J. Org. Chem., 74, (**2009**), 4943.

[22] Junming Ho., Michelle L. Coote, A universal approach for continuum solvent pKa calculations : Are we there yet ? Theor. Chem. Acc., 125, (**2010**), (1-2), 3.

[23] Josefredo R. Pliego Jr. Thermodynamic cycles and the calculation of pKa, Chem. Phys. Lett., 367, (**2003**), 145.

[24] Kelly C.P., Cramer, C.J., Truhlar, ΔG Aqueous solvation free energies of ions and ion-water clusters based on an accurate value for the absolute aqueous solvation free energy of the proton. J. Phys. Chem. B., 110(32), (**2006**), 16066.

[25] Tissandier M.D., K.A. Cowen, W.Y. Feng, E. Gundlach, M.J. Cohen, A.D. Earhart, J.V. Coe, The proton's absolute aqueous enthalpy and Gibbs free energy of solvation from cluster-ion solvation data, J. Phys. Chem. A., 102, (**1998**), 7787.

[26] Brown T.N., Mora-Diez, N. Computational determination of aqueous pK(a) values of protonated benzimidazoles (part 1). J. Phys. Chem. B., 110(18), (**2006**), 9270.

[27] Gramatica P. Principles of QSAR models validation : Internal and external. QSAR & Combinatorial Science, 26(5), (**2007**), 694.

[28] Hawkins D. M., Basak, S. C., Mills, D., Assessing model fit by cross validation. Journal of Chemical Information and Computer Sciences, 43(2), (**2003**), 579.

[29] Golbraikh A., A. Tropsha, Beware of q2 !, J. Mol. Graph. Model., 20, (**2002**), 269.

[30] **(a)** Tropsha A., P. Gramatica, V.K. Gombar, The importance of being earnest : validation is the absolute essential for successful application and interpretation of QSPR models, QSAR Comb. Sci., 22, (**2003**), 69. **(b)** Tropsha A. Best practices for QSAR model development, validation, and exploitation. Molecular Informatics, 29, (**2010**), 476.

[31] Afantitis A., G. Melagraki, H. Sarimveis, P.A. Koutentis, O. Igglessi-Markopoulou, G. Kollias, A combined LS-SVM & MLR QSAR workflow for predicting the inhibition of CXCR3 receptor by quinazolinone analogs, Mol. Divers., 14, (**2010**), 225.

[32] A. Tropsha A. Golbraikh, Predictive QSAR modeling workflow, model applicability domains, and virtual screening, Curr. Pharm. Des., 13, (**2007**), 3494.

[33] Rucker C., G. Rucker, M. Meringer, y-Randomization and its variants in QSPR/QSAR. Journal of Chemical Information and Modeling, 47(6), (**2007**), 2345.

[34] Lindgren F., B. Hansen, W. Karcher, M. Sjöström, and L. Eriksson, Model validation by permutation tests : Applications to variable selection, J. Chemometrics, 10, (**1996**), 521.

[35] Chemdraw is a molecule editor © (2008). http ://www.cambridgesoft.com/

[36] Becke A. D., Correlation energy of an inhomogeneous electron gas : A coordinate-space model, J. Chem. Phys., 88, (**1988**), 1053.

[37] Becke A. D. Density-Functional Exchange-Energy Approximation with Correct Asymptotic Behavior, Phys. Rev. A., 38, (**1988**), 3098.

[38] Gaussian 09, Revision A.1, Frisch, M. J., Trucks, G. W., Schlegel, H. B., Scuseria, G. E., Robb, M. A., Cheeseman, J. R., Scalmani, G., Barone, V., Mennucci, B., Petersson, G. A., Nakatsuji, H., Caricato, M., Li, X., Hratchian, H. P., Izmaylov, A. F., Bloino, J., Zheng, G., Sonnenberg, J. L., Hada, M., Ehara, M., Toyota, K., Fukuda, R., Hasegawa, J., Ishida, M., Nakajima, T., Honda, Y., Kitao, O., Nakai, H., Vreven, T., Montgomery, J. A., Jr., Peralta, J. E., Ogliaro, F., Bearpark, M., Heyd, J. J., Brothers, E., Kudin, K. N., Staroverov, V. N., Kobayashi, R., Normand, J., Raghavachari, K., Rendell, A., Burant, J. C., Iyengar, S. S., Tomasi, J., Cossi, M., Rega, N., Millam, M. J., Klene, M., Knox, J. E., Cross, J. B., Bakken, V., Adamo, C., Jaramillo, J., Gomperts, R., Stratmann, R. E., Yazyev, O., Austin, A. J., Cammi, R., Pomelli, C., Ochterski, J. W., Martin, R. L., Morokuma, K., Zakrzewski, V. G., Voth, G. A., Salvador, P., Dannenberg, J. J., Dapprich, S., Daniels, A. D., Farkas, Ö., Foresman, J. B., Ortiz, J. V., Cioslowski, J., Fox, D. J. Gaussian, Inc., Wallingford CT, (**2009**).

[39] Marenich V., C. J. Cramer, and D. G. Truhlar, "Universal solvation model based on solute electron density and a continuum model of the solvent defined by the bulk dielectric constant and atomic surface tensions," J. Phys. Chem. B., 113, (**2009**), 6378.

[40] Minitab 16 Statistical Software, (**2010**), State College, PA : Minitab, Inc. www.minitab.com

[41] David R. Lide, CRC Handbook of Chemistry and Physics 90th Edition, (**2010**), Edition CRC Press.

[42] Pytela O., J. Kulhànek, Ortho Effect in Dissociation of Benzoic Acids with Electron-Accceptor Substituents Using the AISE Theory, Relation topara Substitution and Solvent, Collect. Czech. Chem. Commun., 67, (**2002**), 596.

[43] Pytela O., J. Kulhànek, M. Ludwig, V., Chemometrical Analysis of Substituent Effects. III. Additivity of Substituent Effects in Dissociation of 3,4-Disubstituted Benzoic Acids in Organic Solvents, Collect. Czech. Chem. Commun., 59, (**1994**), 627.

