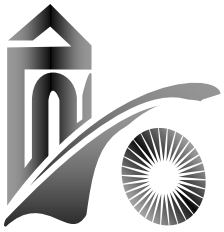
République Algérienne Démocratique et Populaire
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

UNIVERSITE YAHIA FARES DE MEDEA



Faculté de Technologie

Département de Génie des Procédés et Environnement

Mémoire de Fin d'Etude Pour l'Obtention du Diplôme de **Master** en **Génie des Procédés**

Option : **Génie des Procédés de l'Environnement**

THEME

Modélisation d'un paramètre pertinent des phénomènes de transport lié à la pollution atmosphérique et/ou la pollution des rejets industriels

Réalisé par :

M. HASSNAOUI Takieddine
M^{elle}. BENMOUSSA Samia

Proposé et dirigé par :

Dr. KHAOUANE Latifa
Pr. HANINI Salah
M. ROUBEHIE FISSA Mohamed

Promotion: 2019-2020

الحمد لله كما ينبغي

لجلال وجهه وعظيم سلطانه

Dédicace

A l'homme de ma vie, mon exemple éternel, mon soutien moral et source de joie et de bonheur, celui qui s'est toujours sacrifié pour me voir réussir, à toi mon père.

A la lumière de mes jours, la source de mes efforts, la flamme de mon cœur, ma vie et mon bonheur ; maman que j'adore.

À mes frères Ahmed, Ali, Mustapha, Sofiane, Amine, Brahim et mes sœurs Hamida, Samia.

Je dédie ce travail dont le grand plaisir leurs revient en premier lieu pour leurs conseils, aides, et encouragements.

À MES CHERS PETITS Nièces et neveux

Aucune dédicace ne saurait exprimer tout l'amour que j'ai pour vous, Votre joie et votre gaieté me comblent de bonheur.

A mes belles sœurs

A mes sœurs de cœur

Qui m'ont toujours aidé et encouragé, qui étaient toujours à mes côtés, et qui m'ont accompagné durant mon chemin mes loulous.

À MES CHERS ONCLES, TANTES,

A MES CHERS COUSINS COUSINES

Je vais remercier mon binôme Takieddinne de m'avoir fait confiance et d'avoir tenté l'aventure avec moi, Ce travail n'aurait pas pu voir le jour sans toi.

Sans oublier toute la promotion de Master Génie des Procédés de l'Environnement qui ont passé avec moi de bons moments

BENMOUSSA Samia

Dédicace

Après de longues années d'études, sachant l'importance de l'aide des êtres qui m'aiment, je voudrais humblement leurs dédier ce modeste travail tout en sachant qu'ils méritent le meilleur qui soit.

Je dédie ce travail :
À mes premiers professeurs dans la vie, ma tendre mère et mon cher père, qui mon comblés d'amour et d'affection, qui mon toujours encouragé et soutenus durant toutes mes années d'études et dans les moments les plus difficiles.

À mes chères sœurs Zahra et Sara

À toute ma famille maternelle et paternelle.

À ma binôme Samia et sa famille

À Mes chères amis Ali, Rim et Hanane

À ma promotion 2019/2020 sans exception.

À toutes les personnes que je n'ai pas citées mais que je porte dans mon cœur.

A tous ceux qui m'ont aidé, de près ou de loin, même qu'il soit un mot D'encouragement ET de gentillesse

HASSNAOUI Takieddine

Remerciements et gratitude

*Tout d'abord, nous remercions **ALLAH** de nous avoir donné la santé, la patience et les moyens, à fin que nous puissions accomplir ce travail.*

*Nous avons eu le grand plaisir de travailler sous la direction du **Dr. KHAOUANE latifa** Maitre de Conférence classe A à l'université de Médéa nous remercions la pour sa patience, sa disponibilité, sa modestie, sa gentillesse, sa richesse d'expériences, qui nous ont guidés dans notre travail.*

*Un grand merci à notre Co-promoteur **Monsieur ROUBEHIE FISSA Mohamed**, ce travail était d'autant plus agréable grâce à sa gentillesse et ses encouragements, ses efforts pour le bon déroulement de ce travail, et pour son aide précieuse à l'élaboration de ce mémoire.*

*Nous sommes profondément reconnaissant au **Pr. HANINI Salah**, d'avoir suivi et guidé avec bienveillance notre travail. Nous le remercions vivement pour son intérêt.*

*Nous tenons à exprimer nos plus vifs remerciements aux **Membres de jury** de nous avoir honorés par leur présence et d'avoir accepté d'examiner notre modeste travail. Nous leurs devons les remerciements les plus sincères.*

*A **Tous les Enseignants** du département de **Génie des Procédés et Environnement**, on vous dit merci pour la qualité de votre enseignement qui restera gravé dans nos mémoires.*

Enfin, nos remerciements vont à toutes les personnes qui ont contribué de près ou de loin à la réalisation de ce mémoire.

ملخص: نمذجة معلمة ذات صلة بظواهر الإنتقال المرتبطة بتلوث الغلاف الجوي و/أو التلوث الناجم عن التصريفات الصناعية

الهدف الرئيسي من هذه الدراسة هو نمذجة معلمة ذات صلة بظواهر الإنتقال المرتبطة بتلوث الغلاف الجوي و/أو التلوث الناجم عن التصريفات الصناعية. أجريت الدراسة لتطوير نماذج رياضية من نوع (RQSP-RNA) و (RNA) ، من أجل التنبؤ بمعامل الإنتشار D_{AB} لملوثات مختلفة في خمسة أنظمة تجريبية مختلفة، باستخدام شبكة العصبونات الإصطناعية- ذات مستقبلات متعدد الطبقات (RNA-PMC) ونهج RQSP. تم فيها تطوير خمسة نماذج للتنبؤ بخمسة معاملات إنتشار مختلفة : D_{AB} للأزواج *ملوثات* في مختلف °T، D_{AB} للأزواج *ملوثات-هواء* في T=25°C، D_{AB} للأزواج *ملوثات-ماء* في T=25°C، D_{AB} للأزواج *ملوثات-هواء* في T=100°C، D_{AB} للأزواج *ملوثات-ماء* في T=100°C. تم استخدام قاعدتي بيانات : الأولى تحتوي على 148 نقطة تجريبية تستخدم في تطوير أربعة نماذج من نوع RQSP-RNA، والثانية تتضمن 727 نقطة تجريبية تستخدم في تطوير نموذج من نوع RNA. تم تقسيم القاعدتين بشكل عشوائي إلى مجموعتين وثلاث مجموعات فرعية على التوالي ؛ 80٪ و 20٪ لمجموعتي التدريب والإختبار لقاعدة البيانات الأولى، ثم 70٪ و 15٪ و 15٪ لمجموعات التدريب، الإختبار والتحقق لقاعدة البيانات الثانية. تم حساب 1970 واصفا جزيئيا بواسطة برنامج *PaDEL Descriptor* عبر الإنترنت، ثم تم إعتماد منهجية تخفيض إحصائية لتحديد الواصفات ذات الصلة. تم تطبيق خوارزمية BFGS لتدريب RNAs. كانت بنية أفضل النماذج التي تم الحصول عليها كالتالي : "7-24-1"، "20-2-1"، "11-2-1"، "16-6-1" و "9-10-1" ؛ على التوالي لنماذج : $D_{AB_\text{للأزواج}}$ عند T° مختلفة، $D_{AB_\text{هواء}}$ عند T=25°C، $D_{AB_\text{ماء}}$ عند T=25°C، $D_{AB_\text{هواء}}$ عند T=100°C، $D_{AB_\text{ماء}}$ عند T=100°C. كانت النتائج مشجعة للغاية أكدتها قيم معامل (R^2) المرتفعة والتي تجاوزت 0.92 للنماذج الخمسة، والقيم المنخفضة لـ (ESP) التي لم تتجاوز 14.2423٪ للنموذج العصبي $D_{AB_\text{للأزواج}}$ عند T° مختلفة، و 6.8527٪ للنماذج الأربعة المتبقية من نوع RQSP-RNA. أظهرت نتائج تحليل الحساسية بوضوح أهمية معلمات الإدخال المستخدمة في كل نموذج. يمكن تطبيق هذا النهج الحسابي في الهندسة البيئية أو المجالات الأخرى ذات الصلة لتحديد معامل D_{AB} للملوثات بدقة من خلال إستخدام المعلمات المتعلقة بالتركيبات الكيميائية والظواهر الفيزيائية.

كلمات مفتاحية: تقدير معلمات D_{AB}، ملوثات، نماذج (RQSP-RNA) و (RNA)، معلمات و واصفات الإدخال.

Résumé : Modélisation d'un paramètre pertinent des phénomènes de transport lié à la pollution atmosphérique et/ou la pollution des rejets industriels

L'objectif principal de cette étude est la modélisation d'un paramètre pertinent des phénomènes de transport lié à la pollution atmosphérique et/ou la pollution des rejets industriels. L'étude a été apportée pour développer des modèles mathématiques de type (RQSP-RNA) et (RNA), afin de prédire le coefficient de diffusion D_{AB} des différents polluants dans cinq systèmes expérimentales distincts, utilisant le Réseau de Neurone Artificiel-Perceptron Multi Couches (RNA-PMC) et l'approche des RQSP. Dans lequel, cinq modèles ont été élaborer pour prédire cinq coefficients de diffusion différents : D_{AB} des couples *Polluants* à différents T°, D_{AB} des couples *Polluants-Air* à T=25°C, D_{AB} des couples *Polluants-Eau* à T=25°C, D_{AB} des couples *Polluants-Air* à T=100°C et D_{AB} des couples *Polluants-Eau* à T=100°C. Deux bases des données ont été utilisées : la première contient 148 points expérimentales utilisées pour développées quatre modèles de type RQSP-RNA, et la deuxième englobe 727 points expérimentales utilisées pour développer un modèle de type RNA. Les deux bases ont été divisée de manière aléatoire en deux et en trois sous-ensembles respectivement ; 80% et 20% pour les ensembles d'apprentissage et de test pour la première base de données, puis 70%, 15% et 15% pour les ensembles d'apprentissage, de test et de validation pour la deuxième base de données. 1970 descripteurs moléculaires ont été calculées à partir du logiciel en ligne *PaDEL Descriptor*, et ensuite une méthodologie statistique de réduction a été adoptée pour sélectionner les descripteurs pertinents. L'algorithme BFGS a été appliqué pour entraîner les RNAs. L'architecture des meilleurs modèles obtenus était : "7-24-1", "20-2-1", "11-2-1", "16-6-1", et "9-10-1" ; respectivement pour les modèles de : $D_{AB_Couples}$ à différents T°, D_{AB_Air} à T=25°C, D_{AB_Eau} à T=25°C, D_{AB_Air} à T=100°C et D_{AB_Eau} à T=100°C. Les résultats étaient très encourageants confirmée par les valeurs de coefficient (R^2) élevées, qui dépassent 0,92 pour les cinq modèles, et les faibles valeurs de (ESP) qui ne dépassent pas 14,2423% pour le modèle neuronal de $D_{AB_Couples}$ à différents T°, et 6,8527% pour les quatre modèles restants de type RQSP-RNA. Les résultats de l'analyse de sensibilité ont montré bien l'importance des paramètres d'entrées utilisées dans chaque modèle. Cette approche computationnelle peut être appliquée en l'ingénierie environnemental ou d'autres domaines reliés pour une détermination précise de coefficient D_{AB} des polluants par l'utilisation des paramètres liés aux structures chimiques et aux phénomènes physiques.

Mots clés : Estimation des Coefficients D_{AB}, Polluants, Modèles (RQSP-RNA) et (RNA), Paramètres et Descripteurs d'Entrées.

Abstract: Modeling of a relevant parameter of transfer phenomena linked to atmospheric pollution and/or pollution of industrial discharges

The main objective of this study is the modeling of a relevant parameter of transport phenomena linked to atmospheric pollution and/or pollution from industrial discharges. The study was made to develop mathematical models of (RQSP-RNA) and (RNA) types, in order to predict the D_{AB} diffusion coefficient of different pollutants in five distinct experimental systems, using the Artificial Neuron Network-Multi Layer Perceptron (ANN-MLP) and the QSPR approach. In which, five models have been developed to predict five different diffusion coefficients : D_{AB} of *Pollutant* pairs at different T°, D_{AB} of *Pollutant-Air* pairs at T=25°C, D_{AB} of *Pollutant-Water* pairs at T=25°C , D_{AB} of *Pollutants-Air* pairs at T=100°C and D_{AB} of *Pollutants-Water* pairs at T=100°C. Two databases were used : the first contains 148 experimental points used to develop four RQSP-RNA models type, and the second includes 727 experimental points used to develop an RNA model type. The two databases were randomly divided into two and three subsets respectively; 80% and 20% for the learning and testing sets for the first database, and then 70%, 15% and 15% for the learning, testing and validation sets for the second database. 1970 molecular descriptors were calculated from the online *PaDEL Descriptor* software, and then a statistical reduction methodology was adopted to select the relevant descriptors. The BFGS algorithm was applied to train the RNAs. The architecture of the best models obtained was : "7-24-1", "20-2-1", "11-2-1", "16-6-1", and "9-10-1" ; respectively for the models of : $D_{AB_Couples}$ at différent T°, D_{AB_Air} at T=25°C, D_{AB_Water} at T=25°C, D_{AB_Air} à T=100°C et D_{AB_Water} à T=100°C. The results were very encouraging confirmed by the high coefficient values of (R^2), which exceed 0.92 for the five models, and the low values of (ESP) which do not exceed 14.2423% for the neural model of $D_{AB_Couples}$ at different T°, and 6.8527% for the four remaining QSPR-ANN models type. The results of the sensitivity analysis clearly showed the importance of the input parameters used in each model. This computational approach can be applied in environmental engineering or other related fields for accurate determination of D_{AB} coefficient of pollutants by using parameters related to chemical structures and physical phenomena.

