MINISTERE DE L'ENSEIGNEMENT SUPERIEUR ET DE LA RECHERCHE SCIENTIFIQUE
UNIVERSITE «ABBES LAGHROUR» DE KHENCHEL

AFACULTE DES SCIENCES ET DE LA TECHNOLOGIE

DEPPARTEMENT SCIENCE DE LA MATIERE

N° de série : …

Mémoire de fin d'études

Pour l'obtention du diplôme de Master (L.M.D)

Spécialité : Chimie Analytique et Environnement

Intitulé:

PRÉDICTION DE LA TEMPÉRATURE D'ÉBULLITION D'UNE SÉRIE D'ALDHYDES

Réalisé par :

- **ASEFSOU Sarra**
- **AZIZI Malak**

Dirigé par : **KERTIOU Noureddine**

Membres de jury:

-	**BEDGHIOU Djohra**	**MCB**	***Présidente***
-	**BOUAKKADIA Amel**	**MCA**	**Examinatrice**

Année Universitaire 2022-2023

Remerciement

Nous remercions tout d'abord **Allah** pour nous

Avoir donné la santé, la volonté, la force, et le courage

Nous remercions également notre cher et excellent encadreur, le **Dr. KERTIOU Noureddine** Pour son aide, son suivi, son support, ses encouragements, et aussi ses précieux conseils.

Nous avons été heureux de travaille avec lui.

Nous tenons à remercier aussi les membres de jury d'avoir Accepté d'examiner ce travaille et de lui apporter leur soutien

Et sans oublier tous les professeures de chimie de notre université pour leur contribution et en nous aidant à Surmonter les étapes les plus difficiles.

Nos profonds remerciements à nos parents de nous

Avoir soutenu moralement et financièrement durant

Ces longues années

Dédicaces

Je dédie ce travail à…

A mon père ABD ELAZIZ

A ma très chère maman NADJIBA

Qui ont toujours été là pour moi, Leur soutien inconditionnel et leurs encouragements ont été d'une grande aide

À mes chers frères

ACHREF, WAIL, MONDHIR

A ma chère sœur

MIRALE

A mes amis proches

SABRINA, SABAH, SARRA

Et tous mes amis en souvenir de ces moments agréables passé ensembles Surtout mon amie, et mon binôme : «SARRA»

MaLaK

Dédicaces

Avec l'expression de ma reconnaissance, je dédie ce modeste travail à ceux qui, quels que soient les termes embrassés, je n'arriverais jamais à leur exprimer mon amour sincère

A mon cher père, **ALi**

A ma chère mère, **Fouzia**

Qui n'ont jamais cessé, de formuler des prières à mon égard, de me soutenir

Et de m'épa ler pour que je puisse atteindre mes objectifs.

A mes très chers frères,

Youcef, Abd eldjalil et **Youness**

A mes chères sœurs

Sanna et **Salma**

Pour ses soutiens moral, Puisse Dieu vous donne santé, bonheur, courage et surtout réussite

A toute **ma famille,**

Et tous mes amis en souvenir de ces moments agréables passé ensembles

Surtout mon amie, et mon binôme : **MaLaK**

SOMMAIRE

Partie théorique

Partie Application

LISTE DES TABLEAUX

LISTE DES FIGURES

SYMBOLES ET ABREVIATIONS

Symboles	Définitions
ACP	Analyse en composantes principales.
AG	Algorithme génétique (Genetic Algorithm).
Du	D total accessibility index / unweighted
ei	Différence entre les valeurs observées et estimées.
ei std	Résidu de prédiction standardisé.
F	Statistique de Fisher
FIV	Facteur d'inflation de la variance.
H	Matrice de projection, ou matrice chapeau.
hii	Eléments diagonaux de la matrice chapeau.
k	Nombre de descripteurs.
LMO	leave – many- out.
LOO	leave – one – out.
MCO	Les moindres carrés ordinaires
MCP	Les moindres carrés partiels.
MLR	Régression linéaire multiple.
MM+	Mécanique Moléculaire.
n	Dimension de la population.
n-p	Nombre de degrés de liberté.
p	Nombre de descripteurs en comptant la constante (Nombre de paramètres).

PLS	Moindres carrés partiels.
PRESS	Somme des carrés des erreurs de prédiction.
Q^2	Coefficient de prédiction.
Q^2ext	Coefficient de prédiction externe
QSAR	Quantitative Structure/ Activity Relationships. Relations Structure/Activités Quantitatives).
QSPR	Quantitative Structure/ Propriety Relationships. Relations Structure/Propriétés Quantitatives).
R^2	Coefficient de détermination.
RMSE	Root Mean Squared Error.
RNA	Réseau de neurones artificiel.
S	Erreur standard.
SCE	Somme des carrés des écarts
SCT	Somme des carrés totale.
SDEC	Standard Deviation Error in Calculation : Déviation standard de l'erreur calculée.
SDEP	Standard Deviation Error of Prediction : Déviation standard de l'erreur de prédiction.
SDEPext	External Standard Deviation Error of Prediction: Déviation standard de l'erreur de prédiction externe.
t	t de Student.
Teb	Température d'ébullition
X	Matrice des valeurs observées des variables explicatives.
X'	Matrice transposée de X.
y	Vecteur de dimension n.
yi	Valeur observée.
$\hat{y}$i	Valeur estimée.

Introduction générale

Introduction générale

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

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

Le développement technologique et industriel rapide au cours des dernières décennies vise à accroître le bien-être des humanités. Cependant, elle n'est pas sans conséquences pour la santé humaine et l'environnement. C'est à cause des produits chimiques différents types de développement, ce qui provoque la fuite de substances toxiques dans l'environnement.

Suite au développement de l'informatique et de l'existence sur le marché de logiciels professionnels adaptés peu coûteux et rapide, il y a une croissance exceptionnelle de l'intérêt pour les relations quantitatives structure-Activité (QSAR), qui utilisent des méthodes d'analyses multidimensionnelles afin de modéliser des Activité en fonction des paramètres structuraux moléculaires (appelés descripteurs).

QSPR (Quantitative structure Propriety Relationship) est parmi les nouvelles techniques de modélisation, mettant en jeu des relations de structure avec la propriété. Elles prennent de plus en plus d'importance dans les études de la conception des précurseurs. Un outil qui permet une prédiction rapide d'une ou plusieurs propriétés biologique. Elle a pu être mise en place dans les laboratoires et utilisée dans l'industrie pharmaceutique. L'objectif d'une modélisation QSPR est de trouver des modèles précis, applicables et robustes afin de trouver une relation entre la structure et la propriété dans un but de prédiction mais également d'interprétation.

Les programmes de modélisation moléculaire sont parmi les programmes les plus importants qui contribuent à l'imagerie du composé chimique sous la forme de tridimensionnalité et donne les propriétés des molécules, et dans cette étude, nous avons utilisé un ensemble de logiciels, qui est tridimensionnalité et nous donne les propriétés des molécules, et dans cette étude, nous avons utilisé un ensemble de logiciels, qui est (HyperChem) et (Dragon) pour la modélisation des

molécules et le calcul des spécifications, ainsi que les programmes mathématiques (Mobydigs) et (minitab) pour la construction des modèles et les calculs des paramètres statistiques .

Dans ce travail, nous nous sommes intéressés à la température d'ébullition des aldéhydes en utilisant le calcul d'un modèle de régression linéaire et la validation externe.

Ce mémoire comporte en plus de la bibliographie, d'une introduction et d'une conclusion générale, deux grandes parties :

Dans la Partie Généralités, nous avons développé tout ce qui a trait au pré-traitement des molécules (introduction des molécules, optimisation de leur géométrie) en vue du calcul des descripteurs moléculaires à l'aide de différents logiciels de modélisation moléculaire, et le traitement statistique pour l'évaluation de la qualité de l'ajustement (robustesse des modèles ; détection des observations aberrantes; test de randomisation; validation externe).

Dans la Partie Application, nous présentons et discutons le modèle calculé.

Partie théorique

I- INTRODUCTION

QSAR (relation quantitative structure-activité) et QSPR (relation quantitative structure-propriété) sont des outils analytiques importants dans les domaines de la chimie bio-organique, industrielle et environnementale. Ces outils visent à comprendre la relation entre la structure des composés chimiques et leurs propriétés physiques, chimiques et biologiques.

L'idée centrale de QSAR/QSPR est d'utiliser les données existantes pour prédire les propriétés de nouveaux composés qui n'ont pas encore été testés. Ces méthodes reposent sur l'utilisation de modèles mathématiques et statistiques pour analyser les données et prédire de nouvelles propriétés pour les composés chimiques.

L'objectif principal du développement de modèles QSAR/QSPR est de fournir un moyen efficace et économique de prédiction. Ces méthodes ont été utilisé avec succès dans la conception de médicaments, le développement de nouveaux produits chimiques et dans les domaines des sciences environnementales.

Le défi de QSAR/QSPR consiste à déterminer les paramètres structuraux clés liés aux propriétés à prédire et à développer des modèles mathématiques appropriés pour prédire ces propriétés avec une grande précision.

Le défi de QSAR/QSPR consiste à déterminer les paramètres structuraux clés liés aux propriétés à prédire et à développer des modèles mathématiques appropriés pour prédire ces propriétés avec une grande précision. Ainsi, cela peut aider à développer de nouveaux produits chimiques et médicaments plus rapidement et plus efficacement. [1]

I.1 Historique de (QSAR)

Il y a plus d'un siècle et demi, en 1863, Cros [2] a observé que le point d'ébullition et le point de fusion des alcanes augmente avec le nombre d'atomes de carbone et la masse moléculaire. Il a observé également une diminution de la solubilité dans l'eau des alcools avec l'augmentation du nombre d'atomes de carbone et la masse moléculaire, cela est considéré depuis comme la première formulation générale en QSAR.

Cinq ans après, en 1868, Crum – Brown et Fraser [3] postulèrent que «l'activité biologique d'une molécule est une fonction de sa constitution chimique ».

Quelques décennies plus tard, en 1893, Richet [4] a montré que la cytotoxicité de certains composés organiques était inversement proportionnelle à leur solubilité dans l'eau.

A la fin du 19ème siècle, Meyer en 1899 et Overton en 1901 [5-7] ont indépendamment observé « une relation linéaire entre l'activité des narcotiques et leur coefficient de partage huile-eau ».

Six ans après, en 1907, Fühner et Neubauer [8] ont montré pour une série de narcotiques homologues, que l'activité augmentait en fonction de la progression géométrique de la série de composés, ceci montrant l'importance de la contribution d'additivité de groupements fonctionnels pour l'activité biologique.

En 1962, Hansen [9] a montré l'existence d'une corrélation entre la toxicité des acides benzoïques substitués et les constantes électroniques « σ » des substituants.

L'année 1964 est considérée comme le début des méthodes QSAR modernes. Hansch et Fujita ont établi les premières corrélations entre les propriétés physico-chimiques (log P, pKa, paramètres stériques et électroniques) et l'activité biologique (activité enzymatique, pharmacologique), Ces méthodes seront appelées par la suite l'analyse deHansch et l'analyse de Free Wilson) [10-11]. Sept ans plus tard, Hansch et Lien ont réalisé une étude QSAR sur différentes familles d'antifongiques : benzoquinones, sels d'alkylpyridinium, imidazoles et phénols. Ils ont observé que quels que soient la famille et le champignon utilisé, l'activité antifongique dépend du coefficient de partage Eau-Octanol, expérimental ou calculé [12].

Ces études ont été extrapolées aux techniques séparatives en corrélant les propriétés physico-chimiques des analytes avec les temps de rétention obtenus expérimentalement : c'est l'étude quantitative des relations structure temps de rétention noté QSAR [13].

I.2 **Définition de QSAR/QSPR**

Une Relation Quantitative Structure à Activité/Propriété (en anglais :Quantitative Structure-Activity/Propriety Relationship) est le procédé par lequel une structure chimique est corrélée avec un effet bien déterminé comme l'activité biologique ou la réactivité chimique. Ainsi, l'activité biologique peut être exprimée de manière quantitative, comme pour la concentration de substance nécessaire pour obtenir une certaine réponse biologique. De plus lorsque les propriétés ou structures physicochimiques sont exprimées par des chiffres, on peut proposer une relation mathématique, ou Relation Quantitative Structure à Activité, entre les deux. L'expression

mathématique obtenue peut alors être utilisée comme moyen prédictif de la réponse biologique pour des structures similaires.

La QSAR la plus commune est de la forme : activité = f (propriétés physico-chimiques et/ou structurales).

Par définition, Une QSAR est une modèle mathématique qui associe un ou plusieurs paramètres quantitatifs dérivés de la structure chimique, à une mesure quantitative d'une activité [14].

I.3 QSAR/QSPR

Le concept de SAR (relation structure-activité) (relation entre structure et activité activités) tentent d'établir des relations entre différents traits de comportement Composés (par exemple, activité et toxicité) et leurs informations structure chimique, en d'autres termes, les études SAR fournissent Équations impliquant une activité/propriétés/toxicité spécifique des produits chimiques En utilisant des informations sur leur structure chimique et ainsi exprimer quantitativement L'activité des produits chimiques définit une étude QSAR. Axiome central de la modélisation QSAR est basé sur la présentation des réponses chimiques en termes de propriétés moléculaires, sachant que chaque propriété contenant des informations chimiques importantes peut être utilisée comme descripteur. Une fois les équations établies, la méthode QSAR permet de prédire l'activité du produit chimique étudié, de plus, la méthode QSAR met l'accent sur la modification de la structure chimique pour obtenir les produits chimiques d'intérêt avec les valeurs de réponse désirées. L'appellation est influencée par le point final modélisé, par conséquence le processus peut être défini comme suit : QSAR/QSPR/QSTR pour activité/propriété/toxicité.

L'équation mathématique : Réponse = f (propriétés structurale/chimiques) [15].

L'analyse QSAR a été créé afin de remplir les objectifs suivants : (a) la prédiction de nouveaux analogues avec une meilleure activité, (b) améliorer la compréhension et l'investigation du mode d'action de produits chimiques et pharmaceutiques, (c) l'optimisation de la molécule type en congénères moins toxiques, (d) la rationalisation des expérimentations humides (QSAR offre une alternative économique et rapide aux essais in vitro à débit moyen ainsi qu'aux essais in vivo à faible débit) [15]

I.4 Principe

Le principe des méthodes QSPR/QSAR (Figure-01) est d'établir une relation mathématique reliant de manière quantitative des propriétés moléculaires, appelées descripteurs, avec une observable macroscopique (activité biologique, toxicité, propriété physico-chimique, etc.), pour une série de composés chimiques similaires à l'aide de méthodes d'analyses de données.

La forme générale d'un tel modèle est la suivante :

Propriété/ Activité = f (D1, D2,… Dn,)

D1, D2, …Dn sont des descripteurs des structures moléculaires.

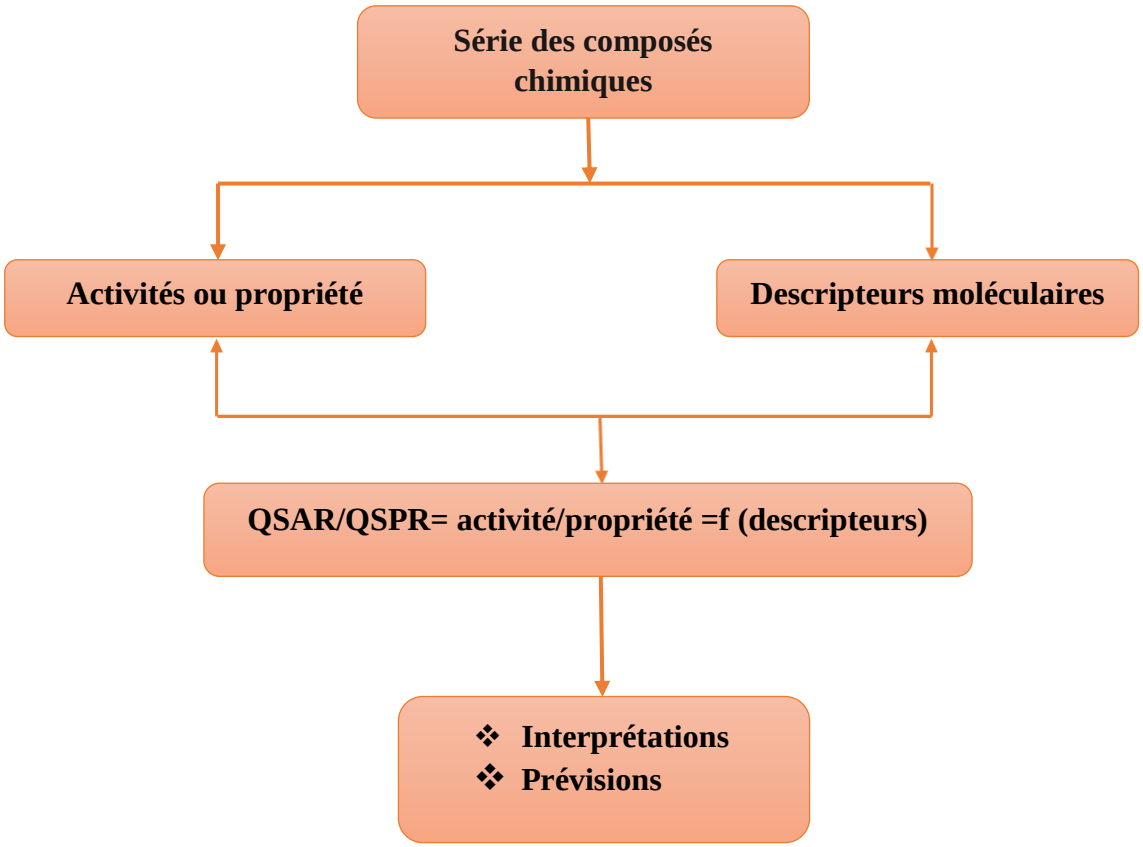


Figure-01 *: Modèle de l'étude de relation structure activité.*

L'objectif de ces études est d'analyser les données structurales afin de détecter les facteurs déterminants pour l'activité ou la propriété étudiée. Pour ce faire, différents types de méthodes statistiques peuvent être employées. L'expression mathématique obtenue peut alors être utilisée comme moyen prédictif de l'activité étudiée pour de nouvelles molécules ou des molécules pour lesquels les données expérimentales ne sont pas disponibles.