[44] Pytela O., J. Kulhànek, M. Ludwig, Chemometrical Analysis of Substituent Effects. IV. Additivity of Substituent Effects in Dissociation of 3, 5-Disubstituted Benzoic Acids in Organic Solvents, Collect. Czech. Chem. Commun., 59, (**1994**), 1637.

[45] Pytela O., J. Kulhànek, Chemometric Analysis of Substituent Effects. VI. A Study of ortho Effect in Dissociation of 2, 6-Disubstituted Benzoic Acids, Collect. Czech. Chem. Commun., 60, (**1995**), 829.

[46] Pytela O., O. Prusek, Chemometric Analysis of Substituent Effects. XII. Application of Relationship Between 2- and 4- Substitution of Benzene Ring to Study ortho Effect in

Selected Compounds with Different Reaction Centres, Collect. Czech. Chem. Commun., 64, (**1999**), 1617.

[47] Pytela O., J. Kulhànek, Chemometric analysis of substituent effects .11. solvent effects on dissociation of 2,6-disubstituted benzoic-acids, Collect. Czech. Chem. Commun., 62, (**1997**), 913.

[48] Rived F., M. Rosés, E. Bosch, Dissociation constants of neutral and charged acids in methyl alcohol. The acid strength resolution, Anal. Chim. Acta., 374, (**1998**), 309.

[49] Rived F., I. Canals, E. Bosch, M. Rosés, Acidity in methanol-water, Anal. Chim. Acta., 439, (**2001**), 315.

[50] Izutsu K., Acid-Base Dissociation Constants in Dipolar Aprotic Solvents, (**1990**), Blackwell Scientific Publications, Oxford.

[51] Chantooni M. K., I. M. Kolthoff, Resolution of acid strength in tert-butyl alcohol and isopropyl alcohol of substituted benzoic acids, phenols, and aliphatic carboxylic acids, Anal. Chem., 51, (**1979**), 133.

[52] Kolthoff I. M., M. K. Chantooni, Substituent effects on dissociation of benzoic acids and heteroconjugation of benzoates with p-bromophenol in acetonitrile, N,N-dimethylformamide, and dimethyl sulfoxide. Intramolecular hydrogen bonding in o-hydroxybenzoic acids and their anions, J. Am. Chem. Soc., 93, (**1971**), 3843.

[53] Kulhànek J., O. Exner, Solvation and Steric Hidrance in Methyl-substituted Benzoic Acids, J. Chem. Soc. Perkin Trans., 2, (**1998**), 1397.

CONCLUSION GÉNÉRALE

Dans cette thèse, nous avons utilisé la méthodologie QSPR pour élaborer des modèles fiables, stables et prédictifs pour deux propriétés physico-chimiques :

• La température de fusion des acides gras.

• Les constantes d'acidité pKa des acides benzoïques dans l'eau.

La méthodologie QSPR standard a été utilisée dans ce travail.

• Elaboration du modèle pour une série d'apprentissage et analyse des paramètres statistiques des modèles obtenus.

• Vérification de la stabilité interne du modèle obtenu avec la validation croisée (R^2_{CV})

• Test la stabilité du modèle avec une série externe.

• Vérification des critères de Tropsha.

• La Y-randomisation pour vérifier que le modèle QSPR obtenu n'est pas du au hasard.

Dans la première application, nous avons utilisé la méthode BMLR (Best Multiple Linear Regression) implémentée dans le logiciel Codessa pour élaborer des modèles QSPR fiables capables de prédire la température de fusion d'une base de données constituée de 62 acides gras dont les valeurs expérimentales sont comprises entre -65.00 et 96.00 $^{\circ}$C. La présente étude montre que le meilleur modèle QSPR est défini en termes de cinq descripteurs moléculaires non corrélés entre eux et qui sont essentiellement de nature électrostatique et topologique. Le modèle QSPR élaboré pourrait être utilisé pour prédire le point de fusion de nouveaux acides gras ou des acides gras pour lesquels la température de fusion expérimentale est indisponible dans la littérature.

Dans la deuxième application, nous avons élaboré des modèles QSPR simples, robustes, stables et prédictifs pour l'estimation des constantes d'acidité pKa en phase aqueuse d'une base de données constituée de 76 acides benzoïques. Les modèles QSPR ont été élaborés avec un seul descripteur pertinent qui est l'enthalpie libre de déprotonation. Le calcul des énergies libres de Gibbs de déprotonation des acides et de leurs anions correspondants, en phases gazeuse et aqueuses, a été effectué

en utilisant la théorie de la fonctionnelle de la densité (DFT) avec la fonctionnelle hybride B3LYP et la base étendue 6-311++G(d,p). Les effets de solvant ont été pris en compte dans les calculs en phase aqueuse avec le modèle de solvatation SMD. Deux cycles thermodynamiques ont été considérés pour décrire le processus de déprotonation dans les deux phases gazeuse et aqueuse. Des modèles QSPR simples, fiables, stables et prédictifs ont été élaborés avec la méthode de régression linéaire simple. Une très bonne corrélation linéaire entre les valeurs de pKa expérimentales des acides benzoïques et les valeurs des énergies libres de Gibbs de déprotonation. La stabilité du modèle et le pouvoir prédictif des modèles QSPR a été testé avec succès sur une série de tests constitués de 25 acides non inclus dans la série d'apprentissage. L'écart absolu moyen entre les valeurs de pKa calculées et expérimentales est inférieur à 0,36 en unités de pKa et le coefficient de corrélation $R^2 = 0,93$. De plus, les 5 critères de Tropsha ont été vérifiés et la Y-randomisation a été effectuée pour s'assurer que les modèles obtenus ne sont pas dus au hazard.

Le travail effectué lors de cette thèse ouvre de nombreuses perspectives. Nous envisageons de poursuivre ce travail par :

• Elaboration des modèles QSPR pour la prédiction des constantes d'acidité et de basicité pour d'autres familles de composés chimiques.

• Utilisation de différentes techniques pour l'élaboration des modèles QSPR telles que les méthodes PCA (principal component analysis), PLS (partial least squares), ANN (Artificial Neural Network) et GA (Genetic Algorithms).