Keywords: Estimation of D_{AB} Coefficients, Pollutants, (QSPR-ANN) and (ANN) Models, Inputs Parameters and Descriptors.

Table des matières

Résumé

Table des matières *i*
Abréviations *v*
Nomenclature et symboles *vi*
Liste des tableaux *viii*
Liste des figures *ix*

Introduction générale 01

Chapitre I. Génialités sur la pollution atmosphérique et la pollution des rejets industriels

I.1. Introduction 04
I.1.1. Importance de l'environnement 04
I.1.2. Echauffement climatique 04
I.1.3. Causes de pollution 04
I.1.4. Grands types de pollution 05
I.1.4.1. Pollution des sols 05
I.1.4.2. Pollution de l'air ou atmosphérique 05
I.1.4.3. Pollution de l'eau 05
I.1.5. Structure de ce chapitre 06
I.2. Classifications des Pollution 06
I.2.1. Pollution de l'eau 06
I.2.2. Pollution du sol 07
I.2.3. Pollution de l'air 07
I.3. Pollution atmosphérique 08
I.3.1. Définition 08
I.3.2. Source de pollution atmosphérique 08
I.3.3. Effet nocive de ces éléments sur l'environnement et les êtres vivants 09
I.3.4. Solutions pour éviter ou réduire la pollution atmosphérique 11
I.4. Pollution des rejets industriels 11
I.4.1. Définition 11
I.4.2. Source des rejets industriels 12
I.4.2.1. Production d'énergie et source de pollution 12
I.4.2.2. Industrie chimique moderne source de polluants variés 12
I.4.2.2.1. Agents polluants 12
I.4.2.2.2. Dispersion planétaire de certains toxiques 13
I.4.2.2.3. Effets nocifs de ces éléments sur l'environnement et sur les êtres vivants 13
I.4.2.3. Solution pour éviter ou réduire la pollution issue des rejets industriels 15
I.5. Phénomènes de transport liés à la pollution atmosphérique et la pollution des rejets industriels 15
I.5.1. Définition 15

I.5.2. Le transfert de chaleur 15
I.5.3. Transfert de matière 16
I.5.4. Transfert de quantité de mouvement 16
I.6. Paramètres lier à la diffusion des polluants dans l'environnement 16
I.6.1. Paramètres de diffusion lies à la pollution 16
I.6.1.1. Coefficients de diffusion des liquides et des gaz polluants 16
I.6.2. Transfert thermique 17
I.6.3. Coefficients de transfert de chaleur 18
I.6.3.1. Différents modes de transmission de la chaleur 18
I.6.3.2. Conduction thermique et la loi de Fourier 18
I.6.3.3. Conductivité thermique 18
I.6.4. Vitesses d'écoulement 19
I.6.5. Viscosité 20
I.6.6. Masse volumique 20
I.7. Conclusion 20

Chapitre II. Techniques de modélisation neuronal et méthodes de RQSP/RQSA

II.1. Introduction 22
II.1.1. Modélisation 22
II.1.2. L'évolution de technique neuronale 22
II.1.2.1 Applications et leurs importances 22
II.1.3. L'évolution des méthodes de type RQSP/RQSA 23
II.1.3.1 Applications et leurs importances 24
II.2 Objectifs de la modélisation 24
II.3. Quelques techniques de la modélisation 25
II.3.1. Régression linéaire multiple 25
II.3.2. Modèle de surface de réponse 25
II.4. Réseaux de neurones artificiels 26
II.4.1. Neurone 26
II.4.2. Bases biologiques 26
II.4.3. Neurone formel 27
II.4.3.1. Définition 27
II.4.3.2. Paramètres 28
II.4.3.3. Principe de fonctionnement 28
II.4.4. Architecture des réseaux de neurones 29
II.4.4.1. Réseaux statiques 29
II.4.4.2. Réseaux dynamiques 29
II.4.4.3. Réseaux autoorganisés 30
II.4.5. Apprentissage 30
II.4.5.1. Définition 30
II.4.5.2. Apprentissage supervisé 30
II.4.5.3. Apprentissage non supervisé 30
II.4.6. Perceptron mono couche 31
II.4.7. Perceptron multicouche 31
II.5. Modélisation par les approches RQSP/RQSA 32
II.5.1. Définition de RQSP 33
II.5.2. Méthodologie de base 33
II.5.3. Notation SMILES 33

II.5.4. Structure chimique des molécules 35
II.5.4.1 Atomes 35
II.5.4.2. Liaisons 35
II.5.4.3. Succursales 35
II.5.5. Descripteurs moléculaires 36
II.5.5.1. Descripteurs 0D 37
II.5.5.2. Descripteurs 1D 37
II.5.5.3. Descripteurs 2D 37
II.5.5.4. Descripteurs 3D 38
II.5.5.5. Descripteurs 4D 38
II.5.6. Calcule des descripteurs moléculaires 38
II.6. Techniques d'analyse et de validation des modèles de RNA 38
II.6.1. Détermination des entrées/sorties du réseau de neurones 39
II.6.2. Choix et préparation des échantillons 39
II.6.3. Elaboration de la structure du réseau 40
II.6.4. Apprentissage 40
II.6.5. Validation et tests 40
II.7. Techniques d'analyse et de validation des modèles de QSPR 41
II.7.1. Coefficients et tests statistiques standards 41
II.7.1.1 Coefficient de corrélation (r) et de détermination (r^2) 41
II.7.1.2. Le coefficient de détermination ajusté (r^2 adj) 41
II.7.1.3. L'erreur quadratique moyenne (MSE) et l'erreur type résiduel (S) 42
II.7.1.4. Facteur d'inflation de la variance (VIF) 42
II.7.1.5. Test de Fisher (F) 42
II.8. Conclusion 43

Chapitre III. Modélisation d'un paramètre pertinent des phénomènes de transport lié à la pollution atmosphérique et/ou la pollution des rejets industriels

III.1. Introduction 44
III.2. Matériels et méthodes 44
III.2.1. Outils informatiques utilisés 44
III.2.2. Méthodologie générale de la modélisation 46
III.2.2.1. Collection et préparation de la base des données 48
III.2.2.2. Méthode de génération des descripteurs moléculaires 51
III.2.2.3. Procédure de prétraitement et sélection des descripteurs 53
III.2.2.4. Choix des paramètres pour le développent des modèles de type RQSP-RNA 54
III.2.2.5. Analyses statistiques et calcules des erreurs des meilleurs modèles développées 57
III.2.2.6. Analyse de sensibilité des paramètres d'entrées pour chaque modèle développé 58
III.3. Résultats et discussions 60
III.3.1. Évolution du nombre de descripteurs au cours de la procédure de sélection 60
III.3.2. Architectures des meilleurs modèles neuronales obtenus 61
III.3.3. Application des modèles obtenus dans la prédiction des coefficients de diffusion des polluants dans différents systèmes 63
III.3.4. Estimation de la performance des résultats d'estimation issus des modèles 65

développés
III.3.4.1. Comparaison graphique entre les valeurs des coefficients de diffusion expérimentales et calculées 65
III.3.4.2 Calculs des paramètres de performance et analyse statistique des modèles obtenus 70
III.3.4.3 Analyse de sensibilité des paramètres d'entrées des cinq modèles développés par la méthode de poids 72
III.4. Conclusion 75

Conclusion générale et perspectives 77

Références biobibliographiques 79

Abréviations

ANN	**A**rtificial **N**eural **N**etwork
BFGS	**B**royden, **F**letcher, **G**oldfarb et **S**hanno
CoMFA	**Co**mparative **M**olecular **F**ield **A**nalysis
CoMSIA	**Co**mparative **M**olecular **S**imilarity **I**ndices **A**nalysis
COV	**C**omposés **O**rganiques **V**olatile
DSR	**D**éviation **S**tandard **R**elative
EPCA	**É**tablissement **P**ublic à **C**aractère **A**dministratif
ERPURS	**E**valuation des **R**isques de la **P**ollution **U**rbaine sur la **S**anté
ESP	**E**rreur **S**tandard de **P**rédiction
FBR	**F**onction de **B**ase **R**adiale
GES	**G**az à **E**ffet de **S**erre
HAP	**H**ydrocarbures **A**romatiques **P**olycycliques
IR	**I**mportance **R**elative
MP	**P**articulate **M**atter
NAAQS	**N**ational **A**mbient **A**ir **Q**uality **S**tandards
ORS	**O**bservatoire **R**égional de **S**anté
PCB	**P**olychlorobiphényles
PM	**P**articules **P**oussières
PMC	**P**erceptron **M**ulti**C**ouche
PRQA	**P**lans **R**égionaux pour la **Q**ualité de l'**A**ir
QSAR	**Q**uantitative **S**tructure **A**ctivity **R**elationships
QSPR	**Q**uantitative **S**tructure **P**roperty **R**elationships
RBFN	**R**adial **B**asis **F**unction
REQM	**R**acine d'**E**rreur **Q**uadratique **M**oyenne
RLM	**R**égression **L**inéaires **M**ultiple
RNAs	**R**éseaux de **N**eurones **A**rtificiel**s**
RQSA	**R**elations **Q**uantitatives **S**tructure **A**ctivités
RQSP	**R**elations **Q**uantitatives **S**tructure **P**ropriétés
RSD	**R**elative **S**tandard **D**eviation
RSM	**R**esponse **S**urface **M**odel
SDC	**S**omme **D**es **C**arrés
SMILES	**S**implified **M**olecular **I**nput **L**ine **E**ntry **S**pecification
SOS	**S**um **O**f **S**quares
VCM	**V**aleur de **C**ontribution **M**oyen

Nomenclatures et symboles

Lecture	Désignation	Unité
a_i	Potentiel du neurone	(-)
w_i	Poids	(-)
A	Surface perpendiculaire au flux thermique	(m^2)
Alg	Algorithme	(-)
b	Bais	(-)
D	Dimension	(-)
D_{AB}	Coefficient de diffusion	(cm^2/s)
F ou f	Fonction : Objectif, Object, Coût ou Fitness	(-)
M	Poids moléculaire	(Kg/Kmol)
N	Nombre des composés dans l'ensemble des données concerné	(-)
R	Coefficient de corrélation ou de régression	(-)
R^2	Coefficient de détermination	(-)
S	Sortie de chaque modèle	(-)
t	Temps	(S)
T ou T°	Température	(°C)
X	Entrée de réseau (les variables)	(-)
Y	Sortie de réseau	(-)
Z_j	Sortie de la couche cachée	(-)

Lettres Grecques

Lecture	Désignation	Unité
α	La pente de la droite de régression	(-)
β	L'interception de la droite de régression	(-)
λ	Conductivité thermique du matériau	($W.m^{-1}.K^{-1}$)

Exposants

Lecture	Désignation
App	Apprentissage
Cal	Calculée
Diff	Différente
Ent	Entière
exp	Expérimentale
Test	Test
Val	Validation

Indices

Lecture	Désignation
i	Elément ***i*** d'un vecteur
ij	Croisement de la ligne ***i*** avec la colonne ***j*** d'une matrice
j	Elément ou dimension ***j*** d'un vecteur
max	Maximum
min	Maximum
moy	Moyen

Liste des tableaux

Chapitre I. Généralités sur la pollution atmosphérique et la pollution des rejets industriels		
Tableau I.01	Type, nature et sources de la pollution des eaux	06
Tableau I.02	Origine des polluants et l'impact sur la santé et l'environnement	09
Tableau I.03	Cause de la pollution, milieu pollué et nature de polluant	12
Tableau I.04	Exemples des maladies reliées à des polluants	13