I.5 **Méthodologie générale d'une étude QSPR/QSAR [16]**

L'approche générale d'une étude QSAR/QSPR est la suivante :

a. Construire des bases de données à partir de mesures expérimentales fiables en déterminant Les propriétés ou activités de chaque composé.
b. Randomiser cette base de données en une série d'apprentissage (ensemble de calibration) contient généralement 2/3 de la base de données et une série de de test (ensemble de validation) se compose du 1/3 restant
c. Sélectionner les descripteurs pertinents à la propriété ou à l'activité à l'étude.
d. Construire des modèles mathématiques à l'aide des séries d'apprentissage.
e. caractérise le modèle développé par son indice de validation interne et Vérifier leur robustesse par des tests.
f. Valider les modèles élaborés en utilisant la série de test et calculer leurs paramètres statistiques de validation externe.
g. Elaborer le domaine d'applicabilité du modèle retenu.
h. Explorer et exploiter les modèles validés pour comprendre les mécanismes et les modes d'action.

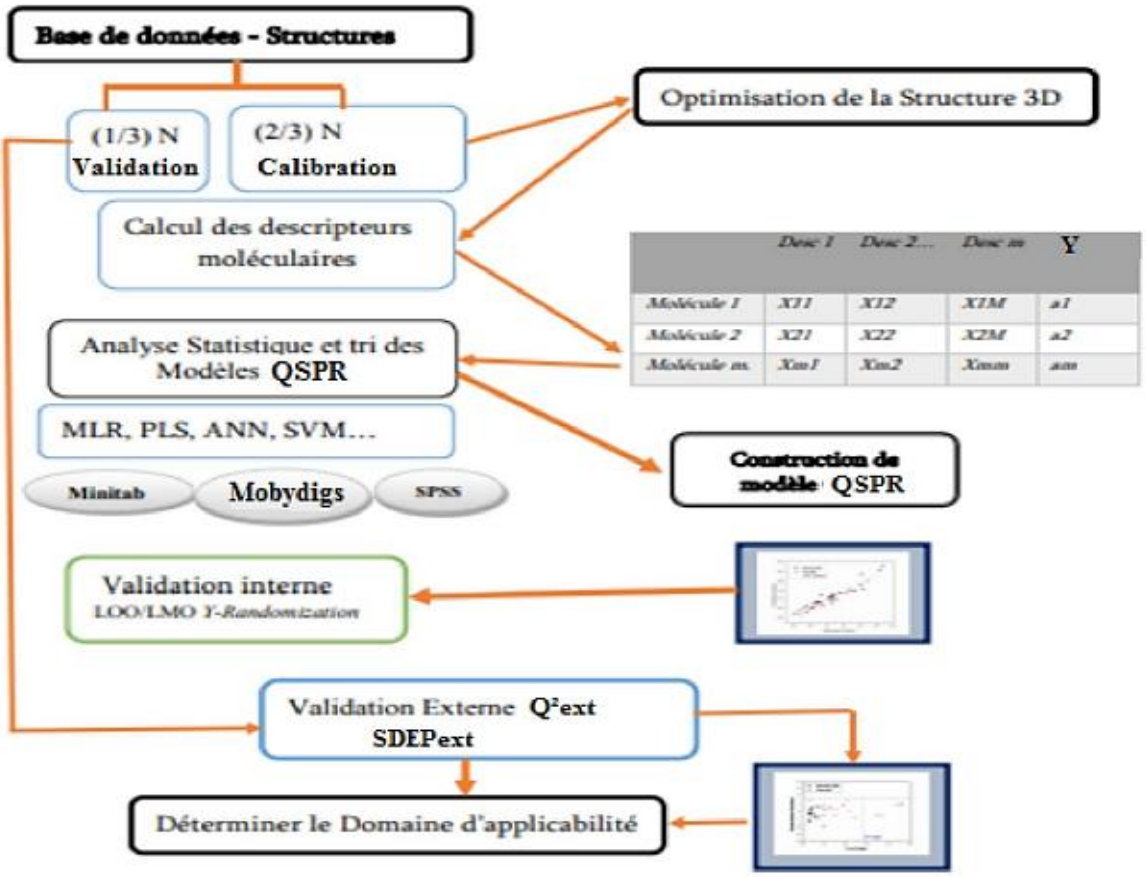


Figure -02 *: Procédure d'obtention et de validation d'un modèle QSPR/QSAR*

I.6 Les applications de l'étude QSAR

- Ces modèles ont de nombreuses applications, par exemple (zynski ; 2010)
- Optimisation de l'activité pharmacologique.
- Conception rationnelle de nombreux autres produits tels que les tensioactifs, Parfums, colorants et chimie fine.
- Identifiez les composés dangereux dès le début du développement Produits ou la projection des stocks de composés existants.
- Prédire la toxicité et les effets secondaires des nouveaux composés.
- Prédire la toxicité des espèces environnementales.
- Sélectionner le composé avec le meilleur profil pharmacocinétique, que ce soit la stabilité ou la disponibilité dans les systèmes biologiques.

- ❖ Prédire diverses propriétés physico-chimiques des molécules.
- ❖ Prédire le devenir des molécules rejetées dans l'environnement.
- ❖ Prédire les effets combinés des molécules, que ce soit dans des mélanges ou des formulations. [17]

I.7 Les méthodes mathématiques utilisés par le model QSPR

Les méthodes utilisées en QSPR sont deux types :

a) Linéaires

- ❖ Régression linéaire simple
- ❖ Régression linéaire multiple MLR.
- ❖ Régressions aux moindres carrées partielles (PLS).

b) Non linéaires

- ❖ Réseau de neurones artificiel RNA.
- ❖ SVM. Arbres de décision.

I.8 La relation structure- propriété quantitative (QSPR)

Les relations de propriété de structure quantitative (QSPR) sont des modèles prédictifs permettant de calculer les propriétés de composés chimiques à partir de leurs seules structures moléculaires. Ils reposent sur une relation mathématique entre la propriété et des descripteurs de la structure moléculaire des composés ciblés. Si ces modèles ont été largement utilisés pour la prédiction d'activités biologiques en toxicologie ou en pharmacologie depuis de nombreuses années, ils trouvent aujourd'hui de plus en plus d'applications pour la prédiction de propriétés physico-chimiques. L'essor de ces approches s'est même récemment accentué avec la mise en place du règlement européen REACH qui recommande l'emploi de tels modèles, une fois validés, pour l'acquisition des données nécessaires à l'enregistrement des substances chimiques [18].

I.9 COLLECTE DES DONNEES

La température d'ébullition est l'un des grandeurs physiques importantes. Les données utilisées dans ce travail, concernent 77 composés, sont réunies Dans le tableau -01-: On a choisi aléatoirement (23) composés pour validation ; et le reste (54) pour la calibration ou à la construction du modèle.

Tableau- 01 *Nomenclature et valeurs de propriété des Aldéhydes étudiés.*

N	Composé	Teb(K)
1	1-Naphthaldehyde	592
2	2,3,4-Trihydroxybenzaldehyde	575
3	2,3,5-Trichlorobenzaldehyde	542.5
4	2,3-Dihydroxybenzaldehyde	513
5	2,4,5-Trimethoxybenzaldehyde	584.5
6	2,4,6-Trihydroxybenzaldehyde	607.7
7	2,4-Dichlorobenzaldehyde	506
8	2,4-Dihydroxybenzaldehyde	624
9	2,4-Dimethoxybenzaldehyde	581
10	2,5-Dihydroxybenzaldehyde	486.5
11	2-Anisaldehyde	516.5
12	2-Bromobenzaldehyde	504
13	2-Chloro-3-hydroxy-4-methoxybenzaldehyde	564.3
14	2-Chloro-4-hydroxycarboxaldehyde	545.5
15	2-Chloro-5-nitrobenzaldehyde	566.5
16	2-Chloro-6-fluorobenzaldehyde	631
17	2-Chlorobenzaldehyde	482
18	2-Fluorenecarboxaldehyde	564
19	2-Fluorobenzaldehyde	514.5
20	2-Hydroxy-1-naphthaldehyde	579
21	2-Hydroxy-3-nitrocarboxaldehyde	569

Tableau - 01 *(suite)*

N	Composé	Teb (K)
22	2-Hydroxybenzaldehyde	470
23	2-Methyl-1-naphthaldehyde	589.5
24	2-Nitrobenzaldehyde	565
25	2-Tolualdehyde	473
26	3,4,5-Trihydroxybenzaldehyde	651.4
27	3,4-Dihydroxybenzaldehyde	596
28	3,4-Dimethoxy-5-hydroxycarboxaldehyde	626
29	3,5-Dibromo-4-hydroxycarboxaldehyde	546.3
30	3,5-Dibromosalicylaldehyde	534.5
31	3-Anisaldehyde	504
32	3-Bromo-4-hydroxycarboxaldehyde	534.5
33	3-Bromobenzaldehyde	508.5
34	3-Chloro-2-fluoro-5-(trifluoromethyl) benzaldehyde	469
35	3-Chlorobenzaldehyde	486.5
36	3-Cyanobenzaldehyde	484
37	3-Ethoxy-2-hydroxycarboxaldehyde	537
38	3-Ethoxy-4-hydroxybenzaldehyde	558
39	3-Fluorobenzaldehyde	467
40	3-Hydroxy-4-methoxybenzaldehyde	581
41	3-Hydroxy-4-nitrobenzaldehyde	576.5
42	3-Hydroxybenzaldehyde	514
43	3-Methoxy-4-hydroxybenzaldehyde	604
44	3-Methoxysalicylaldehyde	538.5
45	3-Nitrobenzaldehyde	560
46	3-Tolualdehyde	496
47	4-(Dimethylamino)benzaldehyde	583
48	4-(Pentyloxy)benzaldehyd	675

Tableau - 01 *(suite)*

N	Composé	Teb(K)
49	4,6-Dimethoxy-2-hydroxybenzaldehyde	643.5
50	4-Acetamidobenzaldehyde	657.5
51	4-Anisaldehyde	521
52	4-Biphenylcarboxaldehyde	599.6
53	4-Bromobenzaldehyde	517
54	4-Butoxybenzaldehyde	558
55	4-Chlorobenzaldehyde	486.5
56	4-Cyanobenzaldehyde	563
57	4-Ethoxybenzaldehyde	528
58	4-Ethylbenzaldehyde	494
59	4-Fluorobenzaldehyde	455
60	4-Hydroxy-1-naphthaldehyde	638
61	4-Hydroxy-3-nitrobenzaldehyde	548
62	4-Hydroxybenzaldehyde	583
63	4-Isopropylbenzaldehyde	576
64	4-Methyl-1-naphthaldehyde	533
65	4-Nitrobenzaldehyde	573
66	4-Phenoxybenzaldehyde	593
67	5-Bromosalicylaldehyde	520.5
68	5-Bromovanillin	625
69	5-Chlorosalicylaldehyde	490.5
70	5-Hydroxy-2-nitrobenzaldehyde	646
71	6-Chloro-2-fluoro-3-methylbenzaldehyde	522
72	Benzaldehyde	540
73	Pentafluorobenzaldehyde	439
74	Phenanthrene-9-carboxaldehyd	678.5

Tableau - 01 *(suite et fin).*

N	Composé	Teb(K)
75	Phenyl-1,3-dialdehyde	519
76	p-Tolualdehyde	477
77	Terephthaldicarboxaldehyde	519.5

II- Optimisation de la géométrie moléculaire et génération des descripteurs moléculaires

II.1 Préparation de base des données :

II.1.1 Calcul du modèle :

Les molécules ont été représentées en utilisant le logiciel ChemDraw (ChemDraw ultra 7.0) et optimisées à l'aide du logiciel HyperChem [19]. Un ensemble de plus de 1600 descripteurs moléculaires a été calculé à l'aide du logiciel informatique Dragon [20]. L'ensemble de données a été aléatoirement divisé en deux sous-ensembles, où 75 % de tous les composés ont été utilisés pour construire le modèle et 25 % ont été réservés pour la validation externe.

En utilisant l'algorithme génétique de la version MobyDigs [21], plusieurs modèles ont été générés et évalués en fonction de leurs performances. Parmi les 100 modèles générés par l'algorithme génétique, le modèle sélectionné présente les meilleures statistiques.

II.1.2 Logiciels « ChemDraw»

ChemDraw fournit des chimistes avec un ensemble d'outils riche, faciles à utiliser pour créer des publications prêt, dessins scientifiquement significatifs de molécules et de réactions [22] .Essayant de concevoir et dessiner de nouvelles molécules en utilisant les différents outils disponibles dans ChemDraw, et de les visualiser, et plus important encore, les enregistrer dans différents formats. ChemDraw est très pratique pour les réactions d'écriture à l'aide de produits chimiques. Vous devriez être en mesure de réaliser des structures dans différents formats disponibles [23]. **Exemple** : Phenyl-1,3-dialdehyde

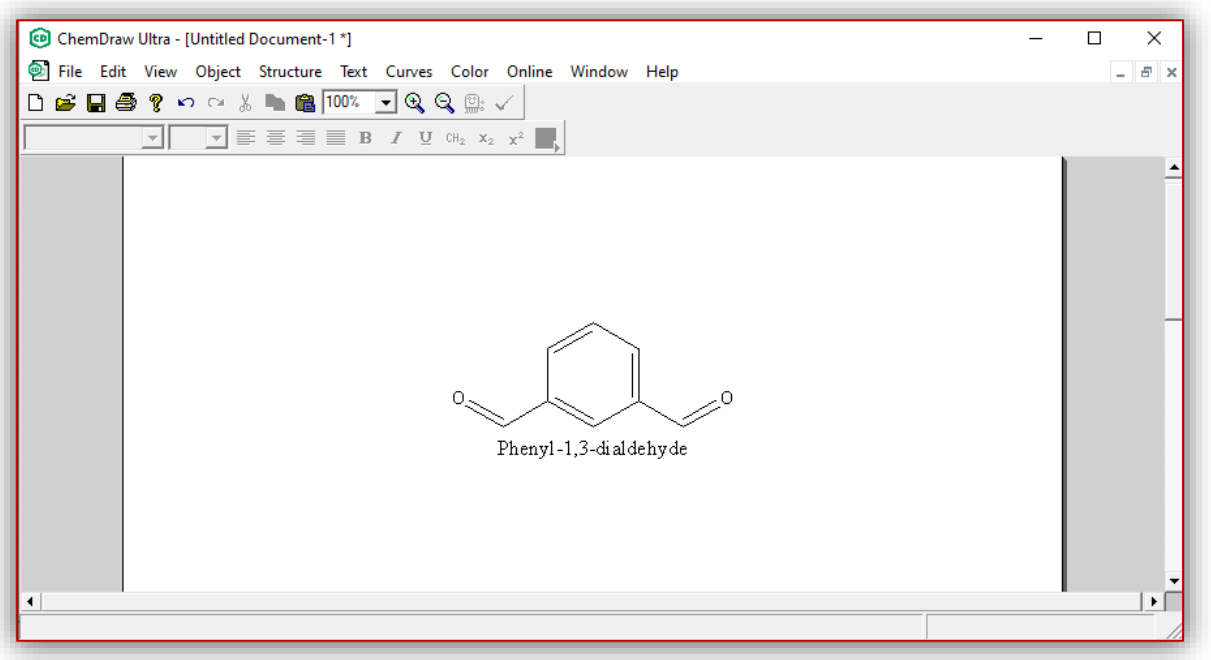


Figure-03 : Représentation des molécules par le ChemDraw.

II.1.3 Le logiciel hyperchem professionnel

HyperChem est un logiciel de modélisation moléculaire chimique développé par Hypercube Inc. Il fonctionne en unissant la visualisation et l'animation 3D avec des calculs de chimie quantique, la mécanique moléculaire et dynamique. Il est facile et flexible.

HyperChem est utilisé dans cette étude pour construire et optimiser les molécules, est enregistrée comme un fichier nommé "Hin" après l'optimisation. Nous avons utilisé la méthode semi empirique MM+ pour l'optimisation [24]. On a 77 molécules donc on obtient 77 fichiers Hin, en suite on calcule les descripteurs moléculaires à partir de ces fichiers par le logiciel Dragon.

II.2 Récupération et stabilisation les molécules de fichier Hin :

II.2.1 Stabilisation structure des molécules (minimisation de l'énergie) :

Pour stabiliser la structure de chaque molécule ou minimisation de l'énergie (ou de la géométrie d'optimisation) on utilise l'HyperChem, en utilisant une variété de méthodes de calcul. Les deux mécanismes moléculaires et les méthodes semi-empiriques sont disponibles. Minimisation de

l'énergie modifie la géométrie ou la forme d'une molécule d'abaisser l'énergie potentielle de la molécule et pour donner une conformation plus stable [25].

II.2.2 Mécanique Moléculaire

Les champs de force mécaniques moléculaires utilisent les équations de la mécanique classique Décrire l'énergie potentielle de surface et les propriétés physiques des molécules.