• Élaboration des modèles QSPR fiables pour prédire différentes propriétés physico-chimiques d'intérêt chimique, biologique, pharmaceutique ou industriel.

ANNEXE A

Prediction of the melting points of fatty acids from computed molecular descriptors : A quantitative structure–property relationship study

Chemistry and Physics of Lipids 165 (2012) 1–6

Contents lists available at SciVerse ScienceDirect

Chemistry and Physics of Lipids

journal homepage: www.elsevier.com/locate/chemphyslip

Prediction of the melting points of fatty acids from computed molecular descriptors: A quantitative structure–property relationship study

Abdelkrim Guendouzi [a,*], Sidi Mohamed Mekelleche [b]

[a] Laboratory of Applied Thermodynamics and Molecular Modeling, Department of Chemistry, Faculty of Science and Technology, University of Saida, PB 138, Saida, 20000, Algeria
[b] Laboratory of Applied Thermodynamics and Molecular Modeling, Department of Chemistry, Faculty of Science, University of Tlemcen, PB 119, Tlemcen, 13000, Algeria

ARTICLE INFO

Article history:
Received 26 July 2011
Received in revised form 1 October 2011
Accepted 3 October 2011
Available online 10 October 2011

Keywords:
Fatty acids
Melting point
QSPR
Multilinear regression
Codessa

ABSTRACT

The aim of these works present in this paper consisted in the development and evaluation of quantitative structure property models (QSPR) for the prediction of the melting points of a series constituted by 62 fatty acids. The best multilinear regressions method (MLR) is used to develop models for the prediction of the melting points. The descriptors of the model are selected among an extended set of more than 500 descriptors (constitutional, topological, geometric, quantum chemical and thermodynamic descriptors). Applicability domains were defined and the predictive power was determined using a set of validations The QSPR models are established using the BMLR method implemented in CODESSA software, It turns out that the best QSPR model ($R^2 = 0.948$, $R^2_{adj} = 0.936$, SD = 0.940 and F-test = 190.90) is obtained with five molecular descriptors.

Published by Elsevier Ireland Ltd.

1. Introduction

The economic importance of the distinctive melting and solidification behavior of fatty acids and their esters was well described by Bailey (1950). Fatty acids and their derivatives constitute an almost infinite variety of long-chain compounds differing in carbon chain length, unsaturation, and isomerism, resulting in a complex and fascinating gamut of physical properties. The transformation from the liquid to the solid state is accompanied by release of heat (latent heat of crystallization signifying an exothermic reaction); the reverse, transformation from solid to liquid, is accompanied by a negative heat effect (endothermic reaction). This forms the basis for the widely used technique of differential scanning calorimetry. Another important phenomenon is melting expansion. Conversion from the solid state to the liquid state results in a melting expansion that is added to the normal thermal expansion. This phenomenon constitutes the basis for the determination of the solid fat index by dilatometry. Protons in the solid state of a fat behave differently from those in the liquid state when subjected to radiofrequency energy when the sample is contained in a magnetic field. This serves as the basis for the determination of the solid fat content in a product by wide-line or pulsed-nuclear magnetic resonance (Harold et al., 2002).

Melting of a fat is an instantaneous reaction, whereas crystallization is usually a slow process. The driving force in crystallization is the degree of supercooling phases: nucleation and crystal growth. A high degree of supercooling will be conducive to nucleation, and many small crystals will be formed. At temperatures closer to the crystallization point, crystal growth will be favored and large crystals will be formed. Another result of a high degree of supercooling is the formation of mixed crystals, also known as solid solutions. Molecules with a range of melting points may crystallize together. As a result, rapidly cooled fats may have higher solid fat content than the same fats that cooled more gradually. These mixed crystals will partially melt when the fat is subjected to temperature variations below its melting point, a phenomenon known as tempering (Moziar et al., 1989).

Svenstrup et al. (2005) have shown that the number of melting points and the melting temperature are correlated with the cooling rate for pork fat, lard, and leaf fat in three different products: extracted fat, raw fat, and fat as an ingredient in liver pate, a rapid cooling leads to lowering of the melting point, assigned to the presence of unstable β′ crystals, and that the melting points vary with the treatment of the fat. The findings suggest that the fraction of unsaturated fatty acids present in the fat is important for both crystallization rate and melting points of α and β crystals in extracted lard, and is less pronounced in liver pate because of the presence of diverse components such as proteins. The identification of the various levels of structure present in fat crystal networks, and the development of analytical techniques to quantify these levels have

* Corresponding author. Tel.: +213662623590.
E-mail addresses: guendouzzi@yahoo.fr (A. Guendouzi), sidi_mekelleche@yahoo.fr (S.M. Mekelleche).

0009-3084/$ – see front matter. Published by Elsevier Ireland Ltd.
doi:10.1016/j.chemphyslip.2011.10.001

Table 1
Experimental and calculated values of Mp of fatty acids A, B, and C correspond to the subsets used in cross validation procedure.