Chapitre III Modélisation d'un paramètre pertinent des phénomènes de transport lié à la pollution atmosphérique et/ou la pollution des rejets industriels		
Tableau III.01	Description quantitatives et qualitatives de la première base des données	48
Tableau III.02	Partie d'une 20 molécule de la première base des données utilisées dans ce travail	48
Tableau III.03	Description quantitatives et qualitatives de la deuxième base des données	50
Tableau III.04	Partie d'une 20 molécule de la deuxième base des données utilisées dans ce travail	50
Tableau III.05	Caractéristiques des descripteurs générées par l'ancien version de logiciel en ligne "PaDEL Descriptor"	53
Tableau III.06	Évolution du nombre de descripteurs après chaque étape de prétraitement pour les quatre $\mathbf{D_{AB}}$ étudiés de la première contribution.	60
Tableau III.07	Caractéristique des descripteurs retenus pour $\mathbf{D_{AB}}$ des quatre modèles développés dans la première contribution.	60
Tableau III.08	Architectures des meilleurs modèles obtenus pour les deux contributions	62
Tableau III.09	Fonctions d'activation apparues dans cette étude	63
Tableau III.10	Paramètres statistiques des meilleurs modèles choisir pour les cinq coefficients de diffusion $\mathbf{D_{AB}}$ étudiées.	71

Liste des figures

Chapitre I. Généralités sur la pollution atmosphérique et la pollution des rejets industriels		
Figure I.01	Classification des trois grandes classes de pollution causées par les activités humaines et quelques grandes catégories de ces divers polluants	05
Figure I.02	Diverses formes de contamination des sols et leurs conséquences	07

Chapitre II. Techniques de modélisation neuronal et méthodes de RQSP/RQSA		
Figure II.01	Représentation d'un neurone réel	27
Figure II.02	Modèle de base d'un neurone formel	28
Figure II.03	Structure et comportement d'un perceptron mono couche	31
Figure II.04	Modèle du perceptron multi couches	32

Chapitre III. Modélisation d'un paramètre pertinent des phénomènes de transport lié à la pollution atmosphérique et/ou la pollution des rejets industriels		
Figure III.01	Capture d'écran des logiciels utilisés dans cette étude : ChemDraw3D, MATLAB, STATISTICA, Microsoft Office Excel, et PaDEL Descriptor	46
Figure III.02	Organigramme générale de la méthodologie de développement suivez dans ce travail	47
Figure III.03	Capture d'écran du fichier texte du molécule chlorobenzène	52
Figure III.04	Représentation graphique de division de la première base des données	55
Figure III.05	Représentation graphique de division de la deuxième base des données	56
Figure III.06	Capture d'écran pour le choix des fonctions d'activation sur STATISTICA	56
Figure III.07	Organigramme illustrant l'analyse de sensibilité par la méthode de la partition des poids (méthode de poids)	59
Figure III.08	Représentation schématique d'un réseau du modèle de type RQSP-RNA	64
Figure III.09	Comparaison entre les valeurs expérimentales et calculées par le modèle neuronale de coefficient de diffusion $D_{AB_Couples}$ à Différents T° pour la base de données entière	66

Figure III.10	Comparaison entre les valeurs expérimentales et calculées par le modèle RQSP-RNA de coefficient de diffusion D_{AB_Air} à T=25°C pour la base de données entière	67
Figure III.11	Comparaison entre les valeurs expérimentales et calculées par le modèle RQSP- RNA de coefficient de diffusion D_{AB_Eau} à T=25°C pour la base de données entière	68
Figure III.12	Comparaison entre les valeurs expérimentales et calculées par le modèle RQSP- RNA de coefficient de diffusion D_{AB_Air} à T=100°C pour la base de données entière	69
Figure III.13	Comparaison entre les valeurs expérimentales et calculées par le modèle RQSP- RNA de coefficient de diffusion D_{AB_Eau} à T=100°C pour la base de données entière	70
Figure III.14	Contributions relatives des sept (7) entrées pour le modèle neuronal de coefficient de diffusion $D_{AB_Couples}$ à Différents T°	73
Figure III.15	Contributions relatives des vingt (20) descripteurs pertinents d'entrées pour le modèle de type RQSP-RNA de D_{AB_Air} à T=25°C	73
Figure III.16	Contributions relatives des onze (11) descripteurs pertinents d'entrées pour le modèle de type RQSP-RNA de D_{AB_Eau} à T=25°C	74
Figure III.17	Contributions relatives des seize (16) descripteurs pertinents d'entrées pour le modèle de type RQSP-RNA de D_{AB_Air} à T=100°C	74
Figure III.18	Contributions relatives des neuf (09) descripteurs pertinents d'entrées pour le modèle de type RQSP-RNA de D_{AB_Eau} à T=100°C	75

Introduction Générale

L'environnement est compris comme l'ensemble des composants naturels de la planète Terre, comme l'air, l'eau, l'atmosphère, les roches, les végétaux, les animaux, et l'ensemble des phénomènes et interactions qui s'y déploient, c'est-à-dire tout ce qui entoure l'Homme et ses activités.

La pollution est la dégradation d'un écosystème ou de la biosphère par l'introduction généralement humaine, l'entités physiques, chimiques ou biologiques, ou bien par les radiations altérant le fonctionnement de cet écosystème. La pollution a des effets importants sur la santé et la biosphère, comme en témoigne l'exposition aux polluants et le réchauffement climatique, en entraînant l'apparition de maladies inconnues jusqu'alors dans certaines zones géographiques.

La pollution atmosphérique est une altération de la qualité de l'air pouvant être caractérisée par des mesures de polluants chimiques, biologiques ou physiques. Elle peut avoir des conséquences préjudiciables à la santé humaine, aux êtres vivants, au climat, ou bien aux matériels. La pollution atmosphérique est responsable ou aggraver de nombreuses maladies liées au système respiratoire dont notamment :

- L'asthme
- Les bronchites.
- Les allergies aux pollens

La pollution industrielle causée directement par l'industrie lorsqu'elle pénètre de manière biologique, chimique ou organique. Ce type de pollution correspond à une contamination plus ou moins, aussi suivant les compartiments des écosystèmes qui sont principalement l'air, l'eau, le sol, le réseau trophique ou l'être humain.

Les phénomènes de transferts concernent tous les déplacements dans l'espace de la matière, de l'énergie et de la quantité de mouvement qui nous permette de contrôler le déplacement des molécules quelque sois dans l'air ou bien dans l'eau, ce dernier nous permette de connaitre les concentrations exactes de ces molécules et évité leurs augmentations qui doit être pollué notre environnement. La connaissance des propriétés physiques, chimiques et thermodynamiques des polluants joue un rôle important dans la conception et le fonctionnement de presque tous les équipements de l'industrie de l'environnement. Cependant, le calcul des propriétés du polluants dépend des propriétés de ses derniers. Certaines propriétés physiques peuvent être facilement mesurées, par exemple la température, d'autres caractéristiques telles que les propriétés critiques et l'enthalpie d'évaporation ne peuvent pas être facilement mesurées **(Constantinou & Gani 1994 ; Riazi 2005)**.

La modélisation moléculaire est une technique qui permet non seulement de représenter les structures en deux ou en trois dimensions, mais aussi d'aider tout chimiste théorique, de mener des études approfondies et précises des systèmes moléculaires, et de représenter explicitement des modèles sous une forme mathématique. La modélisation moléculaire actuelle est largement utilisée pour développer des modèles fiables pour prédire les propriétés et les activités des phénomènes physique, chimiques ou biologiques **(Jensen 2007)**. L'une de ces techniques est la modélisation des Relations Quantitative Structure Propriété/Activités (RQSP/RQSA) ; en anglais : *Quantitative Structure Property/Activity Relationships* (QSPR/QSAR) **(Charton 2002)**, ce qui permet de prédire les propriétés et/ou les activités des systèmes chimiques et/ou biologiques en fonction de leurs structures moléculaires. Les relations entre les structures moléculaires, leurs propriétés et/ou leurs activités sont généralement établies à l'aide des méthodes de la modélisation d'apprentissage statistiques, telles que les Réseaux de Neurone Artificiels (RNAs) ; en anglais : *Artificial Neural Networks* (ANN) **(Roubehie fissa *et al*. 2019)**.

L'objectif principal de ce travail est de développer des modèles mathématiques de type RQSP-RNA afin de prédire le coefficient de diffusion qui est considérer comme un paramètre très attachée au phénomène de la pollution, dans cinq système de diffusion des polluants différents : **(1)** D_{AB} de différents couple des polluants, **(2)** D_{AB} de différents polluants dans l'Air à T=25°C, **(3)** D_{AB} de différents polluants dans l'Eau à T=25°C, **(4)** D_{AB} de différents polluants dans l'Air à T=100°C et en fin **(5)** D_{AB} de différents polluants dans l'Eau à T=100°C, utilisant le Réseau de Neurone Artificiel-Perceptron Multi Couches (RNA-PMC) en fonction de leurs structures chimiques.

Le présent travail s'articule principalement autour de trois (03) chapitres et une conclusion générale dans lesquelles :

- **Le premier chapitre :** Comprend une étude bibliographique sur la pollution atmosphérique et industrielle, nous avons parlé sur des généralités et des notions de bases lier à ce sujet, tel que la définition, la source et leurs effets sur la santé humaine et l'environnement, les caractéristiques des polluants, la source et lutte contre ces polluants avec des méthodes classique.
- **Le second chapitre :** Est dédier à l'étude de la modélisation, en particulier le concept des Relations Quantitative Structure Propriété/Activités (RQSP/RQSA), et aux techniques des modélisations basée sur l'apprentissage statistiques, telles

que les Réseaux de Neurone Artificielles (RNAs). Nous avons donné un aperçu sur l'approche générale de la modélisation, de ses différents types possibles, des éléments nécessaires au développement de ces modèles, du concept de descripteur moléculaire, de la méthode de calcul et de sélection, et enfin les méthodes de validation. Nous nous appuierons particulièrement sur les outils de développement des modèles de type RQSP-RNA.

- **Le dernier chapitre :** Sera consacré à la modélisation de type Relations Quantitative Structure Propriété/Activités (RQSP/RQSA) par le Réseau de Neurone Artificiel- Perceptron Multi Couches (RNA-PMC) afin de développer des modèles de type RQSP- RNA pour prédire le coefficient de diffusion , dans cinq système de diffusion des polluants différents : **(1)** D_{AB} de différents couple des polluants, **(2)** D_{AB} de différents polluants dans l'Air à T=25°C, **(3)** D_{AB} de différents polluants dans l'Eau à T=25°C, **(4)** D_{AB} de différents polluants dans l'Air à T=100°C et en fin **(5)** D_{AB} de différents polluants dans l'Eau à T=100°C. La méthodologie et les matériels de développement ont été présentés afin d'obtenir des modèles qui ont été analysés par différents outils statistiques, qui ont ensuite fait l'objet de discussions sur les principales parties de ces résultats.

Nous terminerons ce travail par une conclusion générale qui rassemble les principaux résultats obtenus et donne quelques perspectives qui sera un guide de référence pour les prochains travails effectue dans cet axe de recherche.

Chapitre I.
Généralités sur la pollution atmosphérique et la pollution des rejets industriels

I.1. Introduction

I.1.1. Importance de l'environnement

L'environnement, c'est l'ensemble des éléments (plantes, animaux, forêts, étendues d'eau) qui entourent les êtres humains. Il est essentiel à la survie, car les gens ont besoin de respirer, de boire et de se nourrir. Heureusement, la planète peut répondre à ces besoins essentiels.

Cependant, contrairement aux animaux, les êtres humains ont tendance à transformer leur environnement. Ils utilisent ce qui les entoure pour améliorer leurs conditions de vie. Par exemple, ils utilisent les cours d'eau pour laver leurs vêtements ou pour se débarrasser de leurs déchets. S'ils exagèrent, ils ne pourront plus boire l'eau sans risquer de tomber malades. De plus, les poissons qui y vivent et les animaux qui la boivent peuvent mourir, ce qui peut avoir des effets néfastes sur la nourriture des êtres humains. Ainsi, pour survivre, les êtres humains doivent faire attention à leur environnement. Malheureusement, ce n'est pas toujours le cas **(Landry *et al*. 2013)**.

I.1.2. Echauffement climatique

Les changements climatiques sont causés par les modifications de l'atmosphère qui résultent de sa transformation chimique par les gaz à effet de serre (GES). Cette perturbation de l'équilibre atmosphérique s'exprime par une augmentation des températures moyennes sur terre, modification de ses caractéristiques physiques, chimiques et biologiques.

Les impacts sur l'environnement sont multiples, importants et de plus en plus fréquents : sécheresses, fonte des glaciers et de la glace de mer, élévation du niveau des océans, tempêtes tropicales. Ils affectent l'ensemble de la population mondiale et la biodiversité planétaire. Les activités humaines sont les responsables principaux des changements climatiques actuels et de leurs impacts sur l'environnement **(Gingras & C 2011)**.

I.1.3. Causes de pollution

- La production d'énergie, cause principale de pollution.
- La diversification des polluants chimiques et l'accumulation des déchets.
- Pollutions dues à l'agriculture et l'élevage intensif.