Une sorte d'une molécule est décrite comme un ensemble d'atomes en interaction par des fonctions analytiques simples. Cette description s'appelle un champ de force. Une sorte de Les composantes du champ de force sont l'énergie produite par la compression et l'étirement de l'objet obligations [26]. HyperChem comprend quatre domaines de la mécanique des polymères Power : Nouvelles implémentations de technologies développées et publiées par des groupes Recherche respectée mais nous sommes cohérents avec l'approche MM+ dans ce travail Propriétés de la force de champ illustrées dans la (figure -04-) :

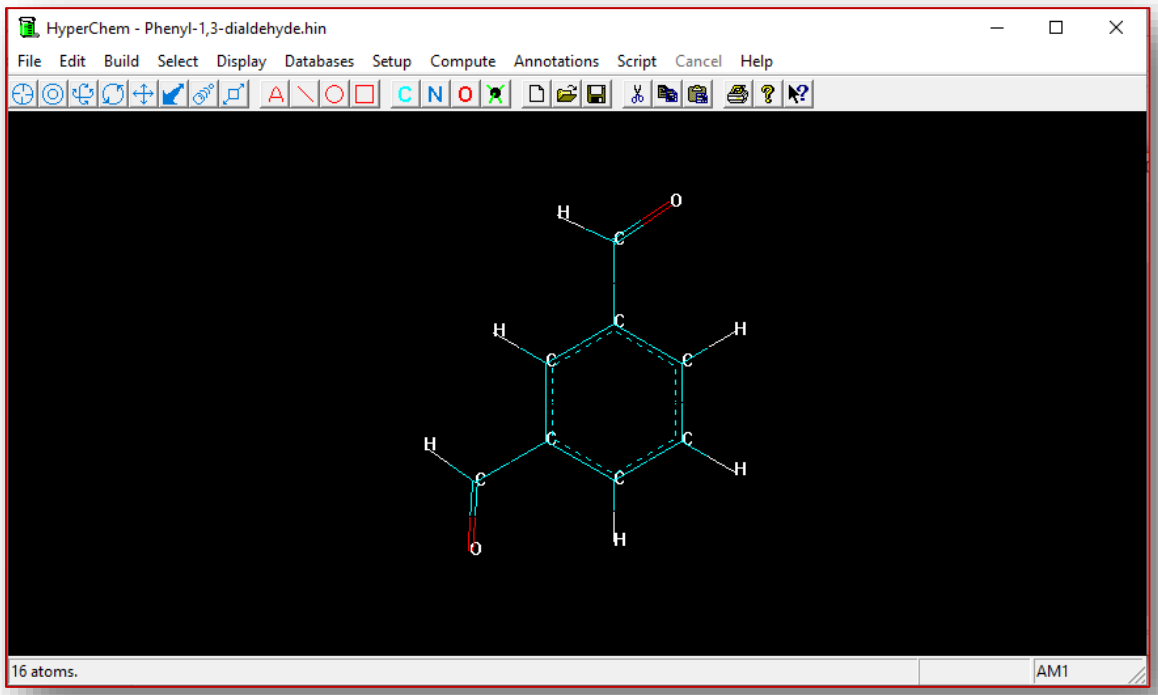


Figure- 04 : Le logiciel Hyperchem

II.2.3 Récupération des fichiers HyperChem HIN

Après avoir construit la structure dans HyperChem, vous pouvez l'enregistrer pour une utilisation ultérieure. C'est une bonne idée car cela vous fait gagner du temps si Vous voulez revoir votre structure plus tard. Pourquoi le construire deux fois ? ! Vous pouvez le faire en allant dans fichier et en enregistrant. Tu dois lui donner un nom hin. Le fichier peut être rappelé à tout moment pour visualisation et manipulation.

Enregistrez-le dans un dossier public dans le répertoire approprié. Dans le calcul Descripteurs moléculaires nécessaires pour optimiser la structure chimique d'un composé. Les structures chimiques des 77 composés de notre ensemble de données ont été établies dans le logiciel HyperChem et pré-optimisées à l'aide du champ mécano-forward MM+. [25]

III-Calcul des descripteurs moléculaires

Afin d'exploiter au maximum les informations contenues dans les structures moléculaires, celles-ci sont traduites en une série de grandeurs (en général scalaires) qui quantifient leurs caractéristiques physico-chimiques et structurelles. Dans la prochaine étape pour tous les composés, les descripteurs moléculaires ont été calculés par le logiciel dragon. Qui peut calculer plus de 1600 descripteurs moléculaires pour chaque structure dans notre jeu des données [27].

III.1 **Le Logiciel DRAGON:**

C'est une application pour le calcul des descripteurs moléculaires. Ces descripteurs peuvent être utilisés pour évaluer l'influence de la structure moléculaire ou les relations propriétés-structure, aussi pour l'analyse de symétrie et la projection des bases de données des molécules [28].

DRAGON fournis 1664 descripteurs moléculaires qui sont divisés en 20 blocs logiques Figure -05- L'utilisateur peut calculer non seulement le descripteurs de type d' atome (Atom J- type), groupe fonctionnel, comptes de fragment, mais aussi des descripteurs topologiques et géométriques .Quelque propriétés moléculaires, comme logP, (LogKoe), molar refractivity, numéro de rotatable bonds, H-donors, H-acceptors, et topological surface area (TPSA) sont calculés par l'utilisation des modèles communs. Pour l'utilisation complète des calculs de DRAGON, structures optimisées 3D avec les atomes d'hydrogène doit utiliser, aussi, DRAGON peut traiter avec Hdepleted molecules et 2D-structures ; dans ce cas, il est apparu que des restrictions au calcul des descripteurs sont adresser.

Figure-05- : le logiciel DRAGON.

III.2 Descripteurs moléculaires

III.2.1 Définition d'un descripteur :

Le descripteur moléculaire est le résultat final d'un processus logique et mathématique qui permet de convertir les informations chimiques chiffrées en une représentation symbolique d'une molécule en un nombre utilisable ou le résultat d'une expérience standard. Les descripteurs moléculaires sont les caractéristiques structurales les plus significatives d'une molécule qui peuvent être utilisées pour développer une "Relation Structure-Propriété". Dans notre cas, la propriété étudiée est la température d'ébullition (Teb). Les descripteurs moléculaires sont proposés à partir de différentes théories et approches dans le but de prédire les propriétés physico-chimiques et biologiques des molécules. [29]

III.2.2 Types de descripteurs :

Il existe différents types de descripteurs moléculaires qui peuvent être utilisés pour caractériser les propriétés d'une molécule. Voici quelques exemples courants :

III.2.2.1 Descripteurs constitutionnels :

Sont un type de descripteurs moléculaires qui sont basés sur la composition et la distribution des éléments chimiques dans une molécule. Ils se basent sur les liaisons et les relations entre les

atomes dans la molécule, et l'analyse de ces descripteurs constitutionnels peuvent être utiles pour comprendre les propriétés physiques et chimiques des molécules.

Quelques exemples de descripteurs constitutionnels incluent :

- Le nombre d'atomes : qui détermine le nombre d'atomes constituant la molécule.
- Le nombre de liaisons : qui indique le nombre de liaisons chimiques entre les atomes dans la molécule.
- La masse moléculaire : qui fait référence à la masse totale de la molécule.
- La distribution des éléments : qui montre comment les éléments chimiques sont répartis à l'intérieur de la molécule, tels que l'oxygène, l'azote, le carbone, etc.

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.-à-d., si on développe des modèles avec ce type de descripteurs seulement, ils peuvent poser problème pour l'interprétation des mécanismes d'interaction mis en jeu pour la propriété étudiée. [30]

III.2.2.2 Descripteurs topologiques :

Ces descripteurs sont basés sur la connectivité atomique et la structure globale de la molécule. Ils incluent des caractéristiques telles que le nombre d'atomes, le nombre de liaisons, la distance entre les atomes, les motifs de cycles, etc. [30]

III.2.2.3 Descripteurs électrostatiques :

Ces descripteurs caractérisent la distribution des charges moléculaires. Ce Calcul des charges partielles empiriques dans les molécules à l'aide de la méthode proposée Zefirov. La méthode est basée sur l'échelle d'électronégativité. Sur la base de ces Les descripteurs électrostatiques suivants sont calculés pour les charges partielles comme suit :

- Charges partielles minimales et maximales dans la molécule (qmin, qmax).
- Charges partielles minimales et maximales des atomes (C, N, O...).
- Indicateurs électroniques topologiques.

Ces descripteurs sont responsables des interactions entre molécules polaires. [30]

III.2.2.4 Descripteurs géométriques :

Le descripteur géométrique de la molécule est dérivé deses atomes dans l'espace, et décrivent des caractéristiques plus complexes ; leurs calculs La modélisation moléculaire est nécessaire pour comprendre la géométrie 3D des molécules Empiriques ou ab initio, ces descripteurs s'avèrent donc relativement coûteux en terme de Calculé, mais fournit plus d'informations et est nécessaire pour la modélisation Dépend des propriétés de la structure 3D. Il existe plusieurs descripteurs importants, Volume moléculaire, surface accessible au solvant, moment d'inertie. Le volume moléculaire est le volume occupé par les molécules en appliquant un maillage 3D d'un cube dans une boite parallélép ipédique de dimensions Xmax, Ymax et Zmax. Là La surface accessible au solvant SAS ou la surface accessible est Molécule accessible aux solvants, généralement mesurée en angströms carrés. Le moment d'inertie est une grandeur physique qui caractérise la distribution de masse dans la molécule. [30]

III.2.2.5 Descripteurs thermodynamiques :

Les descripteurs thermodynamiques sont calculés à partir de la fonction de partition totale Q moléculaire. Les fonctions de partition montrent comment l'énergie d'un système moléculaire est distribuée entre les molécules individuelles. Sa valeur dépend du poids Masse moléculaire, température, volume moléculaire, espacement internucléaire, Mouvement moléculaire et forces intermoléculaires. La fonction de partition est le point Le plus pratique entre les propriétés microscopiques des molécules individuelles (horizontal énergie, moment d'inertie) et propriétés macroscopiques (chaleur spécifique, entropie). Les molécules peuvent augmenter leur énergie en translation, vibration et rotation sont en fait indépendantes. [30]

Ces types de descripteurs peuvent être combinés pour former des ensembles de descripteurs plus complets, fournissant ainsi une représentation détaillée et informatique de la molécule pour l'analyse et la prédiction de ses propriétés.

Figure -06 Représentation des descripteurs moléculaires utilisés à la modélisation QSAR [31].

III.2.3 Groupe des descripteurs moléculaires

En effet, les descripteurs moléculaires sont basés sur plusieurs théories différentes, tel que quantum chimie, théorie de l'information, chimie organique, théorie du graphique, et ainsi de suite, et est utilisé pour modeler des plusieurs propriétés différentes de produits chimiques dans les champs du scientifique tel que toxicologie, chimie analytique, chimie physique…etc. [32].Actuellement, il est possible de calculer plus de 1600 descripteurs moléculaires qui sont représentés dans le tableau suivant qui peuvent être classés en 20 classes (blocs) logiques :

Tableau- 02 *Quelques blocks des descripteurs calculés par logiciel dragon*

Classe	Sous classe
Descripteurs Constitutionnels	❖ Dénombrement des atomes ou des liaisons. ❖ Descripteurs basés sur les masses atomiques
Descripteurs géométriques	❖ Descripteurs liés à la distance. ❖ Descripteurs liés à l'aire de la surface. ❖ Descripteurs liés au volume. ❖ Descripteurs du champ stérique moléculaire
Descripteurs liés à la distribution de charge	❖ Charges atomiques partielles. ❖ Moments électriques moléculaires ❖ Polarisabilités moléculaires. ❖ Descripteurs du champ électrique moléculaire
Descripteurs topologiques	❖ Indices topologiques (connectivité). ❖ Descripteurs théoriques d'information. ❖ Descripteurs topo-chimiques.
Descripteurs température Dépendants	❖ Fonctions thermodynamiques. ❖ Descripteurs facteurs de Boltzmann pondérés
Descripteurs liés aux orbitales Moléculaires	❖ Energie des OM frontières ❖ Ordres de liaison ❖ Indices de réactivité de Fukui.

Tableau - 02 *(suite et fin)*

Classe	Sous classe
Descripteurs mixtes	❖ Descripteurs topographiques. ❖ Descripteurs électro-topologiques. ❖ Descripteurs de la charge partielle de l'aire de la surface.
Descripteurs de solvatation	❖ Energie électrostatique de solvatation. ❖ Energie de dispersion de solvatation. ❖ Enthalpie libre de formation de cavité. ❖ Descripteurs de liaison hydrogène. ❖ Entropie de solvatation. ❖ Descripteurs d'énergie de solvatation linéaire théorique.

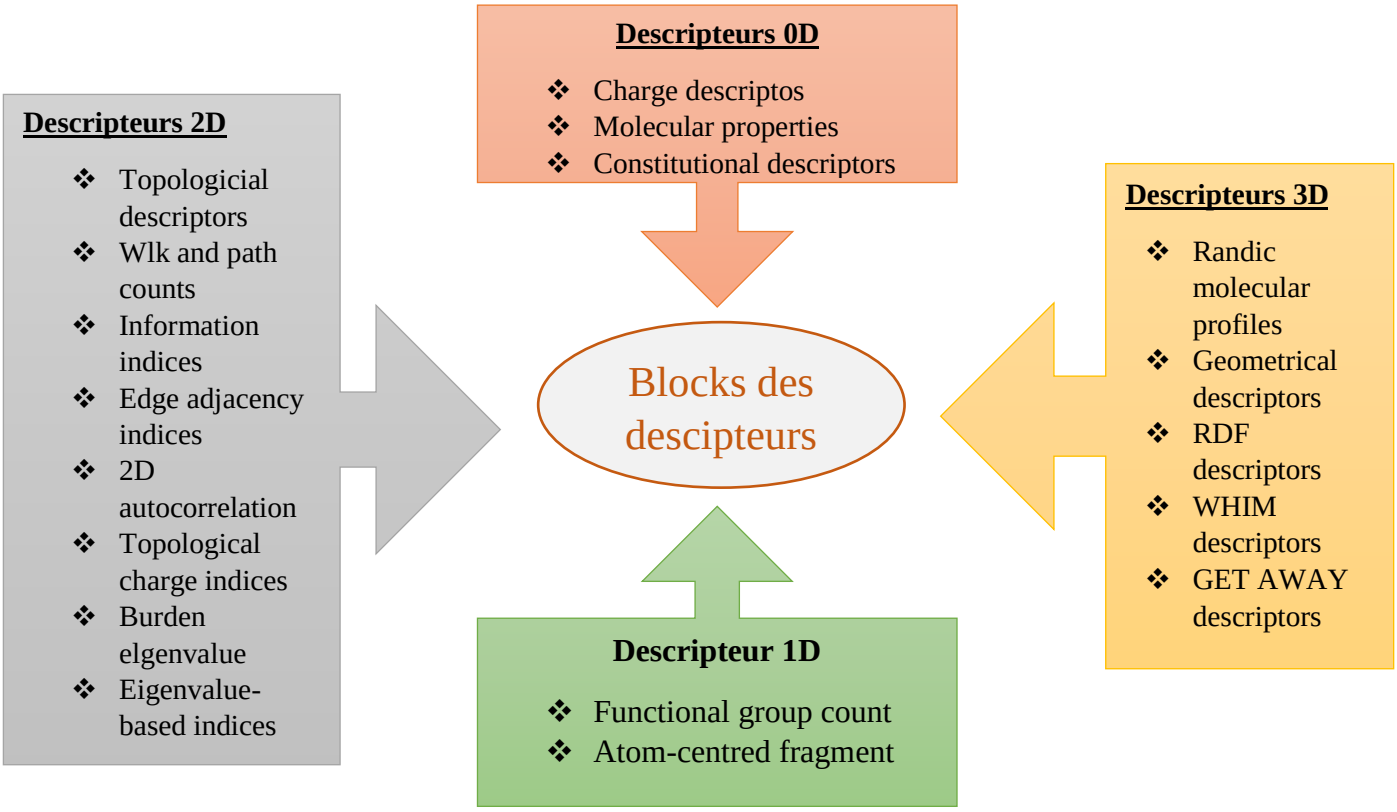


Figure -07 : Représentation des blocs des descripteurs moléculaires [33].

III.3 **Importance des descripteurs**

Les descripteurs moléculaires jouent un rôle important en chimie, en science Médecine, protection de l'environnement, recherche et contrôle de la santé masse, lorsque la molécule est convertie en Représentation moléculaire qui permet certains traitements mathématiques.

Descripteur Les molécules sont importantes pour :

- Indiquons une description de la configuration moléculaire à étudier.
- Décrivons tous les paramètres descriptifs de la molécule.

Les descripteurs moléculaires sont utilisés pour la connaissance statistique, Des principes de chimiométrie et des méthodes QSAR/QSPR sont nécessaires, en plus de connaissance spécifique du problème [34].

III.4 **Les étapes de prédiction :**

Wiener a développé les premières expériences de modélisation QSPR et depuis les echniques de modélisation d'apprentissage d'abord linéaires, puis non linéaires Mise en œuvre de multiples méthodologies, dont la plupart sont basées sur la recherche relation entre un ensemble de nombres réels, des descripteurs de molécules et des propriétés ou l'activité que l'on souhaite prédire [35].