		Structure names	Cal. Mp	Exp. Mp	Residual
1	A	3-7-11-15-Tetramethylhexadecanoic acid	−58.5	−65.0	6.5
2	B	cis-cis-cis-cis-6-9-12-15-Octadecatetraenoic acid	−30.6	−57.0	26.4
3	C	cis-cis-cis-cis-5-8-11-14-Eicosatetraenoic acid	−46.8	−49.0	2.2
4	A	cis-cis-cis-cis-cis-cis-4-7-10-13-16-19-Docosahexaenoic acid	−48.6	−45.0	−3.6
5	B	Pentanoic acid	−23.2	−33.0	9.8
6	C	3-Methylbutanoic acid	−32.9	−29.0	−3.9
7	A	cis-cis-cis-9-12-15-Octadecatrienoic acid	−27.6	−11.0	−16.6
8	B	cis-cis-9-12-Octadecadienoic acid	−15.5	−7.0	−8.5
9	C	Heptanoic acid	2.8	−7.0	9.8
10	A	Butanoic acid	−5.9	−5.0	−0.9
11	B	cis-9-Tetradecenoic acid	7.3	−4.0	11.3
12	C	cis-cis-5-13-Docosadienoic acid	23.1	−4.0	27.1
13	A	Hexanoic acid	−9.3	−3.0	−6.3
14	B	cis-9-Hexadecenoic acid	13.2	0.0	13.2
15	C	12-Hydroxy-cis-9-octadecenoic acid	17.7	5.0	12.7
16	A	Nonanoic acid	19.6	12.0	7.6
17	B	cis-9-Octadecenoic acid	17.0	13.0	4.0
18	C	cis-11-Octadecenoic acid	18.6	15.0	3.6
19	A	Octanoic acid	9.2	16.0	−6.8
20	B	cis-trans-9-11-Octadecadienoic acid	2.0	20.0	−18.0
21	C	trans-cis-10-12-Octadecadienoic acid	11.4	23.0	−11.6
22	A	cis-11-Eicosenoic acid	23.0	24.0	−1.0
23	B	cis-9-Eicosenoic acid	26.0	24.0	2.0
24	C	9-Decenoic acid	−1.3	26.0	−27.3
25	A	cis-5-Eicosenoic acid	28.7	27.0	1.7
26	B	Undecanoic acid	32.2	28.0	4.2
27	C	cis-6-Octadecenoic acid	18.3	29.0	−10.7
28	A	Decanoic acid	27.8	31.0	−3.2
29	B	cis-12-13-Epoxy-cis-9-octadecenoic acid	25.5	32.0	−6.5
30	C	trans-trans-cis-9-11-13-Octadecatrienoic acid	53.3	32.0	21.3
31	A	cis-11-Docosenoic acid	26.2	33.0	−6.8
32	B	cis-13-Docosenoic acid	28.1	34.0	−5.9
33	C	trans-trans-cis-8-10-12-Octadecatrienoic acid	51.1	40.0	11.1
34	A	Tridecanoic acid	45.6	41.0	4.6
35	B	cis-15-Tetracosenoic acid	34.4	43.0	−8.6
36	C	Dodecanoic acid	39.6	43.0	−3.4
37	A	trans-11-Octadecenoic acid	43.1	44.0	−0.9
38	B	cis-trans-cis-9-11-13-Octadecatrienoic acid	51.3	45.0	6.3
39	C	trans-9-Octadecenoic acid	43.1	45.0	−1.9
40	A	cis-trans-trans-9-11-13-Octadecatrienoic acid	41.2	49.0	−7.8
41	B	Pentadecanoic acid	56.5	52.0	4.5
42	C	Tetradecanoic acid	51.1	54.0	−2.9
43	A	Heptadecanoic acid	63.7	61.0	2.7
44	B	trans-13-Docosenoic acid	58.9	61.0	−2.1
45	C	Hexadecanoic acid	57.7	62.0	−4.3
46	A	Nonadecanoic acid	67.8	69.0	−1.2
47	B	Octadecanoic acid	64.3	69.0	−4.7
48	C	trans-trans-trans-9-11-13-Octadecatrienoic acid	69.9	71.0	−1.1
49	A	Eicosanoic acid	70.7	76.0	−5.3
50	B	Pentacosanoic acid	83.9	77.0	6.9
51	C	Tricosanoic acid	79.3	79.0	0.3
52	A	Docosanoic acid	77.1	81.0	−3.9
53	B	Heneicosanoic acid	73.8	82.0	−8.2
54	C	cis-trans-trans-cis-9-11-13-15-Octadecatetraenoic acid	74.1	86.0	−11.9
55	A	Heptacosanoic acid	88.5	87.0	1.5
56	B	Tetracosanoic acid	81.0	87.0	−6.0
57	C	Hexacosanoic acid	85.7	88.0	−2.3
58	A	Nonacosanoic acid	93.0	90.0	3.0
59	B	Octacosanoic acid	90.4	90.0	0.4
60	C	Hentriacontanoic acid	97.6	93.0	4.6
61	A	Triacontanoic acid	94.8	93.0	1.8
62	B	Dotriacontanoic acid	102.8	96.0	3.1
		Minimum	−58.5	−65.0	–
		Maximum	102.8	96.0	–
		Mean	34.9	34.8	0.0
		Median	33.3	33.5	−0.9

been reviewed by Narine and Marangoni (1999, 2002). The types, formulations, functionality, and processing required for the production of lipid-shortening systems, as well as their crystallization, structural elucidation, and mechanical modeling of fat crystal networks have been reviewed by Ghotra et al. (2002). Also, Humphrey et al. (2004) have compared the lipid-shortening functionality as a function of molecular ensemble and shear: crystallization and melting.

Alternatively, the quantitative structure–property relationship (QSPR) provides a promising method for estimating the melting point of fatty acids based on descriptors derived solely from the molecular structure to fit experimental data. The QSPR is based

Table 2
Descriptors involved in the best two and five-parameter models (62 compounds), the corresponding regression coefficients X, the Errors ΔX, t-test, and p-values.

Descriptor	X	ΔX	t-Test	p-Value
Model # 1: R^2 = 0.813; Adjusted R^2 = 0.807, F = 281.67; Std. Error of estimate: 17.8843				
Intercept	−51.981	9.963	−5.22	
D1	−21.211	1.914	−11.08	$<10^{-4}$
D2	8.8311	0.7884	11.20	$<10^{-4}$
Model # 2: R^2 = 0.946; Adjusted R^2 = 0.942, F = 197.93; Std. Error of estimate: 9.9436				
Intercept	227.18	20.84	10.90	
D3	−1.7122	0.1227	−13.95	$<10^{-4}$
D4	17.002	1.253	13.57	$<10^{-4}$
D5	−164.12	10.82	−15.17	$<10^{-4}$
D6	3.1603	0.4531	6.97	$<10^{-4}$
D7	15601	4745	3.29	0.002
Model # 3: R^2 = 0.917; Adjusted R^2 = 0.913, F = 215.469; Std. Error of estimate: 12.0377				
Intercept	−38.595	8.383	−4.604	
D3	0.448	0.038	11.626	$<10^{-4}$
D4	0.737	0.038	19.349	$<10^{-4}$
D6	−0.571	0.038	−14.962	$<10^{-4}$
Model # 4: R^2 = 0. 937; Adjusted R^2 = 0. 933, F = 213.730; Std. Error of estimate: 19.52687				
Intercept	75.676	19.526	3.875	
D3	0.487	0.034	14.149	$<10^{-4}$
D4	0.757	0.034	22.269	$<10^{-4}$
D5	−0.170	0.038	−4.422	$<10^{-4}$
D6	−0.412	0.038	−10.706	$<10^{-4}$