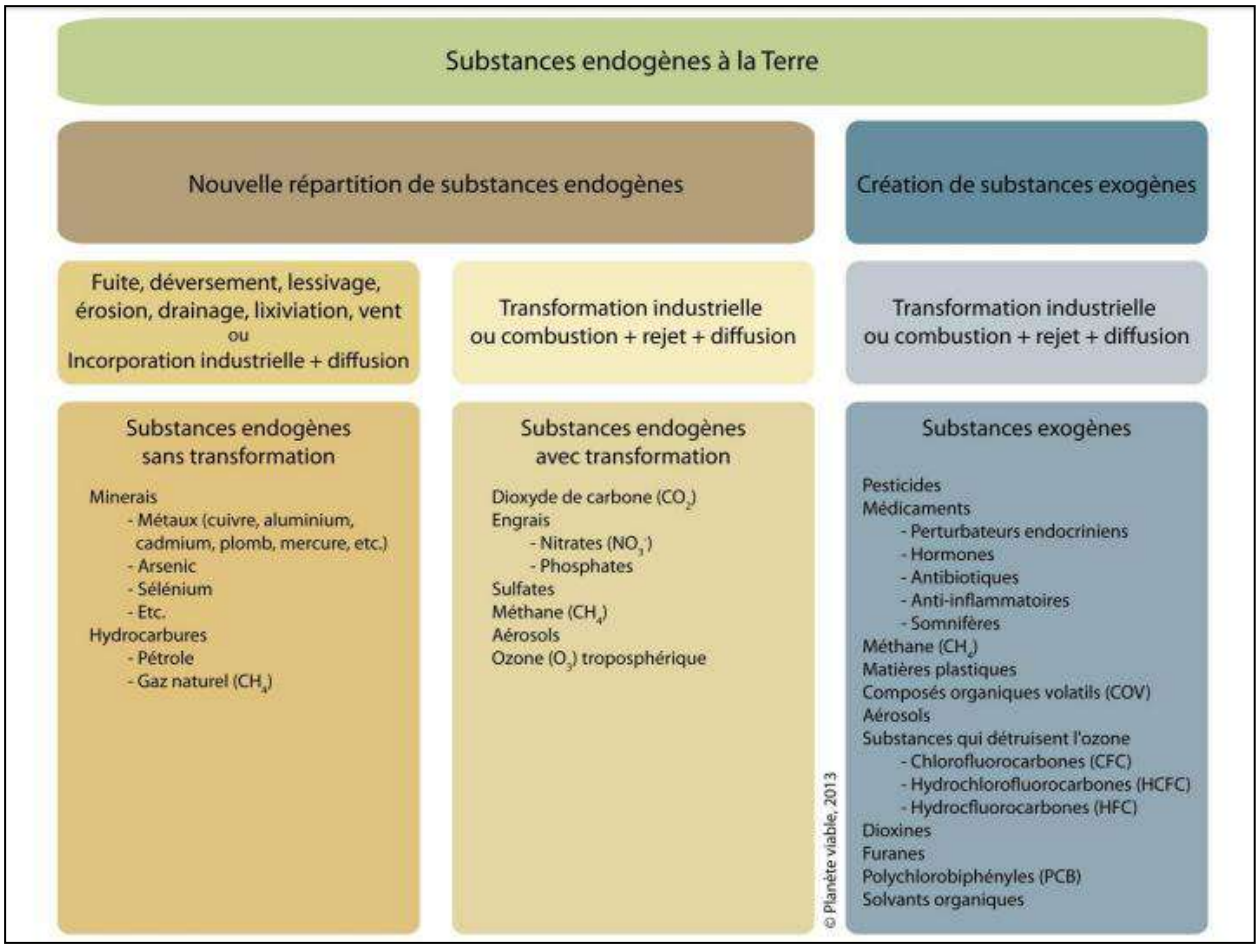


Figure I.01. Classification des trois grandes classes de pollution causées par les activités humaines et quelques grandes catégories de ces divers polluants.

I.1.4. Grands types de pollution

I.1.4.1. Pollution des sols

La pollution du sol peut être diffuse ou locale, d'origine industrielle, agricole (suite à l'utilisation massive d'engrais ou de pesticides qui s'infiltrent dans les sols). Ces pollutions agricoles peuvent avoir plusieurs impacts **(Wassim 2016)**.

I.1.4.2. Pollution de l'air ou atmosphérique

La pollution atmosphérique peut être définie par la présence de polluants (gazeux ou particules) dans l'atmosphère, pouvant provoquer des effets nocifs sur l'environnement et la santé **(Arqués 1998)**.

I.1.4.3. Pollution de l'eau

La pollution de l'eau décrit généralement l'introduction ou la présence des substances nocives ou inacceptables dans l'ampleur suffisante pour modifier les indices de qualité de l'eau

naturelle **(Benson 2008)** ; La pollution de l'eau douce (par exemple par le biais de l'eutrophisation, l'acidification, et la pollution des eaux souterraines) est celle qui diminue sa pureté **(Park & Allaby 2007)**.

I.1.5. Structure de ce chapitre

Dans ce chapitre nous allons présenter des généralités sur la pollution atmosphérique et la pollution des rejets industriel, leurs causes, et donner des solutions pour protéger notre environnement.

I.2. Classifications des Pollution

La pollution est la modification physico-chimique ou biologique d'un écosystème par l'introduction d'un élément extérieur qui créait un danger pour l'être vivant. Il existe trois types de pollution : pollution de l'eau, du sol et pollution de l'aire.

I.2.1. Pollution de l'eau

L'eau, partie intégrante et importante de l'environnement, est exposée à différentes sources de pollution, notamment de nature domestique, agricole et industrielle. Il convient d'indiquer à ce propos que l'activité hospitalière génère des rejets variés. Pollution qui peut mettre en cause la vie humaine en rendant l'accès à cette denrée impossible pour les générations futures **(Bjijou 2016)**. Le **tableau I.01**, résume les différents polluants des eaux et leurs sources.

Tableau I.01. Type, nature et sources de la pollution des eaux **(Moilleron *et al*. 2017)**.

Type de pollution	Nature	Source
Physiques : ***Pollution thermique*** ***Pollution radioactive***	Rejets d'eau chaude Radio-isotopes	Centrale thermique Installation nucléaire
Matière organique	Glucides, lipides, protides Ammoniac, nitrates	Effluents domestiques, agricoles... Elevages et pisciculture
Chimiques : ***Fertilisants*** ***Métaux et métalloïdes*** ***Pesticides*** ***Organochlorés*** ***Composés org. De synthèse*** ***Détersifs*** ***Hydrocarbures***	Nitrates, phosphates Hg, Cd, Pb, Al, As... Insecticides, herbicides... PCB, solvants Nombreuses molécules Agents tensio-actifs Pétrole et dérivés	Agriculture, lessives Industries, agriculture, pluies acides Agriculture, industries Industries Industries Effluents domestiques Industrie pétrolière, transports
Microbiologique	Bactéries, virus, champignons	Effluents urbains et d'élevage

I.2.2. Pollution du sol

Jusqu'à récemment les problèmes de dégradation des sols ont été perçus essentiellement d'un point de vue quantitatif qu'il s'agisse des relations agriculture-forêt, agriculture-urbanisation, ou désertification. Mais le développe- ment économique, à travers l'industrialisation et ses déchets, et aussi du fait de l'intensification de l'agriculture, pose un problème de nature qualitative. L'ampleur croissante des dégradations physiques, chimiques et biologiques remet en cause, au moins ponctuellement, le rôle de filtre, de support de vie et aussi d'outil économique que remplit le sol. La première partie présente un schéma général d'analyse économique des sols intégrant les diverses sources de pollution et leurs impacts tant sur les activités, que sur la santé humaine ou les écosystèmes. Pour chacune de ces conséquences les problèmes d'évaluation sont évoqués. La deuxième partie est centrée sur l'agriculture et décrit les mécanismes généraux à l'origine des pollutions. L'accent est mis sur l'intensification avec une analyse en termes d'intrants polluants **(Rainelli 1996)**. La **figure I.02**, résume les principales formes et conséquences de la pollution des sols.

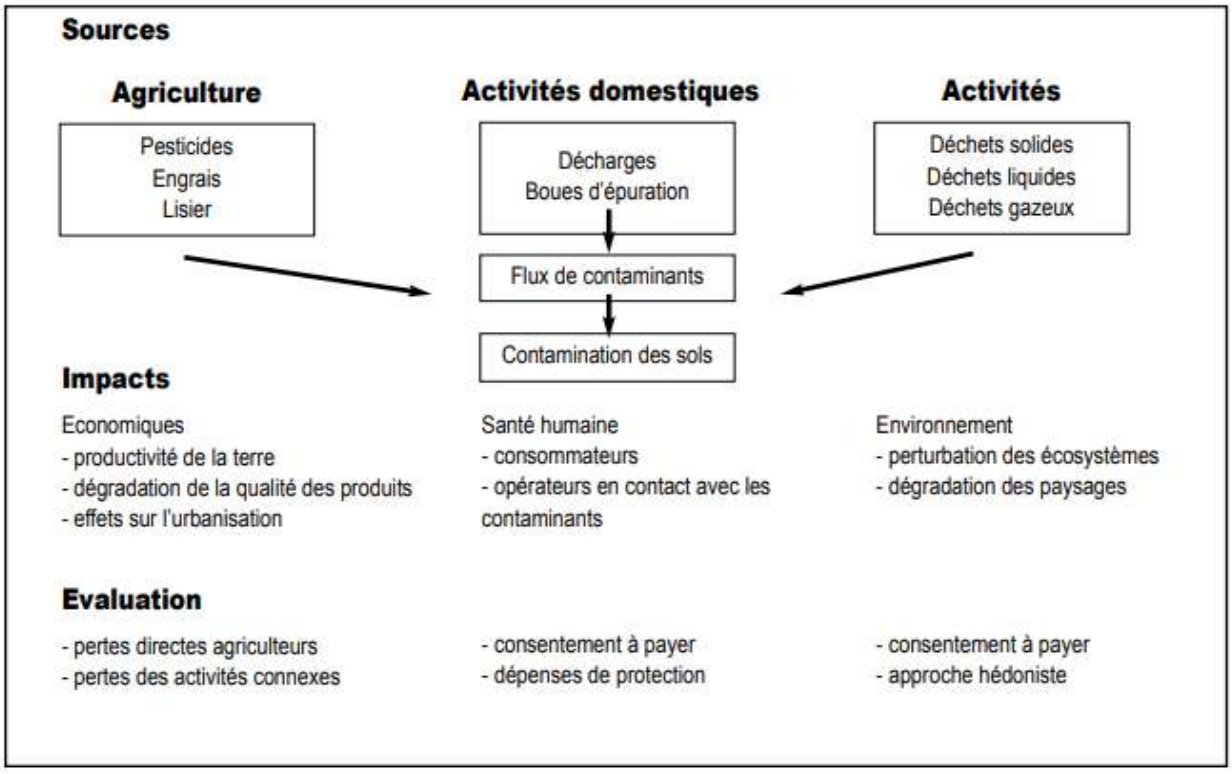


Figure I.02. Diverses formes de contamination des sols et leurs conséquences **(Rainelli 1996)**.

I.2.3. Pollution de l'air

En Algérie, les pouvoirs publics ont définit la pollution atmosphérique, à travers l'article 44 de la Loi N° 03-10 du 19 juillet 2003 relative à la protection de l'environnement dans

le cadre du développement durable, comme suit : « Constitue une pollution atmosphérique au sens de la présente loi, l'introduction, directement ou indirectement, dans l'atmosphère et les espaces clos, de substances de nature à mettre en danger la santé humaine, influer sur les changements climatiques ou appauvrir la couche d'ozone, nuire aux ressources biologiques et aux écosystèmes, compromettre la sécurité publique, incommoder la population, provoquer des nuisances olfactives, nuire à la production agricole et aux produits agro-alimentaires, altérer les constructions et porter atteinte au caractère des sites, détériorer les biens matériels ». Les niveaux de polluants dans l'air dépendent de leur nature et des conditions de rejet de ceux-ci dans l'atmosphère en occurrence de l'intensité de la source et de la météorologie. Ses sources peuvent être soit naturelles tels que feu de forêt et éruption volcanique, soit anthropiques, c'est-à-dire liées à l'activité humaine. **(Mamoudou & Téneman 2018)**.

I.3. Pollution atmosphérique

I.3.1. Définition

La pollution atmosphérique peut être définie par la présence de polluants (gazeux où Particules) dans l'atmosphère, pouvant provoquer des effets nocifs sur l'environnement et la Santé **(Arqués 1998)**. Les sources de cette pollution peuvent être soit naturelles (feu de forêt, Éruption volcanique,), soit anthropiques, c'est-à-dire liées à l'activité humaine **(Mayer 1999)**. Dans ce dernier cas, la pollution est souvent le résultat direct des progrès industriels de ces derniers siècles, comme par exemple l'émission continue et parfois sans précaution des polluants associés aux processus de combustion (véhicules automobiles, installations industrielles, production d'énergie par combustion…) **(Al Barakeh 2012)**.

I.3.2. Source de pollution atmosphérique

Les polluants proviennent majoritairement des activités humaines : des transports, chauffage des bâtiments, l'agriculture, industries et de la production d'énergie et du brûlage à l'air libre des déchets (déchets verts, du BTP…).