La prédiction des propriétés s’effectué par plusieurs étapes intéressantes illustrant dans le diagramme suivant :

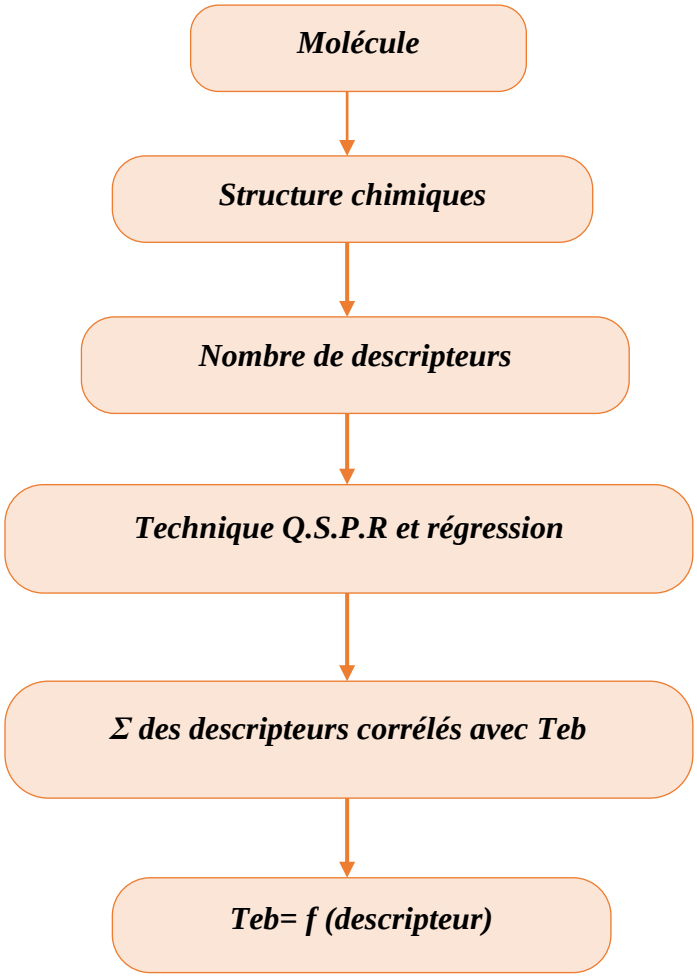


Figure -08 : Diagramme de prédiction par QSPR

III.5 L'objectif de la prédiction

L'objectif principal est de créer un Structures moléculaires, plus précisément descripteurs moléculaires. Le modèle permet Classification selon les composés prédits

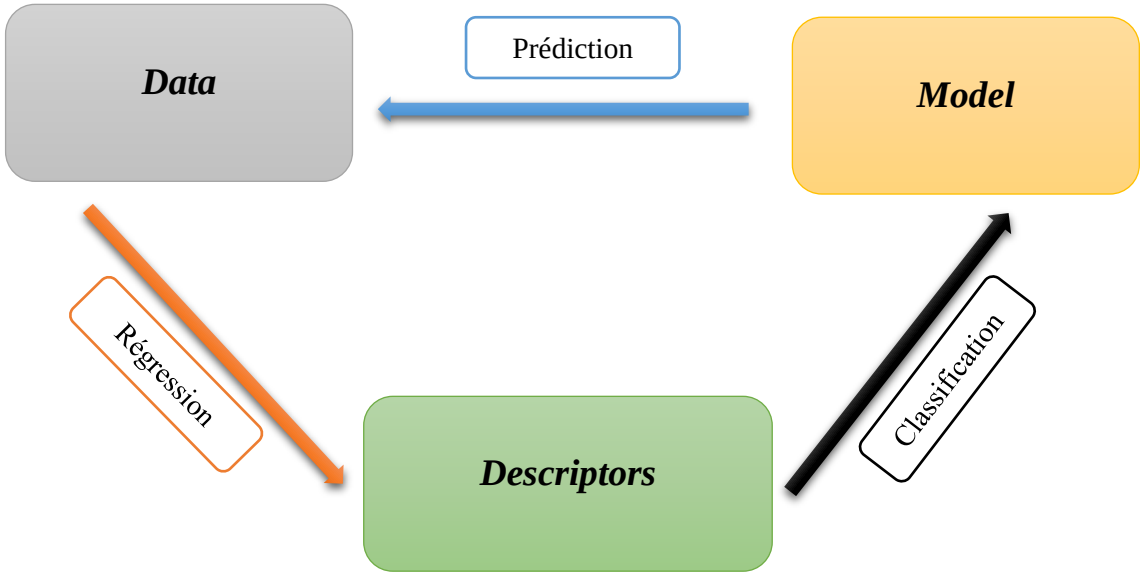


Figure – 09 : Le cycle de prédiction

III.6 Les étapes de travail :

III.6.1 Modélisation :

Le diagramme suivant illustré les étapes du travail et les techniques quant utilisé dans la Procédure de prédiction de la propriété étudiée (la température d'ébullition).

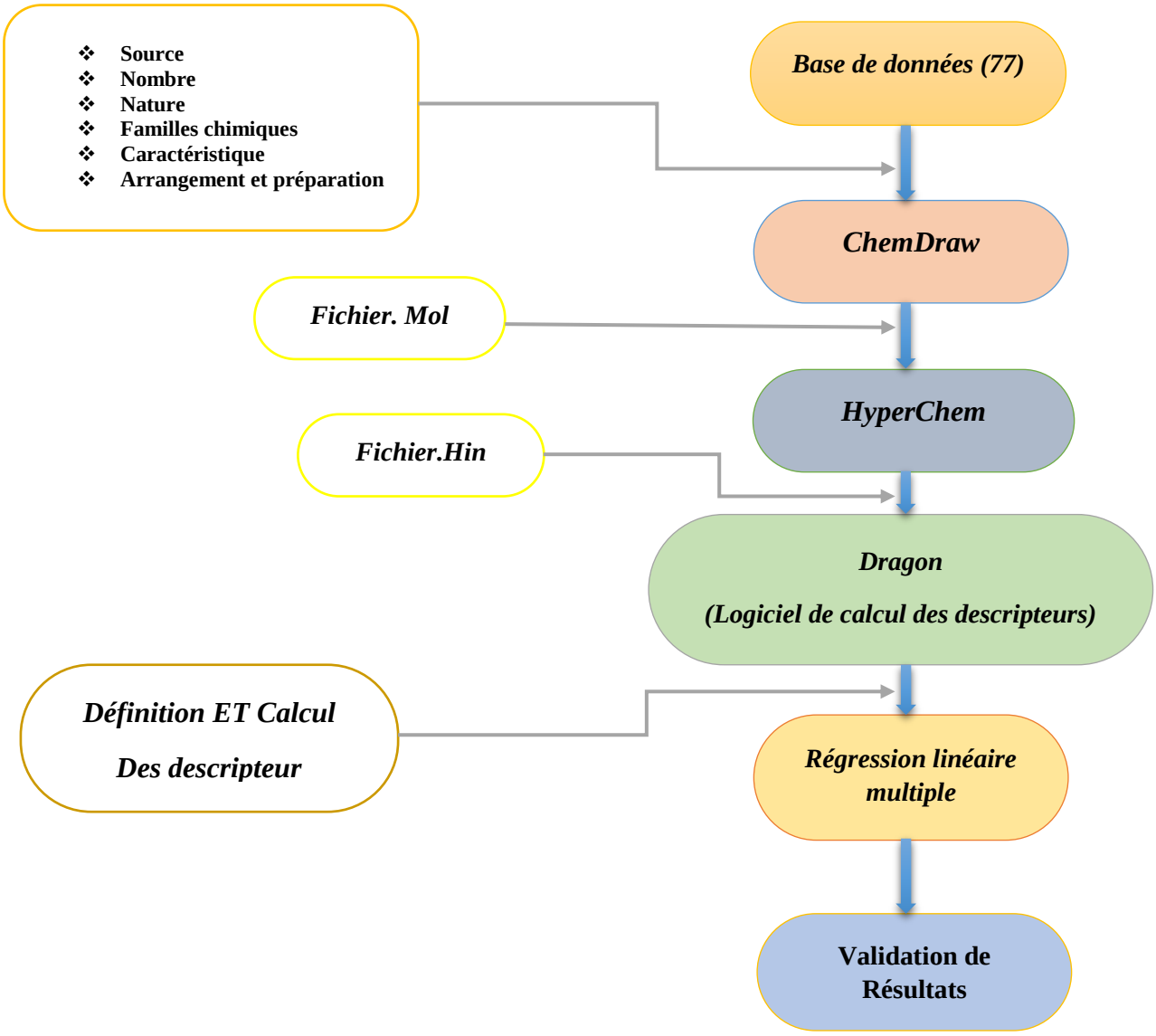


Figure- 10 : Diagramme de notre travail

IV- Méthodes utilisées pour le développement de modèles QSAR/QSPR

IV.1 Introduction

Le développement de modèles QSAR/QSPR pose plusieurs défis [36] .Il existe un grand nombre de descripteurs moléculaires (plus de 3000) proposés au fil des années, sans règles strictes pour

leur sélection. La sélection des descripteurs appropriés est souvent basée sur l'intuition chimique ou la tradition, ce qui rend le processus complexe.

Une autre difficulté réside dans la non-standardisation des gammes de descripteurs. Les constantes d'induction, de résonance et d'effet stérique des constituants ainsi que les échelles empiriques d'effets de solvant comportent des erreurs liées aux mesures expérimentales. De plus, les méthodes quantiques utilisées pour calculer les descripteurs moléculaires peuvent varier, ce qui rend difficile l'utilisation de descripteurs calculés avec différentes méthodes ou pour différents composés.

Une approche systématique pour sélectionner les gammes de descripteurs consiste à utiliser la discrimination statistique entre de larges ensembles de descripteurs.

Différentes approches mathématiques sont utilisées pour développer des modèles QSAR/QSPR. La régression linéaire multiple est la méthode la plus courante, où une équation linéaire est obtenue en régressant les données expérimentales sur un ensemble de descripteurs pré-sélectionnés. Dans certains cas, des formes mathématiques non linéaires peuvent être utilisées lorsque des modèles physiques ou chimiques connus le suggèrent. D'autres méthodes, telles que l'analyse factorielle, l'analyse en composantes principales et la régression par les moindres carrés partiels (PLS), sont également utilisées. [37-38].

Les méthodes d'intelligence artificielle, telles que les réseaux de neurones et les algorithmes génétiques, sont également appliquées au développement de modèles QSAR/QSPR. [39-40]

En résumé, les modèles QSAR/QSPR nécessitent une sélection soigneuse des descripteurs moléculaires, et différentes méthodes mathématiques sont utilisées pour développer ces modèles, y compris des approches statistiques traditionnelles et des techniques d'intelligence artificielle.

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

IV.2 Méthodes de régressions linéaire et multilinéaire

IV.2.1 Aperçu général

Lors de la sélection des descripteurs moléculaires pour les modèles QSAR/QSPR, les chercheurs se basent souvent sur l'intuition chimique, la tradition ou la disponibilité des descripteurs. Cependant, cinq principes peuvent guider cette sélection :

a. Utiliser un maximum de données expérimentales, idéalement toutes, caractérisées par des valeurs de descripteurs complémentaires.
b. Obtenir les valeurs des descripteurs à partir de la même source et, de préférence, en utilisant le même protocole expérimental ou le même logiciel de calcul.
c. Réduire le nombre de descripteurs dans les modèles de régression multiple tout en conservant l'information, en se basant sur des critères statistiques tels que les tests t et F.
d. Dans les modèles de régression linéaire multiple, les descripteurs utilisés doivent être statistiquement orthogonaux.
e. Si tous les autres critères sont similaires, privilégier les descripteurs dont la nature physique ou chimique est la plus proche de la propriété ou du phénomène étudié.

En pratique, il est souvent difficile de respecter pleinement les cinq principes énoncés. Cependant, négliger plusieurs d'entre eux peut entraîner des équations qui sont soit inutiles, soit dotées d'un pouvoir prédictif très limité [41].

IV.2.2 Evaluation préliminaire des données

Avant de commencer le développement proprement dit de l'équation de régression QSPR, il est fortement recommandé de vérifier la qualité statistique des données de départ, à la fois les données à corréler (variables dépendantes) et les descripteurs utilisés dans la corrélation (variables indépendantes).

Une distinction est souvent faite dans ce prétraitement des données entre analyse univariée et analyse bivariée [42-43].

En analyse univariée, il est conseillé de vérifier si les données correspondent à une distribution normale. Une attention particulière doit être portée lors des régressions ultérieures si les valeurs des attributs ou des descripteurs à l'étude n'obéissent pas à la loi de Laplace-Gauss.

Pour un ensemble différent de descripteurs, une analyse de données bivariée est requise, c'est-à-dire le calcul du coefficient de corrélation linéaire R entre chaque paire d'ensembles de descripteurs. Si R est statistiquement significatif (R>0,9), ces deux descripteurs ne peuvent pas être utilisés simultanément dans l'analyse par régression linéaire multiple (MLR).

IV.2.3 Régression linéaire multiple

L'objectif de la régression linéaire simple et multiple : est d'apprendre comment analyser un phénomène quelconque en utilisant des méthodes statistiques dites économétriques.

En effet, la régression linéaire est une relation stochastique entre une ou plusieurs variables.

Elle est appliquée dans plusieurs domaines, tels que la physique, la biologie, la chimie, l'économie…etc.

Un modèle de régression linéaire multiple entre une variable expliquée Y et p variables explicatives X1…,XP, s'écrit pour tout i=1,…, n :

$$\mathbf{y = \beta 0 + \sum_{j=i}^{p} \beta j\ Xij\ + \xi}$$

Où les yi, xi1, xi2…, xip : sont des données respectivement relatives aux variables Y, X1..., XP.

Les estimateurs βj sont calculés en utilisant la méthode des moindres carrés ordinaires. Les variables aléatoires ξ représentent les termes d'erreur non observables du modèle. On peut estimer ces erreurs par les résidus ordinaires ei, différence entre les valeurs observées yi et les valeurs estimées ŷi.

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

a. Les résidus (E) ont une espérance mathématique nulle :

$$E(e) = 0$$

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

c. Les résidus sont indépendants entre eux :

$$E(ei, ej) = 0 \text{ Si } i \neq j$$

Leurs covariances sont nulles.

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

Par ailleurs, l'emploi de tests statistiques pour analyser la variation expliquée par la régression conduit à admettre que : Les résidus suivent une distribution normale (de Laplace-Gauss).

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

Pour minimiser l'influence des erreurs de détermination des valeurs explicatives (ou répresseurs) sur la précision des résultats de la régression 5 données (variables dépendantes, ou encore observations) doivent, à la limite, être associées à chaque variable explicative. Le nombre de degrés de liberté final (n-p-1) doit être [44] tel que :

$$n - p - 1 \geq 10$$

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

IV.2.4 Algorithme génétique

La modélisation de processus génétiques a initié le développement des algorithmes génétiques, qui peuvent être exploités dans une grande variété de problèmes d'optimisation [45].Dans un algorithme génétique adapté à l'optimisation, une solution potentielle est considérée comme un individu dans une population. La valeur de la fonction de coût associée à une solution mesure « l'adaptation » de l'individu associé à son environnement. Un algorithme génétique simule l'évolution, sur plusieurs générations, d'une population initiale dont les individus sont mal adaptés au moyen d'opérateurs génétiques de reproduction et de mutation. Après un certain nombre de générations, la population est constituée d'individus bien adaptés, autrement dit des solutions supposées « bonnes » au problème d'optimisation.

Dans ce travail les sélections des descripteurs en utilisant le logiciel de calcul statistique MINITAB version 16.2.0 [46] ; et par algorithme génétique, dans la version MOBY DIGS de Todeschini [47].

IV.3 Paramètres d'évaluation de la qualité de l'ajustement

Deux paramètres sont couramment utilisés :

Le coefficient de détermination multiple :

Pour comprendre la qualité de l'ajustement obtenu, nous avons calculé le coefficient de détermination R^2, qui représente la fraction de la variation de régression Y (=température d'ébullition) ″ expliquée ou "raisonnable". Ce paramètre correspond au carré du coefficient de corrélation, entre 0 et 1, exprimé en Toujours un pourcentage.

Si la valeur de R^2 est proche de 1 ou 100% ; on a donc un excellent ajustement qualité ; en revanche, si la valeur du R^2 est faible et proche de 0 ou 0 %, elle est mal ajustée.

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

Où ŷi est la valeur estimée du paramètre physique, et ȳ la moyenne des valeurs expérimentales.

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

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

IV.4 Facteur d'inflation de la variance [FIV]

Le facteur d'inflation de la variance sert à détecter si descripteur présente une association linéaire forte avec les prédicteurs restants (présence de multi colinéarité parmi les prédicteurs). Le facteur d'inflation de la variance donne une mesure de l'accroissement de la variance d'un coefficient de régression estimé s'il existe une corrélation entre prédicteurs (multi colinéarité).FIV =1 indique qu'il n'y a pas de relations, si non FIV est supérieur à 1 le facteur FIV le plus grand parmi tous les prédicteurs sert souvent d'indicateur de multi colinéarité importance, Si le FIV > 5-10 la qualité de l'estimation des coefficients de régression est faible [48].

IV.5 **Test de randomisation**

Ce test permet de mettre en évidence des corrélations dues au hasard. Il consiste à générer un vecteur « propriété considérée » par permutation aléatoire des composantes du vecteur réel. On calcule alors sur le vecteur obtenu (considéré comme vecteur expérimental réel) un modèle QSPR, selon la méthode habituelle. Ce procédé est répété plusieurs fois (100 dans notre cas).

IV.6 **Validation externe**

Il est intéressant, pour juger de la qualité du modèle, de considérer la racine de l'écart Quadratique moyen (RMSE, pour Root Mean Squared Error), calculée sur différents ensembles :

- Ensemble d'estimation (appelée SDEC)
- Ensemble de validation croisée (appelée également SDEP)
- Ensemble de prédiction externe (désignée par SDEPext).

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

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

$$SDEP_{ext} = \sqrt{\frac{\sum_{i=1}^{n_{ext}}(y_i - \hat{y}_i)^2}{n_{ext}}}$$

La validation croisée du « leave-one-out » (LOO) [49] consiste à recalculer les modèles (n-1) observations et utiliser le modèle résultant pour calculer la quantité d'intérêt Composés rejetés, notés ŷ(i). Répétez le processus pour chaque quantité d'intérêt.

La somme des erreurs de prédiction au carré, désignée par le symbole PRESS, est une mesure Dispersion estimée. Il est utilisé pour définir des coefficients de prédiction :

$$Q^2_{LOO} = \frac{SCT - PRESS}{SCT}$$

Contrairement à R^2 au coefficient qui augmente avec le nombre de paramètres du modèle, le Facteur Q^2_{LOO} affiche une courbe avec maximum (ou avec palier) obtenu pour un certain nombre de descripteurs, puis décroît de façon monotone. Ce fait confère une grande Signification du coefficient Q^2_{LOO}. Les valeurs Q^2_{LOO} > 0,5 sont considérées comme satis faisantes, et les valeurs supérieures à 0,9 sont excellentes [50]. Si une valeur plus petite de Q^2_{LOO} indique un modèle plus faible, caractérisé par un pouvoir prédictif interne plus faible, pas nécessairement l'inverse. En effet, si une valeur élevée de Q^2_{LOO} est une condition nécessaire à la robustesse et éventuellement au pouvoir prédictif élevé du modèle, alors cette condition seule n'est pas suffisante et peut conduire à une surestimation du pouvoir prédictif du modèle.