on the assumption that the variation of the behavior of the compounds, as expressed by any measured physicochemical properties, can be correlated with numerical changes in structural features of all compounds, termed "molecular descriptors" (Devillers, 1999; Karelson, 2000). The advantage of this approach lies in the fact that it requires only the knowledge of the chemical structure and is not dependent on any experimental properties. Once a correlation is established, it can be applicable for the prediction of the property of new compounds that have not been synthesized or found. Thus the QSPR approach can expedite the process of development of new molecules and materials with desired properties. The QSPR has been successfully used to investigate the spectral properties of various systems, for example, the prediction of the Melting Point of a ionic liquids the estimation of boiling points of organic molecules based on density functional theory (DFT) calculations the prediction of viscosity of imidazolium-based ionic liquids the modeling of relative fluorescence intensity ratio of Eu (III) complex the prediction of some proprieties: Solubility, Boiling Points, Cloud Point (Katritzky et al., 1996a,b; Paul et al., 2002), the analysis of antimicrobial peptides (Stefano and Erminia, 2007).

In this study, we present a comprehensive database of melting points of fatty acids, which was collected from the literature and from original measurements (David R., 2009). We also describe newly developed methodology for the fast selection of descriptors in quantitative structure–property relationships (QSPR) analysis. We propose a model that correlates the MPs property with 62 compounds fatty acids.

2. Melting and crystallization of fatty acids

The internal structure of solid fats in the crystalline state is fairly well-known. Crystals are closely packed systems of molecules or atoms with a regular three-dimensional repeating order. The molecules or atoms are held together by strong forces, and crystals have well-defined spacing's between repeating groups. These short and long spacing's are characteristic of different fat crystals. Although it can be assumed that the internal structure of liquids is represented by molecules in complete disorder, there is evidence that a limited degree of order does exist in liquids. According to Bailey (1950), long-chain compounds in the liquid state contain limited areas of orderly packing. However, these regions of orderly packing extend only over a very limited number of molecules, whereas in a crystal the lattice arrangement extends throughout the crystal. Bailey speaks of the quasicrystalline character of liquids as "short range order" and of crystals as "long range order" Studies by Hernqvist (1984) shed light on the structure of triglycerides in the liquid state and how this affects crystallization. On the basis of X-ray diffraction and Raman spectroscopy studies, he suggested a gradual decrease in size with increasing temperature of the melt (Schema 1). The order in the melt is constant even at 40 °C above the melting point. The order is related to chain length; a long chain is more disordered at the methyl end-group plane than a short one. When the temperature is decreased, the lamellar units increase in size until crystallization occurs (Schema 2).

3. Computational details and data set

3.1. Dataset

The reported experimental data of melting point for 62 fatty acids (see Schema 3 for the structures of FAs studied in this work), are reported in Table 1 (David R., 2009). Table 1 gives the names and melting point of some important fatty acids. It includes most of the acids that are significant constituents of naturally occurring oils. The values of Tm are ordered in the ascending order.

3.2. Geometry optimization procedure and quantum chemistry calculations

The three-dimensional structures of the molecules have been drawn using the graphical interface of the ChemBioDraw Ultra (Chemdraw, 2008). The equilibrium geometries of all systems were optimized using the PM6 (Stewart, 2007) semi-empirical method implemented in MOPAC program (Stewart, 2008). Quantum chemistry and thermodynamic properties were calculated using the Becke's-three-parameter hybrid density functional B3LYP (Becke, 1988) method with the standard 6-31G basis set Single point and frequency calculations were carried out using Gaussian 03 package (Frisch et al., 2007). All calculation outputs were loaded into Codessa software (Katritzky et al., 1997a). Overall, more than 500 theoretical descriptors were calculated.

Table 3
Correlation matrix of all descriptors involving in models # 1–4.

		Number of double bonds	PPSA-3 atomic charge weighted PPSA	Number of aromatic bonds	Average complementary information content (order 1)	Balaban index	YZ shadow	Min atomic orbital electronic population	Melting points
D1	Number of double bonds	1.00							
D2	PPSA-3 atomic charge weighted PPSA	−0.11	1.00						
D3	Number of aromatic bonds	−0.28	−0.03	1.00					
D4	Average complementary information content (order 1)	0.40	−0.77	−0.14	1.00				
D5	Balaban index	0.45	0.05	0.20	0.01	1.00			
D6	YZ shadow	0.57	−0.01	0.16	−0.03	0.49	1.00		
D7	Min atomic orbital electronic population	0.11	−0.11	−0.14	−0.44	0.02	−0.18	1.00	
	Melting points (property)	0.20	0.13	0.25	0.68	−0.28	−0.52	−0.22	1.00

3.3. QSPR analysis

Methodology for a general QSAR/QSPR approach has been developed and coded as the CODESSA software package. CODESSA enables the calculation of numerous quantitative descriptors solely on the basis of molecular structural information Hansch-type approach (Katritzky et al., 1995). Research using CODESSA has successfully correlated and predicted various physical and biological properties reported by (Katritzky et al., 1997b, 2004, 2005a,b; Basak and Mills, 2005; Karelson et al., 1999; Thakur, 2005) including gas chromatographic properties, melting and boiling points, solvent scales, refractive indexes, human breast milk and antimalarial activity.

In the present study we used Codessa software for the prediction of melting point; we believe the modules integrated in this software package provide an optimum way to high quality results. QSPR modeling, by multilinear regression utilized in the Codessa program which applies up to 500 different constitutional, geometrical, topological, electrostatic, quantum chemical, and thermodynamic molecular descriptors. Topological descriptors contain information about the number, type, and connectivity of atoms in the molecule. Such descriptors include atom, bond, and path counts. Information about the electronic aspects of the structures is encoded by electronic descriptors. Examples of such descriptors include the dipole moment and the sum of the negative charges in the molecule. Geometric descriptors are used to encode information about the three-dimensional structure of a compound. One such descriptor is the molecular volume. Finally, hybrid descriptors such as charged partial surface area (CPSA) descriptors combine information about both the geometric and electronic aspects of a molecule. In our treatment all descriptors used were derived solely from molecular structure and do not require experimental data.