Ces activités émettent également des gaz à effet de serre : gaz carbonique, méthane, protoxyde d'azote, gaz fluorés…Les émissions de gaz à effet de serre ne cessent d'augmenter depuis 150 ans. Leur concentration dans l'atmosphère entraîne un réchauffement de la planète, ce qui modifie le climat. Réduire nos émissions de gaz à effet de serre est essentiel tout comme lutter contre la pollution de l'air. En modifiant nos comportements, nous pouvons répondre à ces deux majeurs à la fois **(Charpin *et al*. 2016)**.

I.3.3. Effet nocive de ces éléments sur l'environnement et les êtres vivants

Tableau I.02. Origine des polluants et l'impact sur la santé et l'environnement (**ASQA 2010**).

Polluants	Origine	Impact sur l'environnement	Impacte sur la santé
Oxyde d'Azote (NO_x)	Toutes combustions à haute température de combustible fossiles (charbon, fioul, essence…) le monoxyde d'azote (NO) rejeté par les pots d'échappement s'oxyde dans l'aire et se transforme en dioxyde d'azote (NO_2) qui est 90% un polluant <secondaire>.	• Rôle de précurseur dans la formation d'ozone dans la basse atmosphérique. • Contribuent aux pluies acides qui affectent les végétaux et les sols. • Contribuent à la concentration de nitrates dans les sols.	• NO_2 : gaz irritant pour les bronches (augmente la fréquence et la gravité des crises chez les asthmatiques et favorise les infections pulmonaires infantiles), • NO non toxique pour l'homme aux concentrations environnementales.
Hydrocarbures aromatiques polycycliques (HAP) et composés organiques volatile (COV)	Combustions, utilisation de solvants (peintures, colles) et de dégraissants, produits de nettoyage, remplissage de réservoirs, automobiles, de citernes…	• Précurseurs dans la formation de l'ozone, • Précurseurs d'autres sous-produits à caractère oxydant (PAN, acides nitrique, aldéhydes…).	• Effets divers selon es polluants dont irritation et diminution de la capacité respiratoire, • Considéré pour certains comme cancérogènes pour l'homme (benzène, benzo-(a)pyrène), • Nuisances olfactives fréquentes.
Ozones (O_3)	Polluant secondaire, produit dans l'atmosphère sous l'effet du rayonnement solaire par des réactions complexes entre certaine polluants primaire (NOx, CO et COV) et principal indicateur de l'indicateur de l'intensité de la	• Perturbe la photosynthèse et conduit à une baisse de rendement des cultures (5 à 10% pour bé en Ile-de-France, selon I'INRA), • Nécroses sur la feuille et les aiguilles d'arbres forestiers,	• Gaz irritant pour l'appareil respiratoire et les yeux, • Associé à une augmentation de la mortalité au moment des épisodes de pollution (études ERPURS/ORS Ile-de-France).

	pollution photochimique.	• Oxydation de matériaux (caoutchoucs, textiles, …), • Contribue à l'effet de serre,	
Particules poussières en suspension (PM)	Combustions industrielles ou domestique, transport routier diesel, origine naturelle (volcanisme, érosion...). <u>Classement en fonction de leur taille :</u> • PM10 : particules de diamètre inférieur à 10 µm (Retenue au niveau du nez et des voies aériennes supérieures) • PM2.5 : particules de diamètre inférieur à 2.5µm (Pénètrent profondément dans l'appareil respiratoire jusqu'aux alvéoles pulmonaires)	• Contribuent aux salissures des bâtiments et des monuments : • Cout du ravalement des bâtiments publics d'Ile-de-France 1.5 à 7 milliards de francs par an (source PRQA Ile-de France) • Cout du nettoyage du Louvre en 1995 : de l'ordre de 30 millions de francs (source PRQA Ile-de-France).	• Irritation et altération de la fonction respiratoire chez les personnes sensibles, • Peuvent être combinées à des substances toxiques voire cancérigènes comme les métaux lourds et des hydrocarbures. • Associées à une augmentation de la mortalité pour causes respiratoires ou cardiovasculaires (ERPURS/ORS Ile-de-France).
Dioxyde de soufre (SO_2)	Combustions de combustibles fossiles (fioul, charbon, lignite, gazole…) contenant du soufre. La nature émet aussi des produits soufrés (volcans).	• Contribue aux pluies acides qui affectent les végétaux et les sols, • Dégrade la pierre (cristaux de gypse et croutes noires de micro particules cimentées).	• Irritation des muqueuses de la peau des voies respiratoires supérieures (toux, gêne respiratoire, troubles asthmatiques).

I.3.4. Solutions pour éviter ou réduire la pollution atmosphérique

Les mesures de réduction de la pollution atmosphérique peuvent être directes, agissant à la source grâce à des dispositifs techniques, ou, indirectes à l'aide d'actions visant à modifier le comportement des acteurs. Les dispositifs techniques aspirent à améliorer les combustibles utilisés ainsi que les installations de combustion. Dans ce cadre, le suivi et le contrôle des émissions s'avèrent essentiels pour toute stratégie de réduction des rejets gazeux polluants. La réduction de la pollution atmosphérique repose également sur la réglementation à travers des mesures fiscales, des instruments économiques et des mesures incitatives. Les réglementations concernent aussi bien les sources fixes, notamment les installations industrielles, que les sources mobiles comme les transports. Elles visent en général, à limiter les flux d'émission et sont élaborées dans le cadre d'accords internationaux, notamment à travers le protocole de Kyoto, ratifié par 156 pays dont l'Algérie par le décret présidentiel N° 04-144 du 28 avril 2004.

Généralement, les autorités en charge de la gestion des risques, sur la base des législations nationales ou supranationales, fixent des seuils d'information et d'alerte pour protéger la santé et la sécurité des biens et personnes. Seul le second seuil doit entraîner des mesures de réduction de la pollution. Ces mesures peuvent concerner, avec diverses graduations possibles, en fonction de la gravité de l'épisode de pollution concerné, des diminutions ou arrêt des émissions des gros émetteurs industriels responsables de la pollution en cours par le ralentissement de la production, le changement du combustible par autre moins polluant ou bien carrément l'arrêt complet de l'activité industrielle. Ces mesures peuvent concerner également les transports automobiles par des ralentissements de la vitesse maximale autorisée des axes déterminés de circulation, ou par une circulation alternée, en fonction du numéro de la plaque minéralogique du véhicule qu'il soit pair ou impair **(Rahal 2015)**.

I.4. Pollution des rejets industriels

I.4.1. Définition

Les effluents industriels regroupent différents déchets provenant des activités socioéconomiques diverses (industrie agro-alimentaire, tannerie, textile, papeterie, industrie chimique, industrie pétrochimie, décharges publiques,). Ce type de pollution a des effets néfastes sur les organismes vivants, sur le pouvoir d'autoépuration de l'eau et conduit à l'accumulation de certains éléments forts nuisibles dans la chaîne alimentaire (métaux lourds, radioactivité,) **(Bjijou 2016)**.

I.4.2. Source des rejets industriels

On peut distinguer deux principales causes de contamination dans la civilisation industrielle : La production de l'énergie ; les activités de l'industrie chimique ; les activités agricoles. Pour chacune de ces causes fondamentales de pollution existent des sources situées en amont, au niveau de la fabrication et en aval à celui de l'utilisation par le consommateur.

I.4.2.1. Production d'énergie et source de pollution

La recherche de ressources énergétiques et leur utilisation par les pays industrialisés, implique le gaspillage de ressources naturelles à la fois peu abondantes et non renouvelables comme les combustibles fossiles, et joue un rôle prépondérant dans la contamination de l'environnement par d'innombrables substances toxiques.

Tableau I.03. Cause de la pollution, milieu pollué et nature de polluant.

Activité	Cause de la pollution	Milieu pollué	Nature de polluants
Extraction	Fuite de puis < off-shore >	Océan	Pétrole brute
Raffinage	Rejet d'effluents gazeux et liquides	Atmosphère, eaux continentales, mers	Divers composés organiques, mercaptans, SO_2
Utilisation	Combustions incomplètes	Atmosphère	SO_2, oxydes d'azote, hydrocarbure

I.4.2.2. Industrie chimique moderne source de polluants variés

I.4.2.2.1. Agents polluants

On ne saurait dresser ici une liste exhaustive des innombrables composés organiques, rarement inoffensifs, rejetés tant en amont qu'en aval de l'activité industrielle moderne. Aldéhydes, phénols, fluorures, amines diverses, solvants chlorés, pesticides, détersifs, etc., sont dispersés dans le milieu naturel et se retrouvent soit dans l'air, soit dans les eaux et contribuent chacun pour leur part à la contamination des divers écosystèmes...

La dispersion dans l'environnement de matières plastiques variées (polyéthylène, chlorure de polyvinyle, polyuréthane, polystyrènes, etc.) est de nature préoccupante à l'heure actuelle. N'oublions pas que ces substances, outre des traces de monomères pas toujours inoffensives, renferment divers stabilisants, polymérisant et agents plastifiants dont la toxicité est fort mal évaluée. La combustion incomplète des matières plastiques, leur rejet dans les eaux

continentales et les océans, semblent jouer un rôle significatif dans la contamination de l'environnement par les Polychlorobiphényles (PCB) substances similaires au DDT11 et par le Cadmium, métal très toxique et carcinogène utilisé comme stabilisateur de certains de ces polymères synthétiques.

I.4.2.2.2. Dispersion planétaire de certains toxiques

On s'est de la sorte beaucoup moins inquiétée jusqu'à une date assez récente de la contamination de l'ensemble de l'écosphère, Océan mondial inclus, par de nombreux produits de la chimie de synthèse. Ainsi, on trouve à l'heure actuelle des fragments de matières plastiques entraînés par les courants et dérivant en plein océan Austral très au sud de la convergence antarctique en pleine mer de Ross, ce qui atteste de l'aspect global de la contamination de la biosphère par ces substances. L'ensemble de l'atmosphère et de l'hydrosphère est peu à peu empoisonné par des composés persistants et de toxicité pernicieuse tels les biphényles polychlorés (PCB) déjà nommés, les dioxines, les redoutables dibenzofurans. On trouve de la sorte des traces de ces composés organohalogénés dans l'organisme de Mammifères du Grand Nord canadien ou dans celui des Poissons pélagiques, et aussi d'animaux antarctiques. Il en est de même d'autres substances, comme les divers dérivés du Mercure, utilisés comme pesticides ou encore dans l'industrie, en particulier pour la catalyse, qui à l'image des composés précédents persistent dans l'ensemble des écosystèmes et contamine de façon insidieuse tous les réseaux trophiques.

I.4.2.2.3. Effets nocifs de ces éléments sur l'environnement et sur les êtres vivants

Quelques conséquences graves sont causées par les industries. En effet, les eaux usées et les déchets peuvent perturber le cycle de l'azote dans le sol et provoquent la montée de sel dans les sols. C'est comme cela que les sols sont stérilisés et que plus aucune végétation ne peut pousser à l'endroit contaminé.

De plus, le déversement de produits pétroliers dans la mer met en danger la vie maritime au risque de voir certaines espèces de la faune et la flore sous-marine disparaître. Des pertes de vies humaines ont même été enregistrées dues à la consommation de produits intoxiqués par les produits pétroliers.

Tableau I.04. Exemples des maladies reliées à des polluants **(Goverse 2013)**.

Impacte sur la santé	Association avec certaines expositions environnementales
Maladies infectieuses	• Contamination de l'eau et de aliments. • Changement lie au changement climatique dans le cycle de vie de agents pathogènes
Caner	• La pollution de l'air. • Certains pesticides. • Toxines naturelles (aflatoxines). • Hydrocarbures aromatiques polycycliques. • Certains métaux (arsenic, cadmium, chrome). • Benzène. • Dioxines.
Maladies cardiovasculaires	• La pollution de l'air. • Carbon monoxyde.
Maladies respiratoire	• Dioxyde de soufre. • Dioxyde d'azote.
Maladies de la peau	• Le rayonnement UV. • Certains métaux (nickel, pentachlorophénol).
Dysfonctionnements productifs	• Cadmium. • Phtalates et autres perturbateurs endocriniens produits pharmaceutiques.
Troubles du développements	• Mercure. • Cadmium. • Certains pesticides.
Maladies de système nerveux	• Méthyl mercure • Manganèse. • Certains solvants. • Organophosphates.

Au niveau de l'air, les gaz qui s'échappent des usines provoquent des trous dans la couche d'ozone. Les conséquences sont immédiates pour les êtres humains avec l'apparition de cancers de la peau. Le problème de la couche d'ozone provoque également le réchauffement climatique qui entraîne la fonte des glaciers au Pôle Nord. Des maladies respiratoires peuvent être de plus en plus fréquentes chez les humains à cause des carrières ou scieries **(BEM Recycling 2020)**.