Dans le cas de vrais composés externes, 2, 3 éléments ou plus peuvent devoir être jetés en même temps, ce qui conduit à une procédure LMO (Leave-More-Out). Cependant, ces procédures sont rarement signalées comme des résultats QSPR communs et sous-utilisées dans les travaux actuels. Dans le cas où on a suffisamment de données qui n'ont pas servi dans la création du modèle ou après collecte de nouvelles, on peut ou on doit procéder à la validation de ce dernier, c'est la validation externe. La statistique se rapportant à ce procédé, notée Q^2ext, est calculée comme suit :

$$Q^2_{ext} = \frac{\sum_{I=1}^{n\ ext}(y_i - \widehat{y_i})^2/next}{\sum_{i=1}^{n}(y_i - \widehat{y})^2/n}$$

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

Partie Application

Nous traiterons dans ce travail la propriété physicochimique envisagée (Teb), en considérant les mêmes ensembles d'estimation et de validation. Rappelant que les 77 composés ont été éclatés aléatoirement en deux sous ensemble comportant 54 pour la calibration et 23 pour la validation. Les résultats ainsi obtenus seront discutés.

1. Sélection des descripteurs

Différentes catégories de descripteurs moléculaires ont été calculées pour l'ensemble de la molécule : descripteurs de construction moléculaire et descripteurs topologiques. Nous présenterons ici quelques résultats obtenus pour toutes les molécules (calibration et validation).

Le modèle a été construit à l'aide d'un algorithme génétique utilisant le logiciel mobydigs. Le niveau de signification a été fixé à 0,05 à la fois pour l'inclusion et l'exclusion des variables.

De plus, les modèles ont été construits pour différentes dimensions en utilisant l'algorithme génétique avec le logiciel mobydigs. Dans les études QSPR, chaque variable explicative doit être associée à au moins cinq composés. Le nombre final de degrés de liberté doit être d'au moins 10. Cette condition est satisfaite pour les différents modèles, car le nombre de composés est de 54, ce qui nous permet d'inclure jusqu'à 10 variables.

Une indépendance globale acceptable des descripteurs sera vérifiée lorsque les facteurs d'inflation de la variance (FIV) calculés pour chacun d'entre eux seront inférieurs à 5.
Parmi les modèles optimaux générés, celui qui présente les paramètres statistiques Q^2, R^2 et Q^2ext les plus élevés, tout en respectant la condition FIV < 5, comprend les cinq descripteurs calculés par le logiciel DRAGON. Les symboles, les classes et les significations de ces descripteurs sont résumés dans le tableau 4.

2. Choix de la taille du modèle :

Permis les plusieurs modèles obtenue, le tableau suivant représente les valeurs de R^2 et Q^2 en fonction du nombre de descripteurs k ; on voit que le bon modèle qui possède le moins de descripteurs est celui à 5, parce qu'après ce dernier les valeurs des R^2 augmentent d'un pas faible.

Tableau -03- les valeurs de R^2 et Q^2 en fonction du nombre de descripteurs k

Nombre de Descripteurs	R^2	Q^2loo
Qneg H3v	0.6313	0.5883
TE1 X1sol Du	0.67	0.6205
TE1 X0sol JGI4 De	0.7039	0.6426
BIC0 BELm6 Mor04v De R3u+	0.8078	0.7672
TE1 Mor04p H3p E1e ESpm11d De	0.8079	0.768
TE1 Mor04p R5u JGI4 E1e Mv RDF060m	0.8077	0.7667
TE1 SMTI Mor04p CIC0 E1e Mv R2m RDF060m	0.8128	0.7683

Tableau -04-Valeurs des descripteurs moléculaires sélectionnés.

N	Composé	Tb	BIC0	BELm6	Mor04v	De	R3u+
1	2,3,4-Trihydroxybenzaldehyde	575	0.353	0.405	-0.316	0.339	0.063
2	2,3,5-Trichlorobenzaldehyde	542.5	0.414	0	-0.129	0.368	0.083
3	2,3-Dihydroxybenzaldehyde	513	0.348	0.325	-0.402	0.334	0.075
4	2,4,5-Trimethoxybenzaldehyde	584.5	0.298	0.92	0.172	0.439	0.029
5	2,4,6-Trihydroxybenzaldehyde	607.7	0.353	0.484	-0.346	0.326	0.059
6	2,4-Dihydroxybenzaldehyde	624	0.348	0.404	-0.395	0.327	0.067
7	2,4-Dimethoxybenzaldehyde	581	0.306	0.572	0.028	0.399	0.046
8	2,5-Dihydroxybenzaldehyde	486.5	0.348	0.382	-0.378	0.383	0.071
9	2-Anisaldehyde	516.5	0.312	0.241	-0.216	0.342	0.077

Tableau-04-(Suite).

N	Composé	Tb	BIC0	BELm6	Mor04v	De	R3u+
10	2-Chloro-3-hydroxy-4-methoxybenzaldehyde	564.3	0.376	0.382	-0.127	0.36	0.066
11	2-Chloro-4-hydroxycarboxaldehyde	545.5	0.398	0.071	-0.352	0.334	0.086
12	2-Chloro-5-nitrobenzaldehyde	566.5	0.443	0.068	-0.027	0.378	0.076
13	2-Chlorobenzaldehyde	482	0.378	0.066	-0.345	0.347	0.089
14	2-Fluorenecarboxaldehyde	564	0.233	0.922	-0.628	0.373	0.066
15	2-Fluorobenzaldehyde	514.5	0.378	0.106	-0.406	0.354	0.088
16	2-Hydroxy-1-naphthaldehyde	579	0.278	0.651	-0.273	0.355	0.063
17	2-Hydroxy-3-nitrocarboxaldehyde	569	0.393	0.328	-0.193	0.349	0.072
18	2-Hydroxybenzaldehyde	470	0.337	0.238	-0.404	0.339	0.083
19	2-Methyl-1-naphthaldehyde	589.5	0.245	0.774	-0.059	0.371	0.06
20	3,4,5-Trihydroxybenzaldehyde	651.4	0.353	0.5	-0.353	0.338	0.059
21	3,4-Dihydroxybenzaldehyde	596	0.348	0.388	-0.386	0.365	0.072
22	3,4-Dimethoxy-5-hydroxycarboxaldehyde	626	0.314	0.631	-0.064	0.391	0.039
23	3,5-Dibromo-4-hydroxycarboxaldehyde	546.3	0.423	0	-0.124	0.327	0.075
24	3,5-Dibromosalicylaldehyde	534.5	0.423	0	-0.125	0.355	0.078
25	3-Anisaldehyde	504	0.312	0.275	-0.195	0.367	0.077
26	3-Bromo-4-hydroxycarboxaldehyde	534.5	0.398	0.037	-0.193	0.34	0.08
27	3-Bromobenzaldehyde	508.5	0.378	0	-0.244	0.356	0.085
28	3-Chloro-2-fluoro-5-(trifluoromethyl)benzaldehyde	469	0.438	0	-0.328	0.43	0.072

Tableau-04-(Suite).

N	Composé	Tb	BIC0	BELm6	Mor04v	De	R3u+
29	3-Chlorobenzaldehyde	486.5	0.378	0	-0.33	0.362	0.086
30	3-Cyanobenzaldehyde	484	0.349	0.27	-0.509	0.367	0.08
31	3-Ethoxy-2-hydroxycarboxaldehyde	537	0.306	0.724	-0.18	0.383	0.133
32	3-Ethoxy-4-hydroxybenzaldehyde	558	0.306	0.723	-0.124	0.388	0.133
33	3-Fluorobenzaldehyde	467	0.378	0	-0.401	0.363	0.087
34	3-Hydroxybenzaldehyde	514	0.337	0.238	-0.404	0.339	0.083
35	3-Methoxysalicylaldehyde	538.5	0.325	0.381	-0.19	0.349	0.072
36	3-Nitrobenzaldehyde	560	0.392	0.271	-0.224	0.361	0.075
37	3-Tolualdehyde	496	0.288	0.273	-0.331	0.392	0.076
38	4-(Dimethylamino)benzaldehyde	583	0.305	0.646	-0.036	0.433	0.061
39	4,6-Dimethoxy-2-hydroxybenzaldehyde	643.5	0.314	0.608	-0.218	0.376	0.045
40	4-Biphenylcarboxaldehyde	599.6	0.239	0.932	-0.006	0.364	0.055
41	4-Cyanobenzaldehyde	563	0.349	0.311	-0.497	0.315	0.08
42	4-Ethylbenzaldehyde	494	0.269	0.559	-0.278	0.429	0.107
43	4-Fluorobenzaldehyde	455	0.378	0	-0.408	0.354	0.087
44	4-Hydroxy-3-nitrobenzaldehyde	548	0.393	0.393	-0.207	0.38	0.064
45	4-Methyl-1-naphthaldehyde	533	0.245	0.834	-0.153	0.367	0.065
46	4-Phenoxybenzaldehyde	593	0.26	0.988	-0.606	0.36	0.056
47	5-Bromosalicylaldehyde	520.5	0.398	0.002	-0.203	0.373	0.08

Tableau-04-(Suite).

N	Composé	Tb	BIC0	BELm6	Mor04v	De	R3u+
48	5-Chlorosalicylaldehyde	490.5	0.398	0.002	-0.311	0.377	0.081
49	5-Hydroxy-2-nitrobenzaldehyde	646	0.393	0.397	-0.159	0.353	0.067
50	Benzaldehyde	540	0.311	0.23	-0.405	0.346	0.087
51	Phenanthrene-9-carboxaldehyd	678.5	0.224	0.954	0.092	0.345	0.056
52	Phenyl-1,3-dialdehyde	519	0.32	0.272	-0.292	0.349	0.075
53	p-Tolualdehyde	477	0.288	0.311	-0.289	0.372	0.074
54	Terephthaldicarboxaldehyde	519.5	0.312	0.593	-0.505	0.343	0.066
55	1-Naphthaldehyde	592	0.255	0.637	-0.35	0.36	0.067
56	2,4-Dichlorobenzaldehyde	506	0.405	0	-0.25	0.34	0.088
57	2-Bromobenzaldehyde	504	0.378	0.037	-0.291	0.34	0.089
58	2-Chloro-6-fluorobenzaldehyde	631	0.439	0	-0.245	0.35	0.085
59	2-Nitrobenzaldehyde	565	0.392	0.238	-0.213	0.382	0.079
60	2-Tolualdehyde	473	0.288	0.239	-0.252	0.404	0.08
61	3-Hydroxy-4-methoxybenzaldehyde	581	0.325	0.458	-0.206	0.351	0.063
62	3-Hydroxy-4-nitrobenzaldehyde	576.5	0.393	0.405	-0.146	0.344	0.065
63	3-Methoxy-4-hydroxybenzaldehyde	604	0.325	0.409	-0.193	0.387	0.065
64	4-(Pentyloxy)benzaldehyd	675	0.25	1.076	0.188	0.433	0.073
65	4-Acetamidobenzaldehyde	657.5	0.336	0.553	-0.024	0.364	0.101
66	4-Anisaldehyde	521	0.312	0.311	-0.215	0.349	0.081

Tableau-04-(Suite et fin).

N	Composé	Tb	BIC0	BELm6	Mor04v	De	R3u+
67	4-Bromobenzaldehyde	517	0.378	0	-0.088	0.33	0.085
68	4-Butoxybenzaldehyde	558	0.262	0.895	0.175	0.416	0.082
69	4-Chlorobenzaldehyde	486.5	0.378	0	-0.263	0.339	0.086
70	Pentafluorobenzaldehyde	439	0.378	0	-0.388	0.392	0.084
71	4-Ethoxybenzaldehyde	528	0.292	0.639	-0.185	0.378	0.134
72	4-Hydroxy-1-naphthaldehyde	638	0.278	0.655	-0.212	0.365	0.065
73	4-Hydroxybenzaldehyde	583	0.337	0.311	-0.394	0.331	0.083
74	4-Isopropylbenzaldehyde	576	0.254	0.684	-0.039	0.4	0.071
75	4-Nitrobenzaldehyde	573	0.392	0.311	-0.185	0.357	0.073
76	5-Bromovanillin	625	0.376	0.284	0.092	0.361	0.055
77	6-Chloro-2-fluoro-3-methylbenzaldehyde	522	0.401	0	-0.179	0.395	0.077

❖ Les 23 derniers composés sont destinés à la validation externe.

Tableau- 05- Classes et significations des descripteurs.

Descripteur	Classe	Signification
BIC0	**Information indices**	**Indices of neighbourhood symmetry** (IC*k*, TIC*k*, SIC*k*, BIC*k*, CIC*k*) are topological information indices calculated for a H-included molecular graph and based on neighbour degrees and edge multiplicity [V.R. Magnuson, D.K. Harriss, S.C. Basak, Topological Indices Based on Neighborhood Symmetry: Chemical and Biological Applications in Studies in *Physical and Theoretical Chemistry*, R.B. King (Ed.), Elsevier, Amsterdam (The Netherlands), pp. 178-191, 1983]. They are calculated by partitioning graph vertices into equivalence classes; the topological equivalence of two vertices is that the corresponding neighbourhoods of the k*th* order are the same.
BELm6	**Burden eigenvalue descriptors**	DRAGON provides the first 8 highest eigenvalues BEHwk and the first 8 lowest eigenvalues BELwk (absolute values) for each matrix, w referring to the atomic property and k to the eigenvalue rank.
Mor04v	**3D-MoRSE descriptors**	3D-MoRSE (3D-Molecule Representation of Structures based on Electron diffraction) descriptors are based on the idea of obtaining information from the 3D atomic coordinates by the transform used in electron diffraction studies for preparing theoretical scattering curves [J.H.Schuur, P.Selzer, J.Gasteiger, J. Am. Chem. Soc. 1996, 36, 334-344].
De	**WHIM descriptors**	D total accessibility index / unweighted WHIM descriptors (Weighted Holistic Invariant Molecular descriptors) are geometrical descriptors based on statistical indices calculated on the projections of the atoms along principal axes

Tableau- 05- (Suite et fin).

Descripteur	Classe	Signification
R3u+	**GETAWAY descriptors**	Both R matrix average row sum (RARS) and Randic-type R matrix connectivity (RCON) are based on the row sums of the influence/distance matrix since these encode some useful information that could be related to the presence of significant substituents or fragments in the molecule. In effect, it has been observed that larger row sums correspond to terminal atoms that are located very next to other terminal atoms such as those in substituents on a parent structure. Moreover, the RCON index is very sensitive to the molecular size as well as to conformational changes and cyclicity.

3. Calcul des corrélations entre les différents descripteurs

Le coefficient de corrélation, r, de Bravais-Pearson a servi pour mettre en évidence les relations possibles entre les différents descripteurs des 54 composés, la matrice de corrélation obtenue à l'aide de la commande "corrélation" du logiciel MINITAB, montre que les descripteurs sont entre eux plus ou moins corrélés. Les couples des descripteurs qui présentent des valeurs de $r > 0.90$, sont très fortement corrélés et apportent la même information, ce qui fait qu'ils ne peuvent apparaître dans une même équation de régression et ce n'est le cas dans notre cas.

Tableau- 06 Corrélations Teb avec les 5 descripteurs

Correlations: Teb; BIC0; BELm6; Mor04v; De; R3u+

	Teb	BIC0	BELm6	Mor04v	De
BIC0	-0.273				
	0.046				
BELm6	0.593	-0.836			
	0.000	0.000			
Mor04v	0.381	-0.059	0.217		
	0.004	0.669	0.116		
De	-0.097	-0.181	0.258	0.392	
	0.486	0.190	0.060	0.003	
R3u+	-0.527	0.217	-0.374	-0.257	-0.081
	0.000	0.116	0.005	0.060	0.560

Cette matrice nous permette de voir que le descripteur (BELm6) est bien corrélé avec la propriété, par contre les descripteurs R3u+ > Mor04v > BIC0 > a une corrélation très petit mais ils portent un complément pour le modèle. Egalement les autres descripteurs ne sont pas corrélés entre eux à l'exception de la plupart qui présentent un p> 0.05

4. Equation de régression

L'équation de régression du modèle calculé est la suivante :

Teb = 627 + 572 BIC0 + 186 BELm6 + 89.8 Mor04v - 754 De - 645 R3u+

Tableau-07- paramètre de régression.

Predictor	*Coef*	*SE Coef*	*T*	*P*	*VIF*
Constant	*627.34*	*86.97*	*7.21*	*0.000*	
BIC0	*572.2*	*138.0*	*4.15*	*0.000*	*3.594*
BELm6	*185.81*	*27.38*	*6.79*	*0.000*	*4.110*
Mor04v	*89.82*	*27.47*	*3.27*	*0.002*	*1.305*
De	*-754.5*	*164.8*	*-4.58*	*0.000*	*1.238*
R3u+	*-644.9*	*247.7*	*-2.60*	*0.012*	*1.249*

5. Analyse de régression

Les valeurs de T des descripteurs d'une façon générale sont presque proches, cela nous a permet de de dire qu'il y a une bonne homogénéité de la contribution descripteurs des dans notre modèle.

Les valeurs des VIF (< 5) suggèrent que ces descripteurs sont faiblement corrélés les uns avec les autres. Ainsi, le modèle peut être considéré comme une équation de régression optimale.

Pour la robustesse du modèle est assurée par la valeur de Q^2_{LOO}>76% alors que les valeurs de L'erreur quadratique moyenne de prédiction et de calcul sont petites et proches ; en plus ce modèle est significatif avec une valeur du paramètre de Fisher : (F=40.35).

Le tableau suivant regroupe tous les paramètres statistiques

Tableau- 08-Valeurs des paramètres statistiques pour l'ensemble de calibration

N	54
R^2	80.78
Q^2	76.72
F	40.35
S	25.2806
SDEC	23.8348
SDEP	26.2304

6. Analyse *des points aberrants sur l'axe des Y et X*

Les points aberrants sont localisés loin des valeurs de la température prédite (points aberrants sur l'axe des Y) ou loin des valeurs des descripteurs (points aberrants sur les axes des X).

Généralement, les points possédants des valeurs résiduelles normalisées supérieures à 3 fois de l'écart type sont considérés comme points aberrants sur l'axe des Y.

On peut déterminer ainsi les points aberrant sur les axes des descripteurs en utilisant la valeur de levier h_{ii}, dont les points possédants des valeurs supérieures à $\frac{3(p+1)}{n} = \frac{3(5+1)}{54} = 0.33$ sont considérés comme points à grand levier, avec p et n nombres de descripteurs dans le modèle et nombres d'observations respectivement.