3.4. Multilinear regression (MLR) analysis

Multilinear regression fits a linear model of the form:

$$Y = b_0 + b_1X_1 + b_2X_2 + \ldots + b_kX_k + e \quad (1)$$

where Y is the dependent variable (response) and $X_1, X_2, \ldots, X_k$ are the independent variables (predictors) and e is a random error, b_0, $b_1, b_2, \ldots, b_k$ are known as the regression coefficients, which have to be estimated from the data. The MLR algorithm chooses regression coefficients so as to minimize the squared sum of the difference between predicted values and measured values. MLR is performed either to study the relationship between the response variable and predictor variables or to predict the response variable based on the predictor variables.

3.5. Validation of the correlations

This step consists in testing the accuracy of the correlations by using them to predict the property of interest for a proper validation set. This set is composed of substances similar to those of the training set and for which the property of interest is experimentally known as well. The predictive capability of the correlation is evaluated by comparing the predicted and the experimental value of the property. The regression correlation coefficient R^2 and the cross-validation R^2_{cv} were assumed as an estimate of the predictive capability of the correlations.

The cross validation R^2_{cv} is a popular technique used to explore the reliability of statistical models. Based on this technique, a number of modified data sets are created by deleting in each case one or group of objects (leave-one-out or leave-many-out). For each data set, an input–output model is developed, based on the utilized modeling technique. The model is evaluated by measuring its accuracy in predicting the responses of the remaining data. The

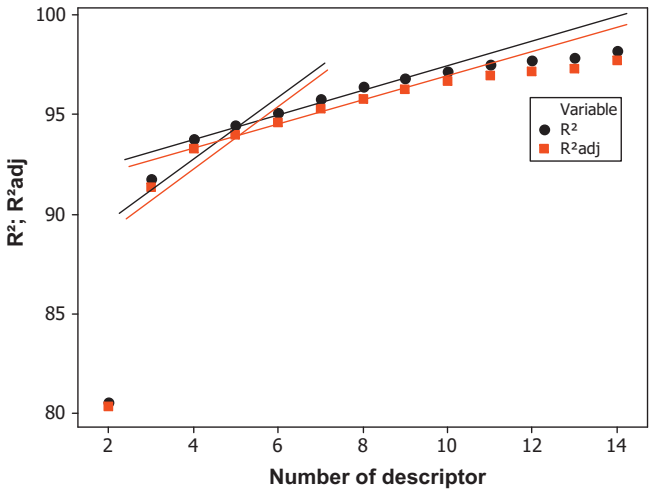


Fig. 1. Influence of the number of descriptors on R^2 and R^2_{adj}.

adjusted coefficient of determination of a multiple linear regression model R^2_{adj} is defined in terms of the coefficient of determination as follows, where n is the number of observations in the data set, and p is the number of independent variables.

$$R^2_{adj} = 1 - (1 - R^2)\frac{n-1}{n-p-1} \quad (2)$$

The cross-validation procedure was carried out using Minitab program (Minitab, 2010).

4. Results and discussions

The Codessa program provides a large variety of non-empirical molecular descriptors. In the present treatment, the preliminary regression analysis was carried out using all the original Codessa descriptors which numbered 322. After the BMLR regression, the pool of the descriptors was reduced to 278. To find the optimum number of descriptors describing (Mp) for the current set of fatty acids, the BMLR correlations performed for all compounds, providing the optimal equations for different numbers of descriptors in the range of 2–14. The influence of the number of the descriptors on the correlation coefficient R^2 and the adjusted coefficient of correlation R^2_{adj} are given in Fig. 1. To avoid the "overparameterization" of the model, an increase of the R^2 values of less than 0.02 was chosen as the breakpoint criterion (Katritzky et al., 2004). The quantitative relationships between MPs and various descriptors are analyzed. In Table 1, we report experimental (David R., 2009) and predicted data for each substance. The parity plot which helps in detecting any outliers and provides a prompt indication of the accuracy of the correlation is given in Fig. 2.

In Table 2, we tabulated the different QSPR models obtained using two, and five parameter models. The models were established using a training set constituted of 62 fatty acids. The best QSPR equations are defined by the following equations:

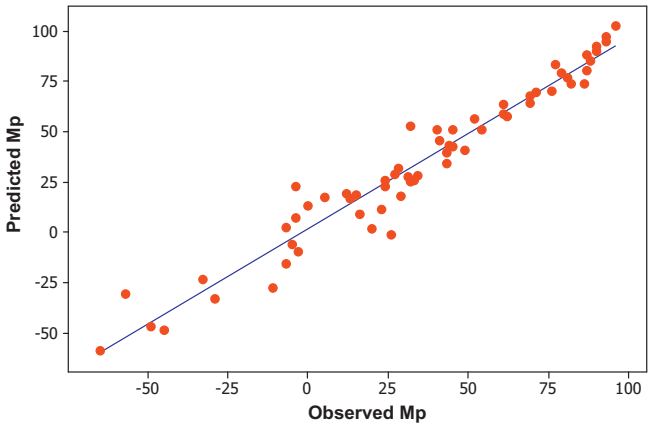


Fig. 2. The plot of experimental versus calculated melting point resulting from model # 2.

Model # 1 (2-descriptors): $R^2 = 0.813$, $R^2_{adj} = 0.807$; $F = 138.75$, SD = 0.017

$$\mathrm{Mp} = -55.5 - 19.9 \times D1 + 0.535 \times (D2) \quad (3a)$$

Model # 2 (5-descriptors): $R^2 = 0.945$, $R^2_{adj} = 0.940$; $F = 197.93$, SD = 0.0994

$$\mathrm{Mp} = -70.1 + 17.6D3 + 37.2 \times D4 - 31.4 \times D5 - 1.70 \times D6 + 267657 \times D7 \quad (3b)$$

The definition of descriptors D1–D7 is given in Table 3. Eq. (3b) shows that MPs is correlated to a combination of molecular descriptors which are essentially Topological and electrostatic nature (see Table 3).