I.4.2.3. Solution pour éviter ou réduire la pollution issue des rejets industriels

Malheureusement, aucune activité industrielle ne peut fonctionner sans polluer un minimum. Cependant, une réglementation beaucoup plus stricte sur le déversement des produits dans la mer, par exemple, pourrait sûrement sauver de nombreuses espèces maritimes. En ce qui concerne la pollution atmosphérique et des sols, seules les lois et les gouvernements peuvent appliquer quelques règles en accord avec les industriels.

Il faut que des actions soient prises pour la pollution industrielle mais également sur toutes les autres sortes de pollution. La prise de conscience de ce problème doit être générale.

Pour conclure, il serait irréel d'éradiquer toutes sortes de pollutions. Néanmoins, celles-ci peuvent être encadrées afin d'être ramenées à des mesures raisonnables. La protection de notre environnement est devenue un impératif pour le monde entier. Tous les pays doivent se sentir impliquer dans cette lutte **(BEM Recycling 2020)**.

I.5. Phénomènes de transport liés à la pollution atmosphérique et la pollution des rejets industriels

I.5.1. Définition

Un phénomène de transfert est un phénomène irréversible durant lequel une grandeur physique est transportée par le biais de molécules et qui a pour origine inhomogénéité d'une grandeur intensive. C'est la tendance spontanée des systèmes physiques et chimiques à rendre uniformes ces grandeurs qui provoquent le transport.

L'étude de chaque phénomène de transport se réfère à une certaine entité (caractéristique) transférée, par exemple : la quantité de mouvement nécessaire pour augmenter la vitesse d'un fluide, la chaleur afin de vaporiser un liquide et la masse du liquide transportée dans une conduite ou la dispersion d'un liquide coloré au sein d'un autre liquide transparent.

Les grandeurs physiques transférées les plus connues sont la chaleur (transfert thermique), la matière (transfert de masse) et la quantité de mouvement. Leurs modes de transfert et attacher avec la pollution atmosphérique et la pollution des rejets industriels **(Debacq 2019)**.

I.5.2. Le transfert de chaleur

On définit le transfert de chaleur comme de l'énergie thermique en transit à cause d'une différence de température. On peut observer les phénomènes de transfert de chaleur aussi bien

dans des situations industrielles (fours, réacteurs, échangeurs de chaleur, colonnes à distiller, chambres froides, presses à injection, coulée des métaux...) que dans notre vie quotidienne (le chauffage et l'isolation de la maison, la cuisson des aliments, les vêtements d'hiver et le facteur de refroidissement éolien, le coup de soleil sur la plage) (**Penot 1980**). Les trois modes de transfert de chaleur sont : la conduction, la convection et le rayonnement.

I.5.3. Transfert de matière

Pour lequel la grandeur transférée est la matière (Concentration massique), ce transfert s'effectue entre deux zones où règnent des concentrations massiques différentes, il se fait toujours de la concentration plus élevée vers la concentration la plus faible. La différence de concentration est appelée : la force motrice du transfert de masse **(Debacq 2019)**.

I.5.4. Transfert de quantité de mouvement

Pour lequel la grandeur transférée est la quantité de mouvement (Vitesse), ce transfert s'effectue entre deux entités qui possèdent des vitesses différentes, il se fait toujours de l'entité qui a la vitesse la plus élevée vers celle qui a la vitesse la plus faible. La différence de vitesse appelée : la force motrice du transfert de quantité de mouvement **(Debacq 2019)**.

I.6. Paramètres lier à la diffusion des polluants dans l'environnement

I.6.1. Paramètres de diffusion lies à la pollution

I.6.1.1. Coefficients de diffusion des liquides et des gaz polluants

La compréhension des phénomènes de diffusion dans les liquides est inférieure à la plupart des égards à celui de la diffusion à l'état gazeux ou à l'état solide. Le manque de compréhension précise de ceux-ci et d'autres phénomènes de transport associés dans les liquides est la conséquence d'une connaissance complète de l'état liquide. En effet, Parmi les paramètres qui influencent plus particulièrement le transport des polluants on pourra citer **(Lemière *et al*. 2001)** : la dispersive, la densité, la solubilité, la température, la viscosité et la masse volumique.

- ***Dispersion***

La dispersion est liée à l'hétérogénéité du milieu poreux à petite et grande échelle ; elle est à l'origine de ((l'étalement N d'un panache de pollution et contribue à diluer les concentrations De même qu'il existe un flux de convection, il existe aussi un flux de dispersion, qui est proportionnel au gradient de la concentration (loi de Fick). Il existe un coefficient de

dispersion longitudinale (suivant le sens de l'écoulement) et un coefficient de dispersion transversale. Les coefficients de dispersion sont proportionnels à la vitesse effective de l'eau **(Lemière *et al*. 2001)**.

- ***Densité du liquide***

Affecte généralement peu son infiltration dans la tranche de sol située au-dessus de la nappe. Par contre, elle contrôle fortement son comportement dans la zone saturée. Si le liquide est immiscible (cas des hydrocarbures, par exemple), le contaminant s'accumulera de préférence en surface de la nappe si sa densité est inférieure à 1, et au plancher de celle-ci si elle est supérieure à un.

II convient de tenir compte de la densité des polluants lors des investigations : les piézomètres utilisés doivent atteindre le plancher de l'aquifère si des polluants de densité > 1 sont suspectés **(Lemière *et al*. 2001)**.

- ***Densité du vapeur et des produits volatils***

Est exprimée par rapport à l'air. Si elle est supérieure à 1, la fraction volatile s'accumulera au toit de la nappe. Si elle est inférieure à 1, cette fraction tendra à se disperser vers l'atmosphère via la zone non saturée (voir O). En principe, on ne tient pas compte de la densité pour un polluant soluble dans l'eau (d= 1). Toutefois, il faut tenir compte de la densité des saumures (NaCI, $CaC1_2$), éventuellement chargées en polluants organiques eu ou métaux lourds, car leur densité est supérieure à l'eau et l'ensemble aura tendance à migrer vers le fond de la nappe **(Lemière *et al*. 2001)**.

- ***Solubilité***

On appelle solubilité, la quantité de substance solide, liquide ou gazeuse, capable de se dissoudre complètement (c'est-à-dire sans qu'une seconde phase apparaisse) dans un volume ou une masse donnée de solvant (ici : l'eau). Lorsque cette quantité est atteinte, l'eau est dite saturée en cette substance La solubilité varie avec la température de l'eau, son pH, et avec la composition chimique du milieu (autres espèces en solution dans l'eau). Les réactions gouvernant les échanges de matière entre solides et solutions aqueuses sont dites de dissolution-précipitation **(Lemière *et al.* 2001)**.

I.6.2. Transfert thermique

Différence de température entre deux milieux de transport d'énergie thermique (transfert thermique). On distingue trois modes de transfert thermique : la conduction, la convection et le rayonnement.

I.6.3. Coefficients de transfert de chaleur

Lorsque deux points dans l'espace sont à des températures différentes, il y a systématiquement transfert de chaleur toujours vers le corps froid. C'est une conséquence directe du deuxième principe de la thermodynamique.

I.6.3.1. Différents modes de transmission de la chaleur

On distingue trois modes différents de transmission de la chaleur **(Taine & Petit 2003)** :

1) La conduction. Transmission provoquée par la différence de température entre deux régions d'un milieu en contact physique. Il n'y a pas de déplacement appréciable des atomes ou molécules.
2) La convection. Transmission provoquée par le déplacement d'un fluide (liquide ou gazeux).
3) Le rayonnement. Transmission provoquée par la différence de température entre deux corps sans contact physique, mais séparés par un milieu transparent tel l'air ou le vide. Il s'agit d'un rayonnement électromagnétique.

I.6.3.2. Conduction thermique et la loi de Fourier

Dans le cas d'un champ de températures à une dimension **(Bianchi *et al*. 2004)**.

$$\dot{Q} = -\lambda A \frac{dT}{dx} \qquad \textbf{(I.01)}$$

- A: surface perpendiculaire au flux thermique
- λ: conductivité thermique du matériau
- $\frac{dT}{dx}$: Le gradient de température au point x considéré, c'est à dire la variation de la température par unité de longueur dans la direction x.

I.6.3.3. Conductivité thermique

La conductivité thermique : flux de chaleur qui traverse une surface unitaire quand le gradient de température est égal à l'unité. La conductivité thermique dépend de **(Taine & Petit 2003 ; Bianchi *et al*. 2004)** :

- La nature chimique du matériau.
- La nature de la phase considérée (solide, liquide, gazeuse).
- La température.
- L'orientation des fibres ou cristaux dans les corps anisotropes (bois, plastiques laminés, etc.).

I.6.4. Vitesses d'écoulement

Décrire le mouvement d'un fluide fait appel à des notions différentes de celles développées en Mécanique du point ou du solide. Le mouvement d'un fluide est un écoulement où il y a déformation continue du fluide On peut, de manière analogue à ce que l'on fait en Mécanique du solide, isoler (par la pensée ou en trouvant un moyen de visualisation, coloration par exemple) une partie restreinte du fluide appelée particule et la "suivre" au cours du temps c'est à dire connaître à chaque instant sa position.

Cette position sera connue, par exemple, par ses cartésiennes :$x_P(t, x_{0p,} y_{0p}, z_{0p})$, $y_P(t, x_{0p,} y_{0p}, z_{0p})$ et $z_P(t, x_{0p,} y_{0p}, z_{0p})$ ou $x_{0p,} y_{0p}, z_{0p}$ représentent les coordonnées de la particule choisie à l'instant t_0, la vitesse de la particule aura pour composantes $v_x = \frac{\partial x_p}{\partial t}$, $v_y = \frac{\partial y_p}{\partial t}$ et $v_z = \frac{\partial z_p}{\partial t}$ Au cours du temps, la particule sera en différents points M, l'ensemble des points M constitue la trajectoire de la particule. Cette façon de faire est appelée méthode de Lagrange, les variables introduites sont appelées variables de Lagrange. La méthode de Lagrange s'avère dans la plupart des cas délicate car il n'est pas facile de suivre les particules : elle est peu employée.

La méthode d'Euler consiste à connaître la vitesse des particules au cours du temps $\boldsymbol{t}$ à un endroit donné déterminé par ses coordonnées, par exemple cartésiennes x, y, z. Elle est plus employée que la méthode de Lagrange, la connaissance du champ des vitesses étant suffisante pour la description du fluide en mouvement. Les composantes du vecteur vitesse $\vec{v}$ sont des fonctions des variables (x, y, z, t), ainsi $\vec{v} = v_{x\vec{i}} + v_y\vec{j} + v_z\vec{k}$où $v_x = v_x(x, y, z, t)$, $v_y = v_y(x, y, z, t)$, et $v_z = v_z(x, y, z, t)$, L'écoulement du fluide est permanent ou stationnaire si ses composantes de vitesse sont indépendantes de la variable temps $\boldsymbol{t}$; il est dit non-permanent ou instationnaire si cette condition n'est pas réalisée.

L'écoulement du fluide est uniforme si ses composantes de vitesse sont indépendantes des coordonnées d'espace ; il est non-uniforme si cette condition n'est pas remplie.

Remarque : Dans la méthode d'Euler, l'accélération d'une particule peut être due, bien sûr, au caractère instationnaire de l'écoulement mais aussi à sa non-uniformité. Ainsi, chacun a pu constater, dans l'écoulement permanent d'une rivière, l'accélération des particules lors du franchissement d'un rétrécissement.

On appelle ligne de courant une courbe dont la direction tangente en chacun de ses points est la direction du vecteur vitesse **(Cengel & Cimbala 2017)**. L'équation d'une ligne de courant se calcule par intégration des équations. $\frac{dx}{v_x} = \frac{dy}{v_y} = \frac{dz}{v_z}$.

I.6.5. Viscosité

La viscosité est une autre propriété des fluides qu'il faut bien comprendre pour décrire divers phénomènes dans les équipements mécaniques, comme les pertes de pression dans la tuyauterie dues au frottement ou à la lubrification des roulements. La viscosité est une mesure de la résistance qu'oppose le fluide à l'écoulement, en raison de son frottement interne. On mesure la viscosité de deux façons : dynamique (absolue) et cinématique. Ces deux paramètres sont liés entre eux, car on obtient la viscosité cinématique en divisant la viscosité dynamique par la masse volumique **(Ravier *et al*. 2000)**.

I.6.6. Masse volumique

La masse volumique (parfois appelée densité) est une autre propriété de base des fluides. La masse volumique (désignée par la lettre grecque rho – ρ) est définie comme étant la masse (m) d'une unité de volume (v). Son unité de base est le kg/m^3. Mathématiquement, cela s'exprime comme suit **(Ravier *et al*. 2000)** : $\rho = \frac{m}{v}$;

I.7. Conclusion

L'environnement dans laquelle nous vivons est exposée quotidiennement à des nombreux facteurs de pollution, qu'elle soit naturelle ou humaine, qui sont pratiqués par des industries où cette dernière nécessite de jeter plusieurs déchets qui à leur tour polluent l'environnement et provoquent plusieurs types de pollutions, une pollution de l'air en augmentant les niveaux des éléments (O_3, N, CO_2). Et d'autres qui causent la pollution de l'air, et il y a aussi la pollution du sol par les engrais utilisés dans l'agriculture et l'exploration pétrolière, en plus de cela, la pollution de l'eau par les déchets d'usine déversés dans les mers et les rivières.