Les résultats obtenus lors de l'analyse des points aberrants et points à grand valeur de levier pour notre modèle sont affichés dans le tableau 09.

Tableau- 09 Les Valeurs expérimentales, calculées, prédites et leurs erreurs pour l'ensemble de calibration

N	Composé	Y Exp.	Y-Calc	Y-Pred	Hat	Err.Calc.	Err.Pred.	Std.Err.Calc.	Std.Err.Pred.
1	2,3,4-Trihydroxybenzaldehyde	575	584.31	584.8	0.05	9.31	9.8	0.38	0.4
2	2,3,5-Trichlorobenzaldehyde	542.5	519.43	517.6	0.074	-23.07	-24.9	-0.95	-1.02
3	2,3-Dihydroxybenzaldehyde	513	551.55	553.43	0.047	38.55	40.43	1.56	1.64
4	2,4,5-Trimethoxybenzaldehyde	650.5	657.07	659.84	0.297	6.57	9.34	0.31	0.44
5	2,4,6-Trihydroxybenzaldehyde	607.7	609.69	609.9	0.097	1.99	2.2	0.08	0.09
6	2,4-Dihydroxybenzaldehyde	624	579.25	575.94	0.069	-44.75	-48.06	-1.83	-1.97
7	2,4-Dimethoxybenzaldehyde	581	593.4	595.05	0.118	12.4	14.05	0.52	0.59
8	2,5-Dihydroxybenzaldehyde	486.5	533.9	536.79	0.058	47.4	50.29	1.93	2.05
9	2-Anisaldehyde	516.5	523.74	524.53	0.098	7.24	8.03	0.3	0.33
10	2-Chloro-3-hydroxy-4-methoxybenzaldehyde	564.3	593.25	594.98	0.056	28.95	30.68	1.18	1.25
11	2-Chloro-4-hydroxycarboxaldehyde	545.5	525.47	524.23	0.058	-20.03	-21.27	-0.82	-0.87
12	2-Chloro-5-nitrobenzaldehyde	566.5	557.7	556.33	0.135	-8.8	-10.17	-0.37	-0.43
13	2-Chlorobenzaldehyde	482	502.04	503.02	0.046	20.04	21.02	0.81	0.85
14	2-Fluorenecarboxaldehyde	564	563.68	563.58	0.236	-0.32	-0.42	-0.01	-0.02
15	2-Fluorobenzaldehyde	514.5	500.33	499.65	0.046	-14.17	-14.85	-0.57	-0.6
16	2-Hydroxy-1-naphthaldehyde	579	582.96	583.17	0.051	3.96	4.17	0.16	0.17
17	2-Hydroxy-3-nitrocarboxaldehyde	569	589.21	590.78	0.072	20.21	21.78	0.83	0.89
18	2-Hydroxybenzaldehyde	470	517.74	520.19	0.049	47.74	50.19	1.94	2.04
19	2-Methyl-1-naphthaldehyde	589.5	599.3	600.49	0.109	9.8	10.99	0.41	0.46

Tableau-09-(Suite).

N	Composé	Y Exp.	Y-Calc	Y-Pred	Hat	Err.Calc	Err.Pred.	Std.Err .Calc.	Std.Err. Pred.
20	3,4,5-Trihydroxybenzaldehyde	651.4	603.82	599.56	0.082	-47.58	-51.84	-1.96	-2.14
21	3,4-Dihydroxybenzaldehyde	596	546.23	544.21	0.039	-49.77	-51.79	-2.01	-2.09
22	3,4-Dimethoxy-5-hydroxycarboxaldehyde	626	612.54	610.95	0.106	-13.46	-15.05	-0.56	-0.63
23	3,5-Dibromo-4-hydroxycarboxaldehyde	546.3	560.2	562.15	0.123	13.9	15.85	0.59	0.67
24	3,5-Dibromosalicylaldehyde	534.5	538.07	538.39	0.081	3.57	3.89	0.15	0.16
25	3-Anisaldehyde	504	514.92	515.77	0.072	10.92	11.77	0.45	0.48
26	3-Bromo-4-hydroxycarboxaldehyde	534.5	533.75	533.7	0.066	-0.75	-0.8	-0.03	-0.03
27	3-Bromobenzaldehyde	508.5	494.89	493.98	0.062	-13.61	-14.52	-0.56	-0.59
28	3-Chloro-2-fluoro-5-(trifluoromethyl)benzaldehyde	469	480.13	484.56	0.285	11.13	15.56	0.52	0.73
29	3-Chlorobenzaldehyde	486.5	481.99	481.71	0.059	-4.51	-4.79	-0.18	-0.2
30	3-Cyanobenzaldehyde	484	504.15	505.75	0.074	20.15	21.75	0.83	0.89
31	3-Ethoxy-2-hydroxycarboxaldehyde	537	547.66	554.74	0.399	10.66	17.74	0.54	0.91
32	3-Ethoxy-4-hydroxybenzaldehyde	558	549.08	542.85	0.411	-8.92	-15.15	-0.46	-0.78
33	3-Fluorobenzaldehyde	467	473.98	474.48	0.066	6.98	7.48	0.29	0.31
34	3-Hydroxybenzaldehyde	514	517.74	517.93	0.049	3.74	3.93	0.15	0.16
35	3-Methoxysalicylaldehyde	538.5	560.75	561.66	0.039	22.25	23.16	0.9	0.93
36	3-Nitrobenzaldehyde	560	563.55	563.73	0.048	3.55	3.73	0.14	0.15

Tableau-09-(Suite).

N	Composé	Y Exp.	Y-Calc	Y-Pred	Hat	Err.Calc	Err.Pred	Std.Err. Calc.	Std.Err. Pred.
37	3-Tolualdehyde	496	471.36	466.76	0.157	-24.64	-29.24	-1.06	-1.26
38	4-(Dimethylamino)benzaldehyde	583	566.11	563.14	0.149	-16.89	-19.86	-0.72	-0.85
39	4,6-Dimethoxy-2-hydroxybenzaldehyde	643.5	599.61	596.24	0.071	-43.89	-47.26	-1.8	-1.94
40	4-Biphenylcarboxaldehyde	599.6	641.24	647.97	0.139	41.64	48.37	1.78	2.06
41	4-Cyanobenzaldehyde	563	550.02	548.64	0.096	-12.98	-14.36	-0.54	-0.6
42	4-Ethylbenzaldehyde	494	472.24	464.97	0.251	-21.76	-29.03	-0.99	-1.33
43	4-Fluorobenzaldehyde	455	479.67	481.3	0.062	24.67	26.3	1.01	1.07
44	4-Hydroxy-3-nitrobenzaldehyde	548	585.47	589.24	0.091	37.47	41.24	1.55	1.71
45	4-Methyl-1-naphthaldehyde	612	601.59	600.63	0.084	-10.41	-11.37	-0.43	-0.47
46	4-Phenoxybenzaldehyde	593	611.57	618.07	0.259	18.57	25.07	0.85	1.15
47	5-Bromosalicylaldehyde	520.5	502.62	501.47	0.06	-17.88	-19.03	-0.73	-0.78
48	5-Chlorosalicylaldehyde	490.5	489.1	489.01	0.065	-1.4	-1.49	-0.06	-0.06
49	5-Hydroxy-2-nitrobenzaldehyde	646	607.29	603.3	0.093	-38.71	-42.7	-1.61	-1.77
50	Benzaldehyde	540	492.93	488.63	0.084	-47.07	-51.37	-1.95	-2.12
51	Phenanthrene-9-carboxaldehyd	678.5	658.54	651.75	0.254	-19.96	-26.75	-0.91	-1.22
52	Phenyl-1,3-dialdehyde	519	524.26	524.56	0.053	5.26	5.56	0.21	0.23

Tableau-09-(Suite et fin).

N	Composé	Y Exp	Y-Cal	Y-Pre	Hat	Err.Ca	Err.Pre	Std.Er Calc.	Std.Er Pred
53	p-Tolualdehyde	477	498.4	501.0	0.108	21.46	24.07	0.9	1.01
54	Terephthaldicarboxaldehyde	580.5	575.7	575.2	0.095	-4.73	-5.22	-0.2	-0.22

7. Diagramme de williams :

On a représenté, sur la même figure 11 ; pour les deux ensembles (calibration et validation) ; Le domaine d'application a été discuté à l'aide de Diagramme Williams qui représente les résidus de prédiction standardisés en fonction des valeurs des leviers hii

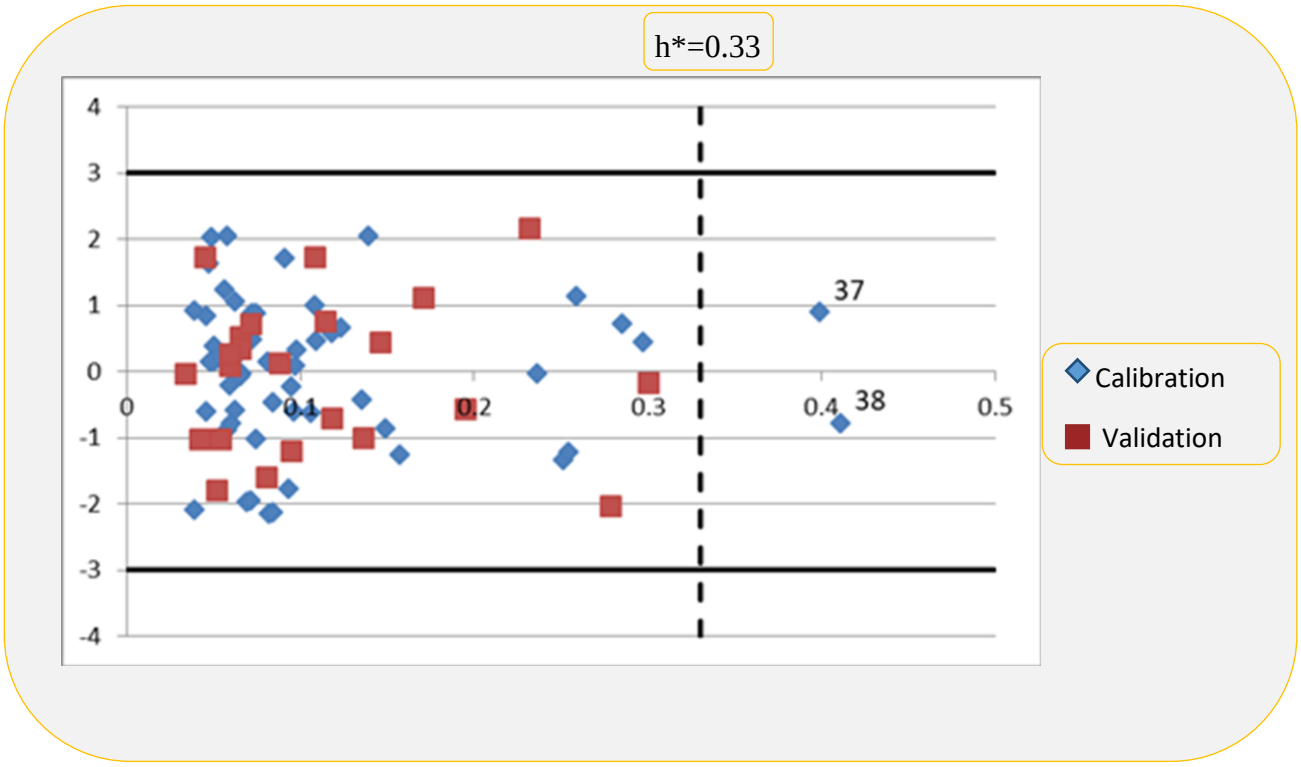


Figure- 11- Diagramme de Williams

D'après le diagramme en voire que toutes les composées sont comprise entre les limites ±3 ; alors en observe l'absence d'un point aberrant à l'exception de 37 et 38 qui représente un point levier supérieure à **h** critique parce que c'est un point influent.

Les deux points cités précédemment appartiennent à l'ensemble de calibration.

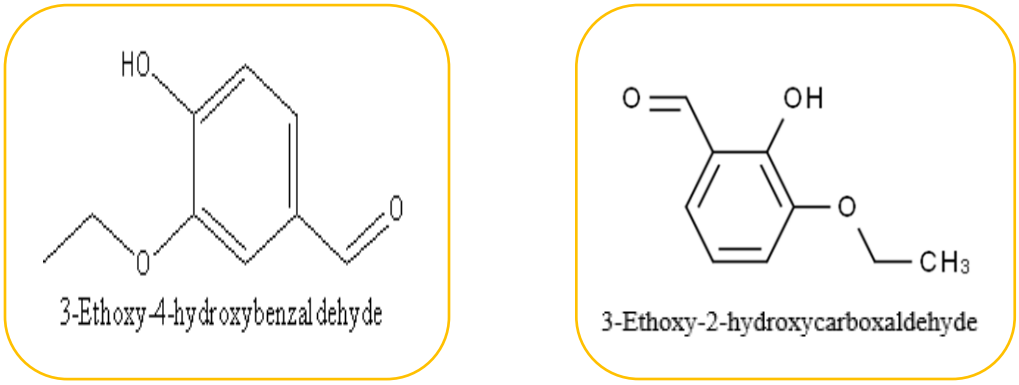


Figure- 12- Les deux composés influents

8. Vérification de la qualité de l'ajustement :

La précision de l'ajustement a été évaluée en comparant les valeurs prédites de Teb à l'aide de notre modèle avec les valeurs observées ou expérimentales. La (figure-13-) illustre un ajustement satisfaisant qui se traduit par une dispersion limitée autour de la première bissectrice pour les deux ensembles (calibration et validation).

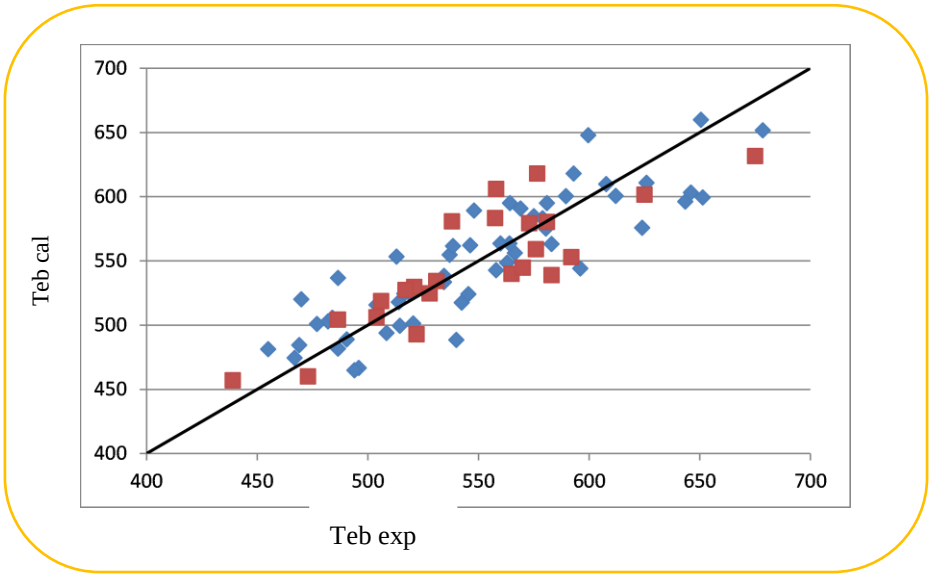


Figure-13 -Graphe des valeurs Teb calculées en fonction des valeurs expérimentales

9. Test de randomisation :

Dans le but d'établir que le modèle obtenu n'est pas dû au hasard ou à une sur spécification, nous avons appliqué le test de randomisation. Ainsi 100 nouveaux vecteurs de la température d'ébullition ont été générés par permutation des positions des composantes du vecteur réel. la figure -1- qui représente le graphe des coefficients statistiques Q^2 et R^2 permet de comparer les résultats obtenus pour les modèles randomisés (triangle vert) au modèle réel de départ (cercle rouge).

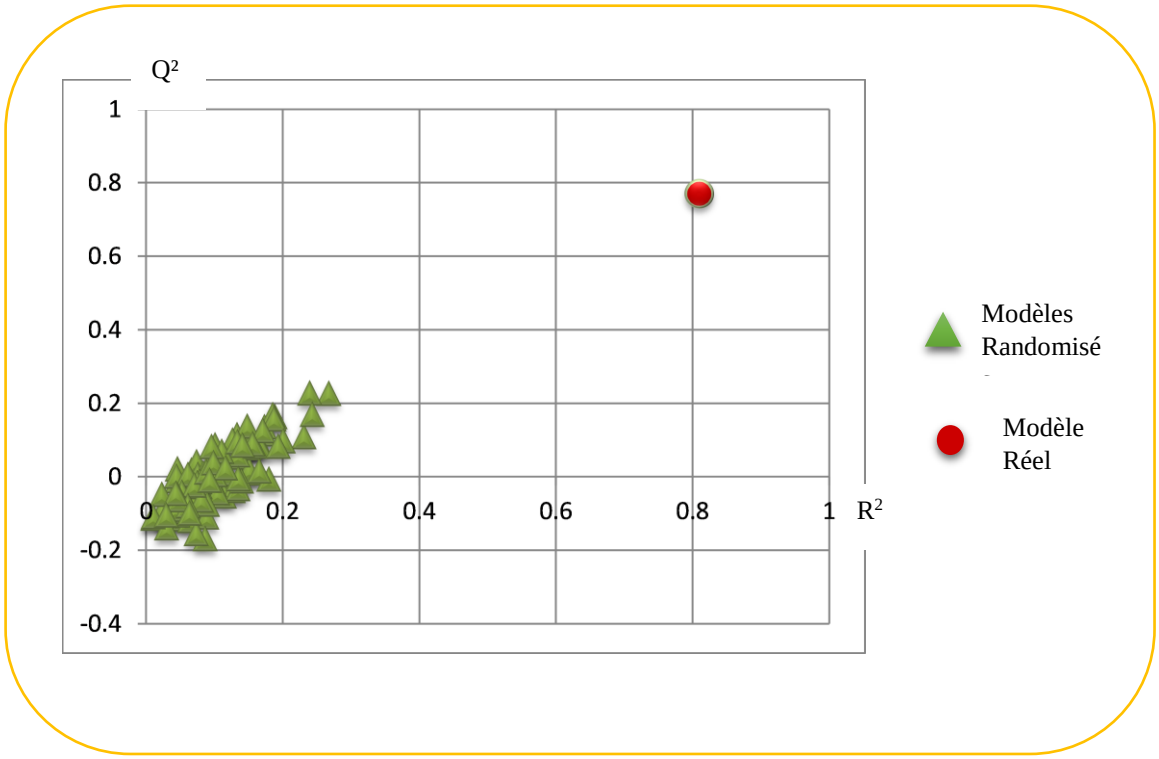


Figure- 14- Test de randomisation associé au modèle QSPR.