The analysis of statistical data of the models #1–#13 (with two up to fourteen descriptors) and the usage of the breakpoint criterion (Katritzky et al., 2004), show that the best one is model #2 defined by Eq. (3b) involving five parameters. The molecular descriptors involved in this model are

D3 (the Number of aromatic bonds), D4 (Average Complementary Information content (order 1)), D5 (Balaban index), D6 (YZ Shadow) and D7 (min atomic orbital electronic population), According to the t-test values ($|t$-test$|$), the importance of the descriptors involved in the model decreases in the following order: D3 > D4 > D5 > D6 > D7.

The correlation matrix of the five selected descriptors of model # 2 is given Table 3. It turns out that the linear correlation coefficient value between two different descriptors is less to 0.5 Thus; we can assume that the five descriptors are mutually independent in the elaborated QSPR model. We note that the 5-parameter model presents satisfactory statistical data correlation coefficient $R^2 = 0.945$, adjusted squared correlation coefficient $R^2_{adj} = 0.94$ standard deviation SD = 0.09, F-test = 197.

In order validate a QSPR model, two approaches are often used (i) to use of only a part of the available data for building the model, keeping the remaining data points for external validation; (ii) the use of all data points to build the model and to apply as

Table 4
Internal validation of the QSPR model # 2.

Training set	N	$R^2_{(Fit)}$	$R^2_{adj(fit)}$	F	Test set	N	$R^2_{(Pred)}$	$R^2_{adj(Pred)}$	F
A + B	42	0.9478	0.9207	130.62	C	20	0.9615	0.9194	69.85
A + C	41	0.9514	0.9003	137.00	B	21	0.9726	0.6502	106.34
B + C	41	0.9507	0.9359	173.56	A	21	0.9540	0.8146	62.25
Average		0.9499	0.9189	147.06			0.9627	0.7947	79.48

validation method only the internal cross validation procedures. In the present QSPR study, we have used the second approach by applying the following steps: (1) all data points were ordered in the ascending order of MPs values. (2) The parent 62 data points were divided into three subsets (A–C): the first, fourth, seventh, etc. data points comprise the first subset (A), the second, fifth, eighth, etc. comprise the second subset (B), and the third, sixth, ninth, etc. comprise the third subset (C). (3) Three new datasets were constructed using all combinations of the binary sums: (A + B), (A + C) and (B + C). (4) The standard QSPR modeling procedure including best multiple linear regression method (B-MLR) was applied to the three datasets obtained in step 3, i.e. for each training set the correlation equation was derived with the same descriptors corresponding to model # 2. (5) The general model was again validated using classical internal cross validation procedures: leave many-out. The procedure described above was applied to the complete data set of 62 points. Three training subsets are constructed with 42 compounds and the remaining 20 compounds were used as external validation datasets. The efficiency of QSPR models to predict MPs values was also estimated using the adjusted R^2. The average values of adjusted $R^2_{(Fit)}$ and $R^2_{(Pred)}$ are very close (0.94 and 0.91, respectively), Table 4 which suggests relatively stable predictivity of the proposed QSPR model.

5. Conclusion

In summary, we have used the best multilinear regression method implemented in Codessa software for elaborating reliable QSPR models capable to predict melting point of fatty acids. The present study shows that the best QSPR model is defined in terms of five noncorrelated molecular descriptors which are essentially electrostatic and topological parameters. Therefore, the elaborated QSPR model could be used to predict the melting point of new fatty acids or for fatty acids for which the experimental melting is unavailable in the literature.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.chemphyslip.2011.10.001.

References

Bailey, A.E., 1950. Melting and Solidification of Fats. Interscience Publishers, Inc., New York.

Basak, S.C., Mills, D., 2005. Prediction of partitioning properties for environmental pollutants using mathematical structural descriptors. ARKIVOC, 60–76.

Becke, A.D., 1988. Phys. Rev. A 38 3098, 1993, J. Chem. Phys. 98 1372, 1993, J. Chem. Phys. 98 5648.

Chemdraw is a molecule editor © 2008. Available from: http://www.cambridgesoft.com/.

David R. Lide, ed., CRC Handbook of Chemistry and Physics, Internet Version 2009, 89th Edition. CRC Press.

Devillers, J.A.T., 1999. In: Balaban (Ed.), Topological Indices and Related Descriptors in QSAR and QSPR. Gordon and Breach, The Netherlands.

Frisch, M.J., et al., 2007. Gaussian 03, Revision C.02. Gaussian Inc., Wallingford, CT.

Ghotra, B.S., Dyal, S.D., Narine, S.S., 2002. Lipid shortenings: a review. Food Res. Int. 35, 1015–1048.

Cook, H.W., McMaster, C.R., 2002. Biochemistp 3' q/Lipid, Lipoproleins am/Membralle, 4th ed.

Hernqvist, L., 1984. On the structure of triglycerides in the liquid state and fat crystallization. Fett. Seifen Anstrichm. 86, 297–300.

Humphrey, K.L., Moquin, P., Narine, S.S., 2004. Phase behavior of a binary lipid shortening system: from molecules to rheology. J. Am. Oil Chem. Soc. 80, 1175–1182.

Karelson, M., 2000. Molecular Descriptors in QSAR/QSPR. Wiley-Interscience, New York.

Karelson, M., Maran, U., Wang, Y., Katritzky, A.R., 1999. QSPR and QSAR models derived using large molecular descriptor spaces. A review of CODESSA applications. Collect. Czech. Chem. Commun. 64, 1551–15570.

Katritzky, A.R., Lan, Mu., Karelson, M., 1996a. A QSPR study of the solubility of gases and vapors in water. J. Chem. Inf. Comput. Sci. 36 (6), 1162–1168.

Katritzky, A.R., Dobchev, D., Hur, E., Fara, D., Karelson, M., 2005a. QSAR treatment of drugs transfer into human breast milk. Bioorg. Med. Chem. 13, 1623–1632.