Nous avons essayé de voir quelques solutions pour lutter contre la pollution qui nous envahie :

- Conversion des énergies les plus émettrices (charbon, pétrole, gaz) vers les énergies renouvelables Conversion des énergies les plus émettrices (charbon, pétrole, gaz) vers les énergies renouvelables.
- Gestion raisonnée des forêts, protection des zones naturelles.
- Orientation de l'agriculture vers l'agroécologie, l'agroforesterie.
- Développer les transports en commun.
- Dans l'agriculture, limiter l'utilisation des produits phytosanitaires (améliorer le ciblage, privilégier les amendements naturels).
- Améliorer les processus industriels et la récupération des polluants.
- Généraliser le traitement des eaux usées.

Chapitre II.
Techniques de modélisation neuronal et méthodes de RQSP/RQSA

II.1. Introduction

II.1.1. Modélisation

La modélisation est un processus par lequel on organise les connaissances portant sur un système donné **(Peric *et al*. 2015)**. Modéliser c'est abstraire de la réalité une description d'un système dynamique **(Fishwick 1992)**. Un modèle est une représentation simplifiée et observable du comportement et/ou de la structure d'un système réel afin de résoudre un problème d'analyse ou de concept **(Varenne *et al*. 2018)**.

II.1.2. L'évolution de technique neuronale

En 1943, Mac Culloch et Pitts ont proposé le premier modèle de neurone biologique, **(Davalo *et al*. 1989)**. Ce dernier, appelé aussi neurone à seuil, a été inspiré des récentes découvertes en biologie. Ce sont des neurones logiques (0 ou 1). Ces deux physiciens ont montré que les neurones formels peuvent réaliser des fonctions logiques.

Les travaux se sont ralentis considérablement jusqu'aux années 80. En 1982, Hopfield a fini par démontrer l'intérêt des réseaux entièrement connectés **(Hopfield 1982 ; Davalo *et al*. 1989)**. Parallèlement, Werbos a conçu un mécanisme d'apprentissage pour les réseaux multicouches de type perceptron : la rétro-propagation (en anglais back-propagation). Cet algorithme, qui permet de propager l'erreur vers les couches cachées, sera popularisé en 1986 dans un livre "Parallel Distributed Processing" par Rumelhart et ces co auteurs **(Rumelhart *et al*. 1986)**.

En 1989 Moody et Darken **(Moody & Darken 1989)** exploitent quelques résultats de l'interpolation multi variables pour proposer le Réseau à Fonctions de base Radiales (RFR), connu sous l'appellation anglophone "Radial Basis Fonction network" (RBF). Ce type de réseau se distingue des autres types de réseaux de neurones par sa représentation locale **(Talbi 2011)**.

II.1.2.1 Applications et leurs importances

Une application dans l'exploration de données filtrage des informations **(Goyena & Fallis 2019)** :

- L'augmentation rapide du volume d'informations disponibles, notamment par voie électronique, rend obligatoire la conception et la mise en œuvre d'outils de filtrage des informations efficaces, qui permettent à l'utilisateur d'accéder uniquement aux informations pertinentes.

- Ces outils répondant aux besoins des professionnels, ils doivent être fiables et convivial. L'utilisateur peut accéder aux informations pertinentes soit en étant fourni avec la machine en texte intégral (recherche de texte), ou en étant fourni extraits de texte ou réponses à des questions (extraction d'informations).
- Une application en prévision : optimiser la probabilité de réussite, estimer une probabilité de réussite en fonction de la stratégie de dépenses utilisée et déduire des résultats la stratégie la plus appropriée.

Aussi y on a autre application comme :

- Les applications dans les tests non destructifs.
- Les Applications dans le contrôle de l'environnement : ozone Pollution et hydrologie urbaine.
- Les applications en robotique mobile.

II.1.3. L'évolution des méthodes de type RQSP/RQSA

Il y a plus d'un siècle et demi, en 1863, Cros **(Schrodinger 1926)** 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 RQSP.

A la fin du 19ème siècle, Meyer en 1899 et Overton en 1901 **(Kohn 1999)**, ont dépendamment observé « une relation linéaire entre l'activité des narcotiques et leur coefficient de partage huile-eau ».

En 1962, Hansen **(Mulliken 1928)** 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 RQSA 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 de Hansch et l'analyse de Free Wilson **(Thomas 1927)**. Sept ans plus tard, Hansch et Lien ont réalisé une étude RQSA 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 **(Dirac 1928 ; Chtita 2017)**.

Ces études ont été extrapolées aux techniques séparatives en corrélant les propriétés physicochimiques des analystes avec les temps de rétention obtenus expérimentalement : c'est l'étude quantitative des relations structure temps de rétention noté RQSR **(Hohenberg & Kohn 1964)**.

Maintenant, des méthodes 3D comme l'étude *CoMFA* (Comparative MolecularFieldAnalysis) et CoMSIA (Comparative Molecular Similarity Indices Analysis) **(Kohn & Sham 1965)** permettent de traiter les relations structure-activité en trois dimensions, 3D-RQSA/RQSP **(Chtita 2017)**.

II.1.3.1 Applications et leurs importances

Il existe un grand nombre d'applications des modèles QSAR en milieu industriel, dans le domaine de la recherche universitaire, en économie et en prévision météorologique **(Puzyn *et al*. 2010)**. Nous rapportons ci-dessous quelques applications possibles des modèles QSAR **(Peric *et al*. 2015)** :

- L'identification rationnelle de nouvelles pistes.
- L'optimisation de l'activité pharmacologique, biocide ou pesticide.
- La conception rationnelle de nombreux produits tels que des agents tensioactifs, des parfums, des colorants et des produits chimiques fins.
- L'identification des composés dangereux aux premiers stades de développement.
- La prédiction de la toxicité et les effets secondaires de nouveaux composés.
- La sélection des composés ayant des propriétés pharmacocinétiques optimales, que ce soit la stabilité ou la disponibilité dans les systèmes biologiques.

II.2 Objectifs de la modélisation

On considère que toutes les idées à la base de la modélisation ont comme vocation d'être remises en cause au contact du terrain, c'est-à-dire d'être volontairement et directement soumises à réfutation **(Etienne *et al*. 2015)** :

- Avoir aucune hypothèse implicite dans l'expérimentation : ceci constitue un objectif en soi de la démarche et implique le développement de procédures de mise en évidence de

telles hypothèses à la fois sur le plan empirique et conceptuel, ce dernier plan consistant à confronter le modèle à des résultats reconnus par la communauté scientifique.

- Prendre en considération, dès les premières étapes, l'impact du processus de recherche sur le terrain, que ce soit en termes d'objectifs définis, de qualité de la démarche, d'indicateurs quantifiables, de suivi ou d'évaluation ;
- Donner une attention particulière aux processus de validation de cette démarche de recherche, sachant qu'il n'existe pas de théorie générale de la validation des modèles et que des procédures autres que celles utilisées dans les modèles bio-physicomathématiques sont à envisager.

II.3. Quelques techniques de la modélisation

Il existe une chaine vaste de techniques de modélisation, et dans ce cas-là on propose des méthodes sont très souvent utilisées dans lesquelles :

II.3.1. Régression linéaire multiple

La régression multi-linéaire (MLR) est la méthode la plus simple et la plus communément employée pour le développement de modèles prédictifs **(Lejeune 2004)**. Elle repose sur l'hypothèse qu'il existe une relation linéaire entre une variable dépendante y (ici, la propriété physicochimique) et une série de n variables indépendantes x_i (ici, les descripteurs moléculaires). L'objectif est d'obtenir une équation de la forme suivante **(Hamad & Khahla 2014)** :

$$y = a_0 + a_1x_1 + a_2x_2 \cdot \cdots + a_nx_n \qquad \textbf{(II.01)}$$

II.3.2. Modèle de surface de réponse

La technique de surface de réponse (RSM) est destinée à fournir un outil de substitution de modélisation qui peut simuler efficacement en temps réel Impacts des MP (particulate matter) pour diverses alternatives réglementaires à utiliser dans les analyses d'impact de la réglementation. Pour exemple, générer des estimations des avantages pour la santé de la réduction des précurseurs de fournir des estimations de niveau de dépistage des impacts des stratégies de contrôle sur les valeurs de conception du NAAQS sont des fonctions prises en charge par RSM. Bien que le RSM ne fournisse pas une image complète de tous les changements nécessaires pour atteindre diverses normes alternatives à l'échelle nationale, il est très utile contexte de fournir des scénarios de contrôle illustratifs pour des zones sélectionnées, et de comprendre contribution de différentes catégories de sources, régions sources et émissions de

polluants à la qualité de l'air au États-Unis Le RSM peut être utilisé de différentes façons **(US Environmental Protection Agency 2006)** :

- Conception et évaluation de la stratégie (Par exemple, comparaison des contrôles urbains et régionaux, comparaison entre les secteurs, comparaison entre polluants).
- Optimisation (développer des combinaisons optimales de contrôles pour atteindre les Coût minimum).
- Sensibilité du modèle (évaluer systématiquement la sensibilité relative des niveaux d'ozone et de particules aux changements dans les émissions.

II.4. Réseaux de neurones artificiels

Les réseaux de neurones artificiels sont présentés comme la vague du futur en informatique. Ce sont en effet des mécanismes d'autoapprentissage qui ne nécessitent pas les compétences traditionnelles d'un programmeur.

Les réseaux de neurones artificiels sont des réseaux fortement connectés de processeurs élémentaires fonctionnant en parallèle. Chaque processeur élémentaire calcule une sortie unique sur la base des informations qu'il reçoit. Toute structure hiérarchique de réseaux est évidemment un réseau **(De la Fuente Aparicio & Cano 1999)**.

II.4.1. Neurone

Le neurone est une cellule composée d'un corps cellulaire et d'un noyau. Le corps cellulaire se ramifie pour former ce que l'on nomme les dendrites. Celles-ci sont parfois si nombreuses que l'on parle alors de chevelure dendritique ou d'arborisation dendritique. C'est par les dendrites que l'information est acheminée de l'extérieur vers le soma, corps du neurone. L'information traitée par le neurone chemine ensuite le long de l'axone (unique) pour être transmise aux autres neurones. La transmission entre deux neurones n'est pas directe. En fait, il existe un espace intercellulaire de quelques dizaines d'Angströms (10-9 m) entre l'axone du neurone afférent et les dendrites (on dit une dendrite) du neurone efférent. La jonction entre deux neurones est appelée la synapse **(Brégains 2007)**.

II.4.2. Bases biologiques

Le neurone est le processeur élémentaire du cerveau. Chaque neurone traite l'information qui lui parvient localement, puis transmet aux autres neurones qui lui sont connectés l'information qu'il a traitée. Ces cellules peuvent apprendre en changeant l'intensité de

leurs connexions avec d'autres cellules ou détruire ou même créer de nouvelles connexions. Elles peuvent aussi changer leur(s) règle(s) de traitement de l'information. Ce processus de changement est appelé apprentissage et joue un rôle fondamental dans le comportement du neurone **(Brégains 2007)**.

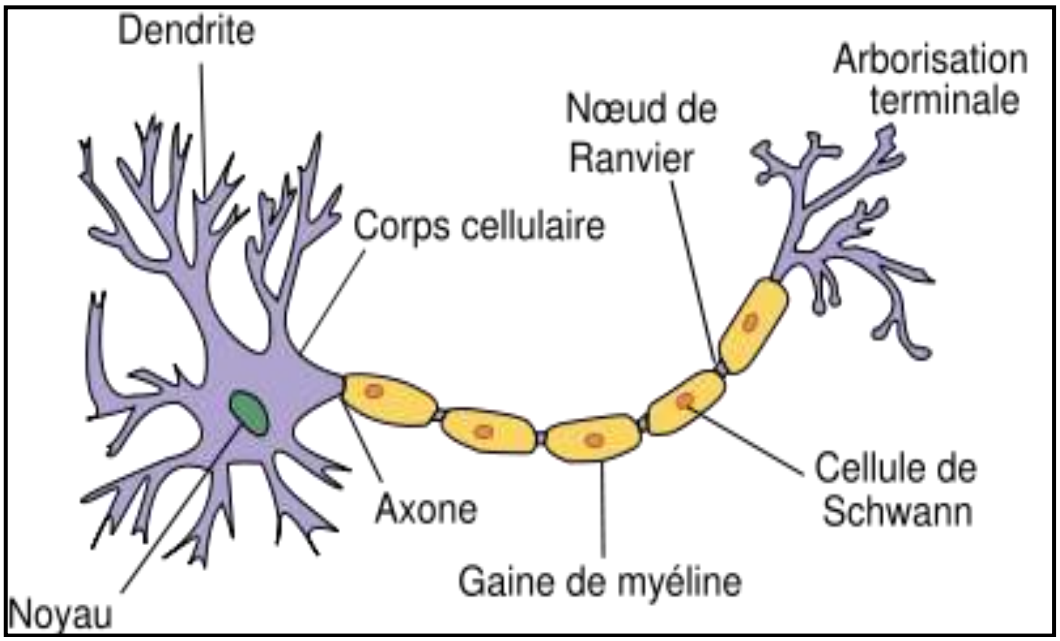


Figure II.01. Représentation d'un neurone réel.