Les triangles verts représentent les températures d'ébullitions ordonnées de façon aléatoire, et le cercle rouge correspond au modèle réel.

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

10. Validation externe

Pour vérifier la généralisation du modèle sélectionné, nous avons effectué une validation externe en utilisant 23 composés sélectionnés au hasard, qui n'étaient pas inclus dans l'ensemble d'entraînement. Les résultats obtenus (tableau 10) indiquent que les valeurs prédites sont très proches des valeurs observées, ce qui confirme que le modèle choisi décrit de manière excellente la relation température d'ébullition et une série d'aldéhydes. Cette validation externe renforce la validité et la fiabilité du modèle pour prédire la température d'ébullition d'autres aldéhydes qui n'ont pas été inclus dans l'ensemble d'entraînement initial.

Egalement ce tableau regroupe les valeurs des érreurs standartiseés et hii qui ont sérvis pour chercher les points abérants (figure -11) pour la validation externe.

Tableau -10- Valeurs expérimentales, prédites et leurs erreurs pour l'ensemble de validation

N	Composé	Y Exp.	Y-Pred	Hat	Err.Pred.	Std.Err.Pred.
1	1-Naphthaldehyde	592	553.11	0.08	-38.89	-1.6
2	2,4-Dichlorobenzaldehyde	506	518.85	0.065	12.85	0.53
3	2-Bromobenzaldehyde	504	506.1	0.059	2.1	0.09
4	2-Chloro-6-fluorobenzaldehyde	531	534.33	0.088	3.33	0.14
5	2-Nitrobenzaldehyde	565	540.04	0.054	-24.96	-1.02
6	2-Tolualdehyde	473	460.21	0.195	-12.79	-0.56
7	3-Hydroxy-4-methoxybenzaldehyde	581	580.4	0.034	-0.6	-0.02
8	3-Hydroxy-4-nitrobenzaldehyde	576.5	617.99	0.108	41.49	1.74
9	3-Methoxy-4-hydroxybenzaldehyde	570	544.89	0.042	-25.11	-1.01
10	4-(Pentyloxy)benzaldehyd	675	631.68	0.278	-43.32	-2.02
11	4-Acetamidobenzaldehyde	557.5	583.54	0.171	26.04	1.13
12	4-Anisaldehyde	521	529.81	0.065	8.81	0.36
13	4-Bromobenzaldehyde	517	527.49	0.146	10.49	0.45

Tableau-10-(Suite et fin).

N	Composé	Y Exp.	Y-Pred	Hat	Err.Pred.	Std.Err.Pred.
14	4-Butoxybenzaldehyde	558	605.99	0.232	47.99	2.17
15	4-Chlorobenzaldehyde	486.5	504.31	0.071	17.81	0.73
16	4-Ethoxybenzaldehyde	528	524.76	0.3	-3.24	-0.16
17	4-Hydroxy-1-naphthaldehyde	538	580.76	0.045	42.76	1.73
18	4-Hydroxybenzaldehyde	583	538.93	0.052	-44.07	-1.79
19	4-Isopropylbenzaldehyde	576	559.26	0.118	-16.74	-0.7
20	4-Nitrobenzaldehyde	573	579.54	0.059	6.54	0.27
21	5-Bromovanillin	625	601.64	0.136	-23.36	-0.99
22	6-Chloro-2-fluoro-3-methylbenzaldehyde	522	493.06	0.095	-28.94	-1.2
23	Pentafluorobenzaldehyde	439	457.16	0.114	18.16	0.76

Toutes les valeurs des paramètres statistiques de validation sont regroupées dans le tableau ci-dessous :

Tableau -11-Valeurs des Q²ext et SDEPex

n	23
Q^2ext	76.45
$SDEP_{ext}$	26.982

La valeur de Q²ext est proche ou même supérieure à celle de la prédiction interne ce qui confirme que le modèle a une bonne capacité prédictive, également pour le SDEPext qui une Valeur proche à celle de SDEP.

Conclusion Générale

CONCLUSION GENERALE

Nous avons appliqué la méthodologie QSPR pour relier la propriété (température d'ébullition) d'un mélange hétérogène d'Aldéhydes, comportant un ou deux cycle(s) benzenique(s) ; à des descripteurs moléculaires théoriques reflétant certaines particularités.

Le modèle QSPR a été établis en utilisant l'analyse de régression multilinéaire MLR.

Les 77 données de base ont été éclatées aléatoirement en deux ensembles disjoints, invariants pour tous les modèles :

- un ensemble principal de 54 composés utilisés pour le calcul et, éventuellement, les essais du modèle ;
- un ensemble de 23 composés pour la prédiction externe.

La taille du modèle est fixée à 5 descripteurs en considérons les meilleurs paramètres statistiques (R^2 et Q^2). La sélection des variables explicatives a été réalisée par algorithme génétique, dans la version MOBYDIGS de TODESCHINI, en maximisant Q^2_{L00}.

Les paramètres statistiques (R^2>80% ; S=25.28 ; F=40.35) calculées établissent la pertinence du modèle QSPR développé.

L'analyse des résidus a permis de détecter deux observations influentes (37 et 38) qui ont des bras de levier supérieurs à la valeur critique (h*=0.33) par contre on a une absence totale des observations aberrantes.

A chaque fois, la qualité de l'ajustement a été vérifiée en procédant à une validation croisée par "leave – one - out". La valeur de Q^2 obtenue (>76 %) reproduise pratiquement celle du coefficient de détermination multiple, ce qui fait ressortir la qualité de l'ajustement du modèle obtenu.

Le test de randomisation montre, que seul le vecteur réel des observations conduit à des valeurs élevées des statistiques R^2 et Q^2, ce qui prouve que le modèle obtenu n'est pas aléatoire.

Les valeurs RMSE variants entre 23 et 27 sont acceptables et proches les unes des autres, ce qui permet de s'assurer de la bonne capacité prédictive et de la possibilité d'extension suffisante du modèle.

Ainsi, la propriété physicochimiques (Teb) peut être prédite à partir de leur structure moléculaire en utilisant la régression multilinéaire.

Enfin, pour améliorer d'avantage ces resultats nous pouvons étendre ce travail par d'autres méthodes non linéaires comme les réseaux de neurones qui peuvent s'avérer plus avantageuses en ce qui concerne la précision et l'interprétation des modèles, et du point de vue de la capacité de généralisation.

Références bibliographiques

Références bibliographiques

N°	Références
[1]	www.openai.com
[2]	A.F.A Cros, Action de l'alcool amylique sur l'organisme‖, thèse de doctorat, faculté de médecine, université Strasbourg, Strasbourg, 1863.
[3]	A.C. Crum-Brown and T.R. Fraser, On the Connection Between Chemical Constitution and Physiological Action, Part I: On the Physiological Action of the Salts of the Ammonium Bases, Derived from Strychnia, Brucia, Thebia, Codeia, Morphia, Nicotia", "Earth and Environmental Science Transactions of the Royal Society of Edinburgh, 25, 1868, 151–203.
[4]	M.C. Richet, " Noté sur le rapport entre la toxicité et les propriétés physiques des corps"," Comptes rendus des séances de la Société de biologie et de ses filiales", Paris, 45, 1893, 775– 6.
[5]	H. Meyer, ZurTheorie der Alkoholnarkose. Erste Mittheilung. WelcheEigenschaft der anäs the ticabeding tihrenar kotische Wirkung‖, Archivfür experimentelle pathologie end Pharmakologie, 42, 1899, 109–118.
[6]	E. Overton, ―Studienüber die Narkos ezugleice in B eitragzur allgemein en Pharmakologie, Ed. G. Fischer, Jena, 1901.
[7]	A - R.L. Lipnick, "Charles Ernest Overton: narcosis studies and a contribution togeneral pharmacology", Trends in Pharmacological Sciences, 7, 1986, 161–164. B - R.L. Lipnick, "Hans Horst Meyer and the lipoid theory of narcosis", Trends in Pharmacological Sciences, 10(7), 1989, 265–269.
[8]	H. Fühner and E. Neubauer, "ămolysedurch Sub stanzen homo logen Reihen", Archiv fürex perimentelle Pathologie and Pharmakologie, 56, 1907, 333–345.
[9]	O.R. Hansen, "Hammett Series with Biological Activity", ActaChemica scan dinavica, 16, 1962, 1593–1600.
[10]	C. Hansch and T. Fujita, «p-σ-π Analysis. A Method for the Correlation of Biological Activity and Chemical Structure", Journal of the American Chemical Society, 86(8), 1964, 1616–1626.
[11]	S.M. Free and J.W. Wilson, "A Mathematical Contribution to Structure-Activity studies", Journal of Medicinal Chemistry, 7(4), 1964, 395–399.

[12]	C. Hansch and E.J. Lien, Structure-activity relationships in antifungial agents. A survey, Journal of Medicinal Chemistry, 14(8), 1971, 653–670.
[13]	S.Y. Tham and S. Agatonovic-Kustrin, Application of the artificial neural network in quantitative structure-gradient elution retention relationship of phenyl thiocarbamy l amino acids derivatives, Journal of Pharmaceutical and Biomedical Analysis, 28(3), 2002, 581-590.
[14]	Hansch C. Lien E. J. Structure-activity relationships in antifungial agents. Surveyl, Journal of Medicinal Chemistry. 14(8). P 653- 670. (1971).
[15]	K. Roy, S. Kar, and R. N. Das, Understanding the Basics of QSAR for Applications in Pharmaceutical Sciences and Risk Assessment. (2015).
[16]	A.Z. Dudek, T. Arodz, J. Gàlvez. Computationml methods in developing quantitative structure-activityrelation ships (QSARs), a review, Combinatorial Chemistry and High Through put Screening, 2006, 9, PP: 213–228.
[17]	K. Dermiche. Etude in sillico de la thalidomide : Apport de la modélisation moléculaire. Thèse en vue de l'obtention du diplôme de doctorat en science. Université Mohamed Boudiaf d'Ouran. P 40. (2016).
[18]	V. Prana, P. Rotureau, G. Fayet, D. André, S. Hub, P. Vicot, L. Rao, C. Adamo, Predction of the thermal decomposition of organic peroxides by validated QSPR models, Journal of HazardousMaterials, 2014, 276, 216-224
[19]	HyperchemTM Release 6.03 for windows, Molecular Modeling system, (2000).
[20]	Todeschini. R, Consonni. V. Et Pavan .M, DRAGON, Software for the calculation of Molecular Descriptors. Release 5.3 for windows, Milano. 2005
[21]	Todeschini. R, Ballabio. D, Consonni. V, Mauri. A, Pavan. M. MOBYDIGS Software for Multilinear Regression Analysis and variable Subset Selection by Genetic Algorithm. Release I.1 for Windows. Milano. (2009).

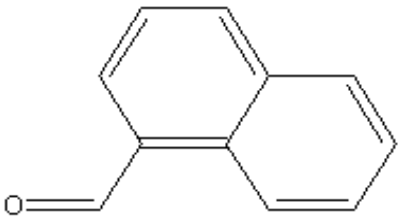
[22]	Searching scientific literature directly with Chem Draw v14 Chemistry World .29 July 2014.
[23]	Allen M. P., TildesleyD.J.,Computer stimulation of liquids .Oxford .1987
[24]	R. Todeschini, V. Consonni, M. Pawan, DRAGON, Software for the calculation of Molecular Descriptors. Release 5.3 for Windows, Milano, (2005).
[25]	B .Hoggas, Ch.Amamiche, Modélisation De La Température D'ébullition Des Alcanes En Utilisant Une Approche QSPR, Mémoire de Master (L.M.D), chimie analytique et environnement : p8, (2016)
[26]	Thomas-Danguin T. Intensité olfactive des composés purs et demélanges:application au masquage des odeurs, Université ClaudeBernard, Lyon, p224. (1997).
[27]	S .Remache, W.Redah, Etude QSRR de la rétention chromtographique des HAP, Mémoire de Master (L.M.D), chimie analytique et environnement, 2018 : p 20.
[28]	Martin G., Laffort P., Odeurs et désodorisations dans l'environnement, Lavoisier, Tec&Doc, Paris. (1991).
[29]	Dragon_ Aide blocs des descripteurs.
[30]	E. T. D. E. L'and R. Scientifique, "Élaboration des modèles QSPR prédictifs des propriétés physico- chimiques à l'aide des descripteurs moléculaires." (2015).
[31]	QSAR analysis: paradigm for drug design," Drug Discov. Today, vol. 21, no. 8, pp. 1291–1302, 2016, doi: 10.1016/j.drudis. 06.013. (2016).
[32]	Le Cloirec P. (2002). Introduction au traitement de l'air, Les techniques de l'ingénieur Traité environnement (G 1700) : 1-8.
[33]	Saadi K., Contribution á l'étude de la Relation structure chimique- odeur Utilisation de la technique Random Forest (Application à la famille des pyrazines), p 36, (2009).

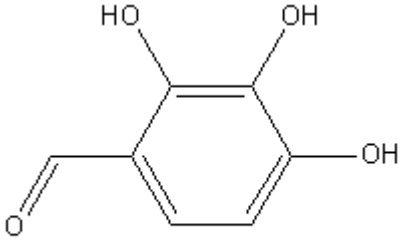
[34]	Saadi Khaled. Contribution l'étude de la Relation structure chimique- odeur Utilisation de la technique Random Forest (Application à la famille des pyrazines), Mémoire de Magister. UNIVERSITE KASDI Merbah Ourgla.p30. (2009).
[35]	AI ACCESS, 91940, Les Ulis, Franonce.
[36]	M.karelson. Molecular descriptors in QSAR/QSPR. Wiley- Inter science, 2000, 385.
[37]	B. Kowalski, R. Gerlach, H. Wold. Systems under Indirect Observation (K. Jörenskog et H. Wold, eds.), North Holland, Amsterdam, 1982, 191-206.
[38]	P. Gelada, B. R. Kowalski, Anal. Chim. Acta, 1986, (1), 185.
[39]	J. A. Burns, G. M. Whiteside. Chem. Rev., 1993, 93, 2583.
[40]	P. C. Jurs, Computer Software Applications in Chemistry. Second Edition, J. Wiely 1996.
[41]	R. Sirid. R. Wahiba Etude. QSRR de la rétention chromatographique des HAP. 2018,26.
[42]	A. R. Katritzky, V. S. Lobanov, M. Karelson. CODESSA Reference Manual. University of Florida, Gainesville, 1994.
[43]	P. Dagnélie. Statistique Théorique Et Appliquée. Tomes 1 et 2. De Boeck & Larcier s. a. 1998.
[44]	R. Tomassone, E. Lesquoy, C. Miller, 1983. La régression : nouveaux regards sur une ancienne méthode statistique. Masson, INRA .variables. Ecology 89(9) : 2623-2632.
[45]	L. Chambers. Practical Handbook of Genetic Algorithms. Lewis Publishing (1995).
[46]	MINITAB Release 16.2.0.0 for Microsoft langagepack 2.
[47]	R. Todeschini. MOBY DIGS Software for Multilinear Regression Analysis and Variable Subset Selection by Genetic Algorithm.Release for Windows. Milano Srl.
[48]	R. Sirid. R. Wahiba Etude. QSRR de la rétention chromatographique des HAP. 2018, 30.

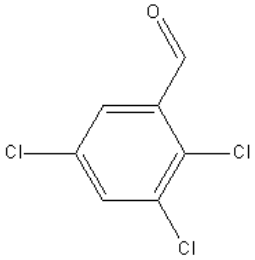
[49]	N.R Draper, H. Smith, Applied Regression Analysis, Third Edition, Wiley series in Probability and Statistics, New york, (1998).
[50]	L. Eriksson, J. Jaworska, A.P. Worth, M.T.D. Perspective, 111(10), 1361- 1375. (2003).