Katritzky, A.R., Dobchev, D., Fara, D., Karelson, M., 2005b. QSAR studies on 1-phenylbenzimidazoles as inhibitors of the platelet-derived growth factor. Bioorg. Med. Chem. 13, 6598–6608.

Katritzky, A.R., Karelson, M., Lobanov, V.S., 1997a. QSPR as a means of predicting and understanding chemical and physical properties in terms of structure. Pure Appl. Chem. 69, 245.

Katritzky, A.R., Lan, Mu., Lobanov, V.S., 1996b. Correlation of boiling points with molecular structure. 1. A training set of 298 diverse organics and a test set of 9 simple inorganics. J. Phys. Chem. 100, 10400–10407.

Katritzky, A.R., Lobanov, V.S., Karelson, M., 1995. QSPR the correlation and quantitative prediction of chemical and physical properties from structure. Chem. Soc. Rev. 24, 279–285.

Katritzky, A.R., Lobanov, V.S., Karelson, M., 1997b. Codessa: Comprehensive Descriptors for Structural and Statistical Analysis, User Manual. University of Florida, Gainesville, Florida.

Katritzky, A.R., Taemm, K., Kuanar, M., Fara, D., Oliferenko, A., Oliferenko, P., et al., 2004. Aqueous biphasic systems. Partitioning of organic molecules: a QSPR treatment. J. Chem. Inf. Comput. Sci. 44, 136–142.

Minitab is statistical software. Minitab Inc ©, 2010. Available from: www.minitab.com/support.

Moziar, C., deMan, J.M., deMan, L., 1989. Effect of tempering on the physical properties of shortening. Can. Inst. Food Sci. Technol. J. 22, 238–242.

Narine, S.S., Marangoni, A.G., 1999. Relating structure of fat networks to mechanical properties: a review. Food Res. Int. 32, 227–248.

Narine, S.S., Marangoni, A.G., 2002. Structure and mechanical properties of fat crystal networks. Adv. Food Nutr. Res. 44, 33–145.

Paul, D., Huibersa, T., Dinesh, 1., Shaha, O., Katritzky, A.R., 2002. Predicting Surfactant Cloud Point from Molecular Structure.

Stefano, P., Erminia, B., 2007. QSAR analysis on antimicrobial peptides. Chem. Phys. Lipids 149 (1), S87.

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

Stewart, J.J.P., 2008. MOPAC2009. Stewart Computational Chemistry. Colorado Springs, CO, USA, Available from: http://openmopac.net/MOPAC2009.html.

Svenstrup, G., Brüggemann, D., Kristensen, L., Risbo, J., Skibsted, L.H., 2005. The influence of pretreatment on pork fat crystallization. Eur. J. Lipid Sci. Technol. 107, 607–615.

Thakur, A., 2005. QSAR study on benzenesulfonamide dissociation constant pKa: physicochemical approach using surface tension. ARKIVOC 14, 49–58.

- ملخص -

العمل المقدم في هذه الأطروحة يهدف إلى تطوير نماذج *QSPR* موثوقة بها ومستقرة للتنبؤ بخصائص فيزيوكيميائية:

1- درجة انصهار سلسلة من الأحماض الدهنية باستخدام الانحدار الخطي المتعدد ومجموعة من الواصفات الجزيئية دو الطبيعة الطوبوغرافية والالكتروستاتيكي.

2- ثوابت الحموضة *pKa* لأحماض البنزويك المائية باستخدام الانحدار الخطي البسيط وواصفة جزيئية كمية واحدة وهي طاقة غيبس الحرة ΔG_{aq} . تحسب على المستوى *B3LYP/6-311++G*** باستخدام الدورتين الحراريين. كما أدرج تأثير المذيبات مع نموذج SMD. وللتحقق من الاستقرار الداخلي من النماذج تم استخدام العديد من التقنيات، وهي *cross-validation*، *Leave-one-out* و *Y-Randomization* . كما تم إختبار القدرة التنبؤية بنجاح باستخدام مجموعة اختبار خارجي التي لم تدرج في السلسلة الأصلية المستخدمة لتطوير النموذج "مجموعة التدريب".

الكلمات المفتاحية: نماذج QSPR ; الأحماض الدهنية ;الاحماض البنزويك.

Résumé

Le travail présenté dans cette thèse a pour objectifs d'élaborer des modèles QSPR fiables, stables et prédictifs pour la prédiction des propriétés physico-chimiques :

1- Températures de fusion d'une série acides gras en utilisant la régression linéaire multiple et d'une combinaison de descripteurs moléculaires de nature topologique et électrostatique

2- Les constantes d'acidité pKa des acides benzoïques en phase aqueuse en utilisant la régression linéaire simple et un seul descripteur moléculaire quantique qui est l'enthalpie libre de déprotonation ΔG_{aq} calculé au niveau *B3LYP/6-311++G(d,p)* en utilisant deux cycles thermodynamiques. L'effet de solvant a été pris en considération avec le modèle SMD. La stabilité interne des modèles élaborés a été confirmée par plusieurs techniques, à savoir, la validation croisée et la technique dite du *Leave-one-out* et la *Y-Randomization*. Le pouvoir prédictif a été testé avec succès avec une série de test non incluse dans la série d'apprentissage.

Mots clés : modèles QSPR ; Acides gras ; Acides benzoïques.

Abstract

The work presented in this thesis aims to develop reliable, stable and predictive QSPR models for the prediction of physico-chemical properties:

1- Melting points of a series of fatty acids using multiple linear regression and combination of molecular descriptors, which are topological and electrostatic nature.

2- The acidity constants pKa of benzoic acids in aqueous phase using simple linear regression and a single quantum molecular descriptors which is the free enthalpy of deprotonation ΔGaq , calculated at the B3LYP/6-311++G(d,p) level of theory, using two thermodynamic cycles. The solvent effect was included with the SMD model. The internal stability of the developed models was verified by several techniques, namely, the cross-validation, the Leave-one-out, and the Y-Randomization. The predictive ability was further performed by using the external test set composed of data not used to develop prediction 'training set'.

Keywords: QSPR models; Fatty acids; Benzoic acid.