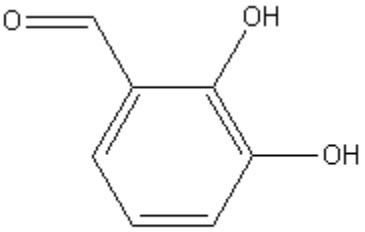
Annexes

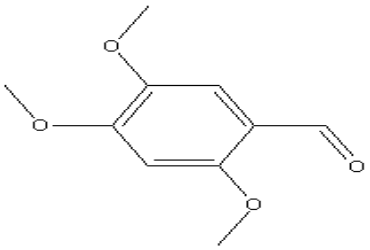
Annexes

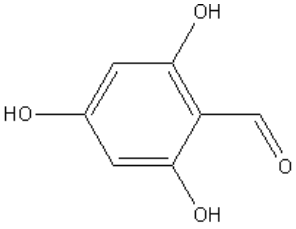

1-Naphthaldehyde

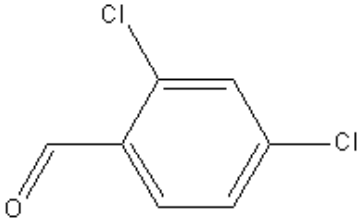

2,3,4-Trihydroxybenzaldehyde

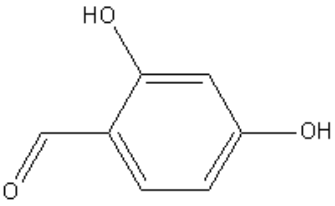

2,3,5-Trichlorobenzaldehyde

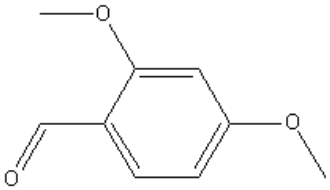

2,3-Dihydroxybenzaldehyde

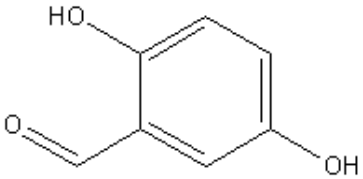

2,4,5-Trimethoxybenzaldehyde

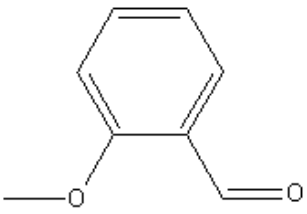

2,4,6-Trihydroxybenzaldehyde

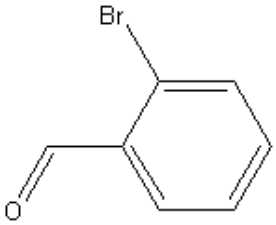

2,4-Dichlorobenzaldehyde

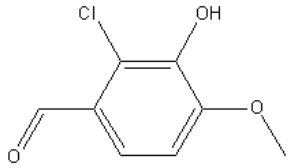

2,4-Dihydroxybenzaldehyde

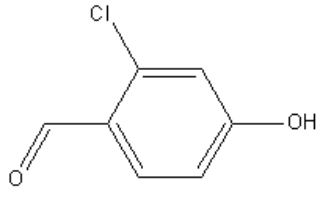

2,4-Dimethoxybenzaldehyde

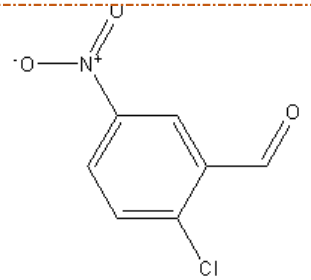

2,5-Dihydroxybenzaldehyde

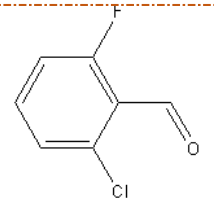

2-Anisaldehyde


2-Bromobenzaldehyde

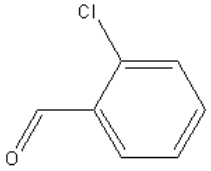

2-Chloro-3-hydroxy-4-methoxybenzaldehyde


2-Chloro-4-hydroxybenzaldehyde

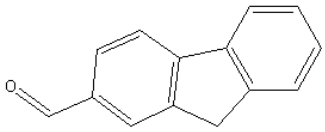

2-Chloro-5-nitrobenzaldehyde



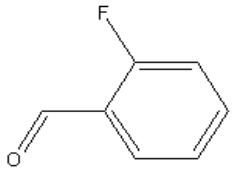
2-Chloro-6-fluorobenzaldehyde



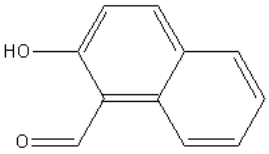
2-Chlorobenzaldehyde



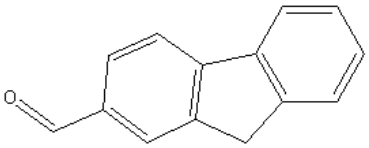
2-Fluorenecarboxaldehyde



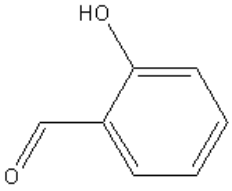
2-Fluorobenzaldehyde



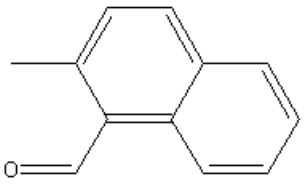
2-Hydroxy-1-naphthaldehyde



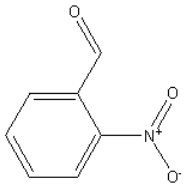
2-Fluorenecarboxaldehyde



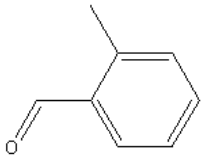
2-Hydroxybenzaldehyde



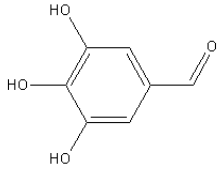
2-Methyl-1-naphthaldehyde



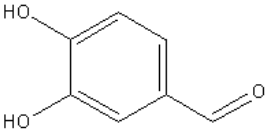
2-Nitrobenzaldehyde



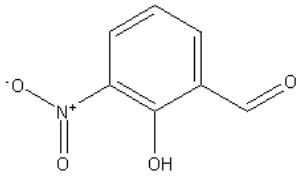
2-Tolualdehyde



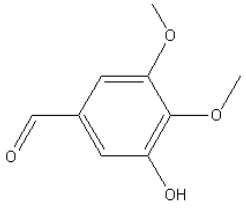
3,4,5-Trihydroxybenzaldehyde



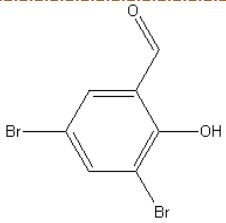
3,4-Dihydroxybenzaldehyde



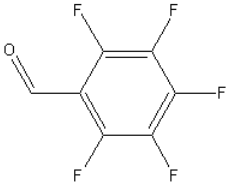
2-Hydroxy-3-nitrobenzaldehyde



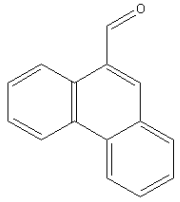
3,4-Dimethoxy-5-hydroxybenzaldehyde



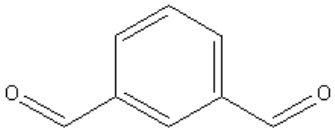
3,5-Dibromosalicylaldehyde



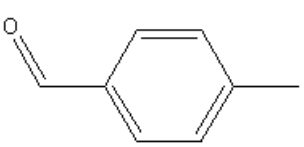
Pentafluorobenzaldehyde



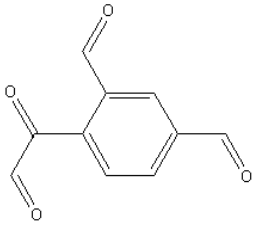
Phenanthrene-9-carboxaldehyd



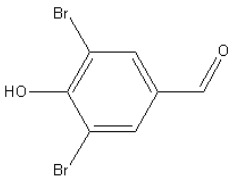
Phenyl-1,3-dialdehyde



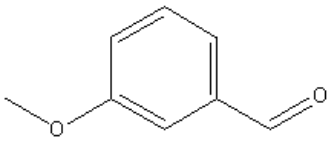
p-Tolualdehyde

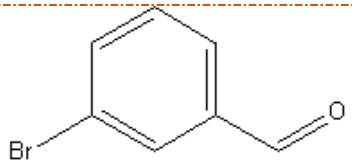


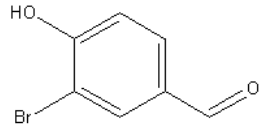
Terephthaldicarboxaldehyde

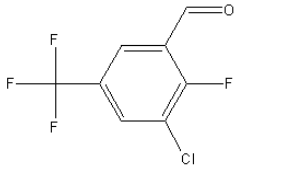


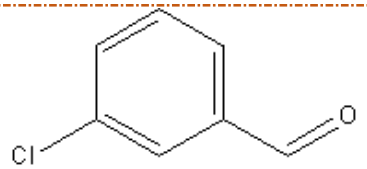
3,5-Dibromo-4-hydroxybenzaldehyde

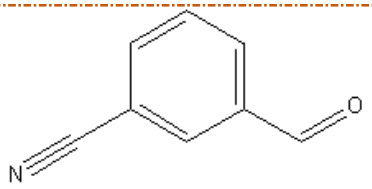

3-Anisaldehyde

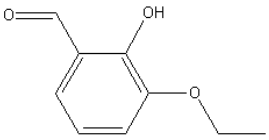

3-Bromobenzaldehyde

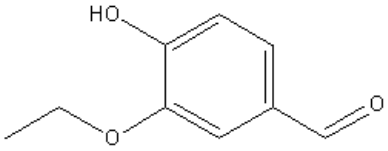

3-Bromo-4-hydroxybenzaldehyde

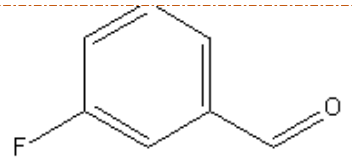

3-Chloro-2-fluoro-5-(trifluoromethyl)benzaldehyde

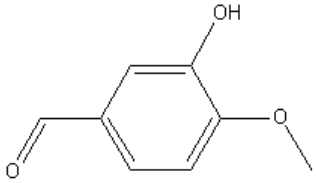

3-Chlorobenzaldehyde

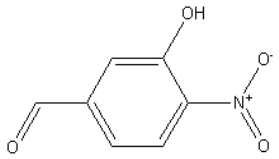

3-Cyanobenzaldehyde

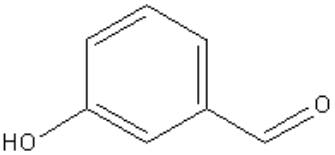

3-Ethoxy-2-hydroxybenzaldehyde

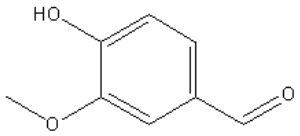

3-Ethoxy-4-hydroxybenzaldehyde

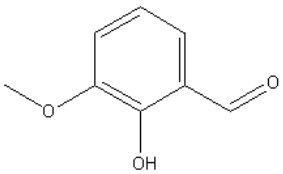

3-Fluorobenzaldehyde

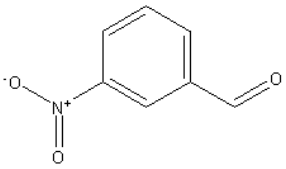

3-Hydroxy-4-methoxybenzaldehyde

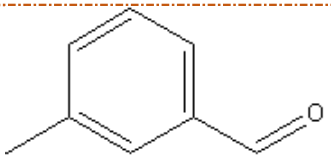

3-Hydroxy-4-nitrobenzaldehyde

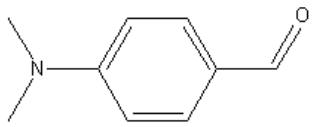

3-Hydroxybenzaldehyde

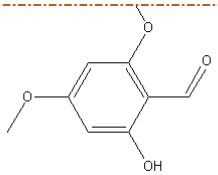

3-Methoxy-4-hydroxybenzaldehyde

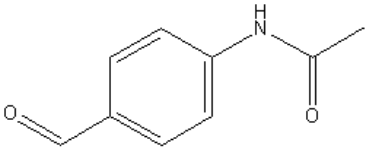

3-Methoxysalicylaldehyde

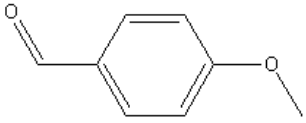

3-Nitrobenzaldehyde

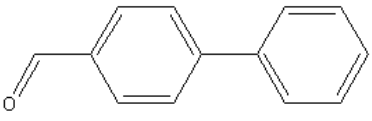

3-Tolualdehyde

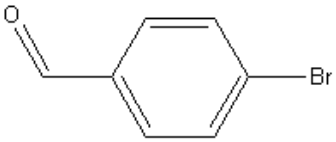

4-(Dimethylamino)benzaldehyde

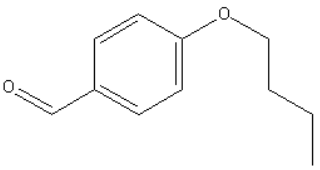

4,6-Dimethoxy-2-hydroxybenzaldehyde

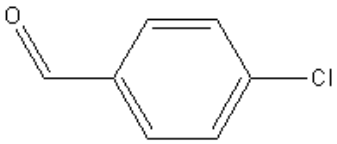

4-Acetamidobenzaldehyde

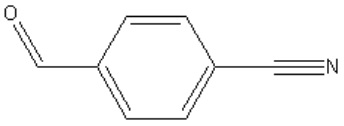

4-Anisaldehyde

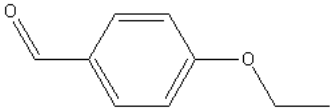

4-Biphenylcarboxaldehyde

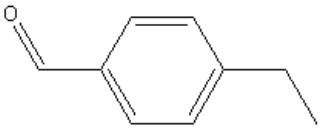

4-Bromobenzaldehyde

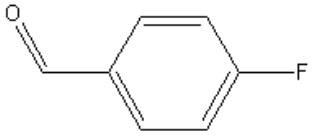

4-Butoxybenzaldehyde

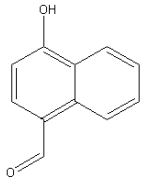

4-Chlorobenzaldehyde

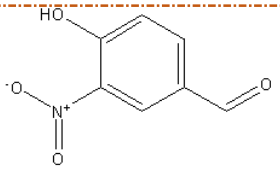

4-Cyanobenzaldehyde

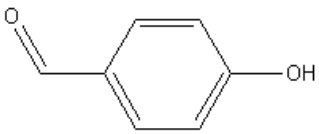

4-Ethoxybenzaldehyde

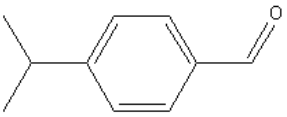

4-Ethylbenzaldehyde

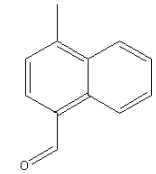

4-Fluorobenzaldehyde

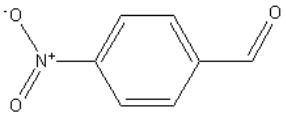

4-Hydroxy-1-naphthaldehyde

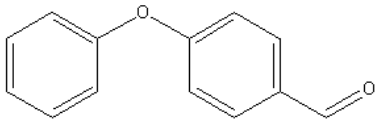

4-Hydroxy-3-nitrobenzaldehyde

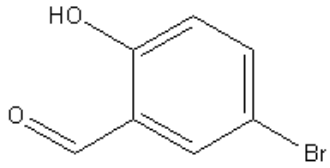

4-Hydroxybenzaldehyde

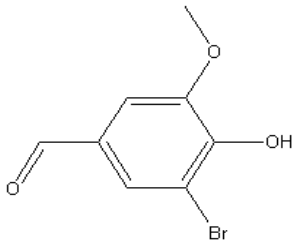

4-Isopropylbenzaldehyde

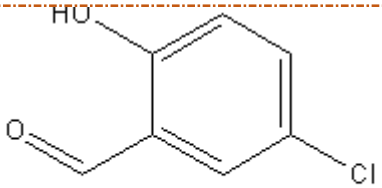

4-Methyl-1-naphthaldehyde

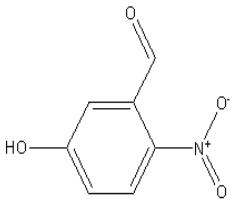

4-Nitrobenzaldehyde

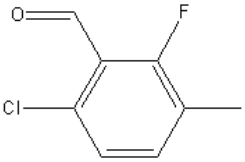

4-Phenoxybenzaldehyde

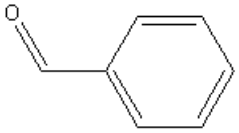

5-Bromosalicylaldehyde


5-Bromovanillin


5-Chlorosalicylaldehyde


5-Hydroxy-2-nitrobenzaldehyde


6-Chloro-2-fluoro-3-methylbenzaldehyde


Benzaldehyde

Résumés

Résumé:

Un modèle QSPR a été développé pour la prédiction de la température d'ébullition d'une série d'aldéhydes comportant un ou deux cycle(s) benzénique(s). Les 77 données ont été séparées en deux sous-ensembles disjoints comprenant respectivement 54 éléments pour le calcul et le test (éventuel) du modèle, et 23 éléments pour sa validation externe.

Des descripteurs moléculaires théoriques ont été calculés en utilisant le logiciel (Dragon) de modélisation moléculaire du commerce. La taille du modèle à été déterminée en prenant la valeur optimale du coeficient de determination R^2, et la sélection des descripteurs réalisée par algorithme génétique.

Les valeurs des paramètres statistiques (R^2, Q^2, SDEC, SDEP, SDEPext) obtenues attestent de la pertinence du modèle QSPR développé.

Mots-clés:

Aldéhydes – Température d'ébullition – Descripteurs moléculaires théoriques – Representation numérique de la structure chimique – Modèle QSPR.

Abstract:

A QSPR model has been developed for the prediction of the boiling temperature of a series of aldehydes comprising one or two benzene ring(s). The 77 data were separated into two disjoint subsets comprising respectively 54 elements for the calculation and (possible) test of the model, and 23 elements for its external validation.

Theoretical molecular descriptors were calculated using commercial molecular modeling software (Dragon). The size of the model was determined by taking the optimal value of the coefficient of determination R^2, and the selection of the descriptors carried out by genetic algorithm.

The values of the statistical parameters (R^2, Q^2, SDEC, SDEP, SDEPext) obtained attest to the relevance of the QSPR model developed.

Key words:

Aldehydes – Boiling temperature – Theoretical molecular descriptors – Numerical representation of the chemical structure – QSPR model.

ملخص:

تم تطوير نموذج QSPR للتنبؤ بدرجة حرارة الغليان لسلسلة من الألدهيدات تشتمل على حلقة (حلقات) بنزين واحدة أو اثنتين. تم فصل البيانات الـ 77 إلى مجموعتين فرعيتين منفصلتين تشتملان على التوالي على 54 عنصرًا للحساب والاختبار (المحتمل) للنموذج، و23 عنصرًا للتحقق الخارجي.

تم حساب الواصفات الجزيئية النظرية باستخدام برنامج النمذجة الجزيئية التجارية .(Dragon) تم تحديد حجم النموذج بأخذ القيمة المثلى لمعامل التحديدR^2 ، واختيار الواصفات بواسطة الخوارزمية الجينية

قيم المعلمات الإحصائية(R^2 ، Q^2، SDEC، SDEP، SDEPext) التي تم الحصول عليها تشهد على أهمية نموذج QSPRالذي تم تطويره.

الكلمات المفتاحية:

الألدهيدات -درجة حرارة الغليان -الواصفات الجزيئية النظرية -التمثيل العددي للتركيب الكيميائي -نموذج.QSPR