Vous pouvez constater que le neurone est composé des parties suivantes **(Brégains 2007)** :

- **Corps cellulaire :** Noyau du neurone.
- **Dendrites :** Récepteurs principaux du neurone.
- **Axone :** Fibre nerveuse de transport pour les signaux émis par le neurone.
- **Synapse :** Connexion entre la dendrite et l'axone de deux neurones.

II.4.3. Neurone formel

II.4.3.1. Définition

Neurone formel est un processeur élémentaire qui réalise une somme pondérée des signaux qui lui parviennent. La valeur de cette sommation est comparée à un seuil et la sortie des neurones une fonction linéaire **(López *et al*. 2008)**.

Un neurone est une fonction paramétrée non linéaire de ses variables d'entrée. Naturellement, un réseau de neurones est la composition des fonctions non linéaires de deux neurones ou plus **(Díez *et al*. 2001)**.

Les variables du neurone sont souvent les valeurs d'entrées du neurone et les valeurs de sa sortie. Les réseaux de neurones artificiels sont des réseaux fortement connectés de processeurs élémentaires fonctionnant en parallèle. Chaque processeur élémentaire calcule une sortie unique sur la base des informations qu'il reçoit **(De la Fuente Aparicio & Cano 1999)**.

II.4.3.2. Paramètres

Les paramètres sont affectés aux entrées des neurones ; la sortie du neurone est une combinaison non linéaire des entrées {xi}, pondérées par les paramètres {Wi}, souvent appelées poids.

Le potentiel v le plus fréquemment utilisé est la somme pondérée des entrées, avec un terme constant supplémentaire appelé « biais » **(Díez *et al*. 2001)**.

Les formules à appliquer pour mettre à jour les poids et biais lorsque l'on entre un individu dans le réseau sont donc :

$$\boldsymbol{w_{new} = w_{old} + zve(p_i)P_i} \text{ et } \boldsymbol{p_{new} = p_{old} + zve(p_i)P_i}$$

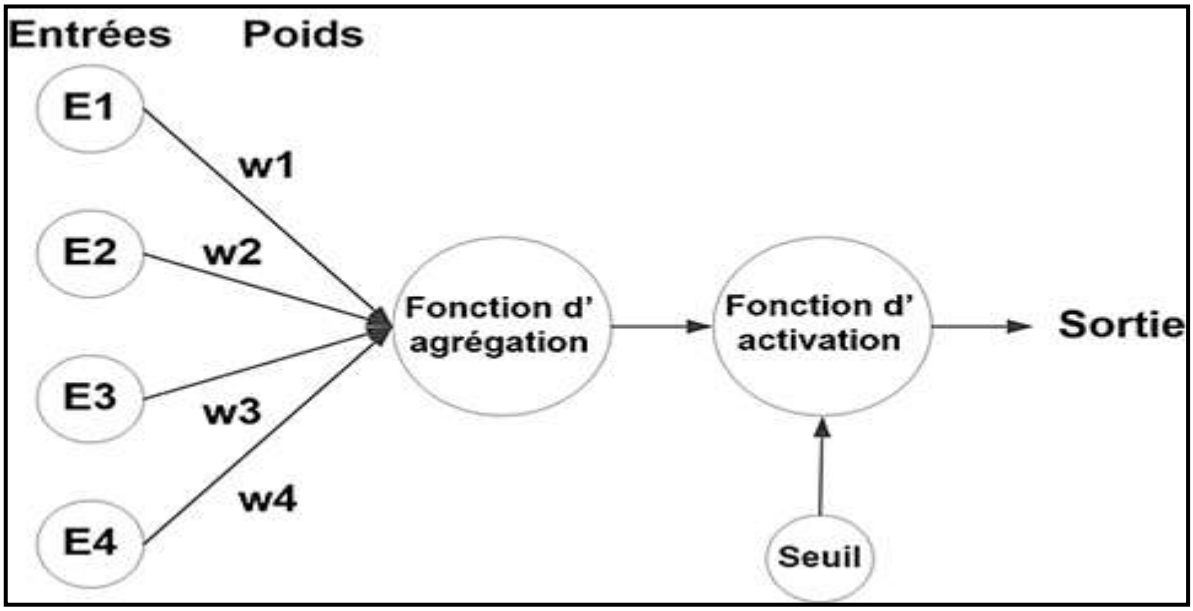


Figure II.02. Modèle de base d'un neurone formel.

Nous pouvons définir un neurone par les trois éléments suivants **(Brégains 2007)** :

- La fonction d'entrée totale qui définit le prétraitement effectué sur les entrées.
- La fonction d'activation (ou d'état) du neurone qui définit son état interne en fonction de son entrée totale.
- La fonction de sortie qui calcule la sortie du neurone en fonction de son état d'activation.

II.4.3.3. Principe de fonctionnement

Un neurone reçoit des entrées et fournit une sortie, grâce à différentes caractéristiques **(Douali *et al*. 2003)** :

- Des **poids** accordés à chacune des entrées, permettant de modifier l'importance de certaines par rapport aux autres.
- Une **fonction d'agrégation**, qui permet de calculer une unique valeur à partir des entrées et des poids correspondants.
- Un **seuil** (ou biais), permettant d'indiquer quand le neurone doit agir.
- Une **fonction d'activation**, qui associe à chaque valeur agrégée une unique valeur de sortie dépendant du seuil.
- Pas de notion temporelle
- Coefficient synaptique : coefficient réel
- Sommation des signaux arrivant au neurone
- Sortie obtenue après application d'une fonction de transfert **(Andrews *et al*. 1995)**.

II.4.4. Architecture des réseaux de neurones

L'architecture d'un réseau de neurones définit son fonctionnement et joue un rôle important dans son comportement. Elle est en fonction de ses couches et de la structure des connexions de ses neurones ; ces paramètres permettent de distinguer les différentes classes et/ou types d'architecture neuronales **(Frisch *et al*. 2009)**. Donc Il existe trois grands types d'architectures de réseaux de neurones :

II.4.4.1. Réseaux statiques

Ce type de réseaux est organisé généralement en couches de neurones. Chaque neurone d'une couche reçoit ses entrées à partir des neurones de la couche précédente ou tout simplement de l'entrée du réseau. Dans tels réseaux il n'existe pas de feed-back (boucles de retour d'informations). Le traitement des données se fait en sens unique et le flux d'informations circule directement de la couche d'entrée à la couche de sortie ; le traitement est donc réalisé en boucle ouverte. Ces réseaux peuvent être utilisés pour les problèmes de classification ou d'approximation des fonctions (**Frisch *et al*. 2009**).

II.4.4.2. Réseaux dynamiques

Les réseaux dynamiques ou bien récurrents, sont les réseaux pouvant comporter des boucles (feed-back) entre les neurones. En général, la sortie de chaque neurone peut être envoyée vers l'entrée de tous les autres neurones du réseau. Ainsi, ces boucles ramènent l'information en

provenance de la couche de sortie sur la couche d'entrée simultanément avec le signal d'entrée présent au même instant. Le réseau dynamique peut donner une sortie différente en lui présentant la même entrée à des instants différents, contrairement aux réseaux statiques qui donnent la même sortie pour une même entrée. Au début, ces réseaux étaient souvent utilisés pour les problèmes de classification et de mémorisation. Actuellement, dans plusieurs travaux la structure interne dynamique de ces réseaux est exploitée pour l'identification ou la commande des systèmes dynamiques **(Frisch *et al*. 2009)**.

II.4.4.3. Réseaux autoorganisés

Les réseaux de neurones autoorganisés sont des réseaux qui changent leurs structures internes pendant l'apprentissage. Ainsi les neurones se regroupent typologiquement suivant la représentation des exemples. Ces réseaux sont des dérivées des modèles de Kohonen **(Frisch *et al*. 2009)**.

II.4.5. Apprentissage

L'apprentissage est vraisemblablement la propriété la plus intéressante des réseaux neuronaux. Elle ne concerne cependant pas tous les modèles, mais les plus utilisés.
Une fois qu'un réseau a été structuré pour une application particulière, ce réseau est prêt à être formé. Pour démarrer ce processus, les poids initiaux sont choisis de manière aléatoire. Ensuite, l'apprentissage commence.

II.4.5.1. Définition

L'apprentissage est la modification des poids du réseau dans l'optique d'accorder la réponse du réseau aux exemples et à l'expérience. Il est souvent impossible de décider à priori des valeurs des poids des connexions d'un réseau pour une application donnée. A l'issu de l'apprentissage, les poids sont fixés : c'est alors la phase d'utilisation **(Andrews *et al*. 1995)**. Au niveau des algorithmes d'apprentissage, Il existe deux approches d'apprentissage :

II.4.5.2. Apprentissage supervisé

Apprentissage supervisé implique un mécanisme fournissant au réseau le résultat souhaité, soit en "évaluant" manuellement les performances du réseau, soit en fournissant les résultats souhaités avec les entrées. La grande majorité des réseaux utilisent l'apprentissage supervisée. Les coefficients synaptiques sont évalués en minimisant l'erreur (entre sortie souhaitée et sortie obtenue) sur une base d'apprentissage **(Andrews *et al*. 1995)**.

II.4.5.3. Apprentissage non supervisé

Apprentissage non supervisé est l'endroit où le réseau doit donner un sens aux entrées sans aide extérieure, il est utilisé pour effectuer une caractérisation initiale des entrées. On ne dispose pas de base d'apprentissage. Les coefficients synaptiques sont déterminés par rapport à des critères de conformité : spécifications générales **(Andrews *et al*. 1995)**.

II.4.6. Perceptron mono couche

Avant d'aborder le comportement collectif d'un ensemble de neurones, nous allons présenter le Perceptron (un seul neurone) en phase d'utilisation. L'apprentissage ayant été réalisé, les poids sont fixes. Le neurone de la figure 3 réalise une simple somme pondérée de ses entrées, compare une valeur de seuil, et fourni une réponse binaire en sortie **(Stewart 2009)**.

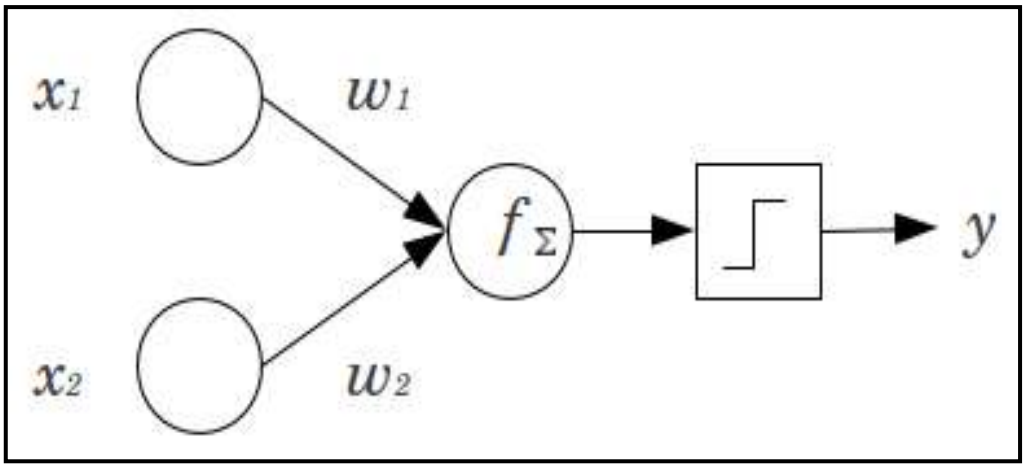


Figure II.03. Structure et comportement d'un perceptron mono couche.

II.4.7. Perceptron multicouche

D'une façon générale, un perceptron multicouche (Multi Layer Perceptron) contient une couche d'entrée, une ou plusieurs couches cachées, chargées des calculs possédant une fonction d'activation et une couche de sortie. La couche d'entrée représente toujours une couche virtuelle associée aux entrées du système, elle ne contient aucun neurone tandis que les couches suivantes sont des couches de neurones. La couche de sortie correspond aux sorties du système **(Stewart 2009)**.

Dans le modèle du Perceptron Multi–Couches, les perceptrons sont organisés en couches. Les perceptrons multi–couches sont capables de traiter des données qui ne sont pas linéairement séparables Avec l'arrivée des algorithmes de rétro propagation, ils deviennent le type de réseaux de neurones le plus utilisé **(HyperCube 2003)**.

L'utilité de plusieurs couches cachées n'a pas été démontrée. La **figure II.04** illustre la structure d'un MLP présentant quatre neurones en entrée, trois neurones sur la couche cachée et deux en sortie **(HyperCube 2003)**.

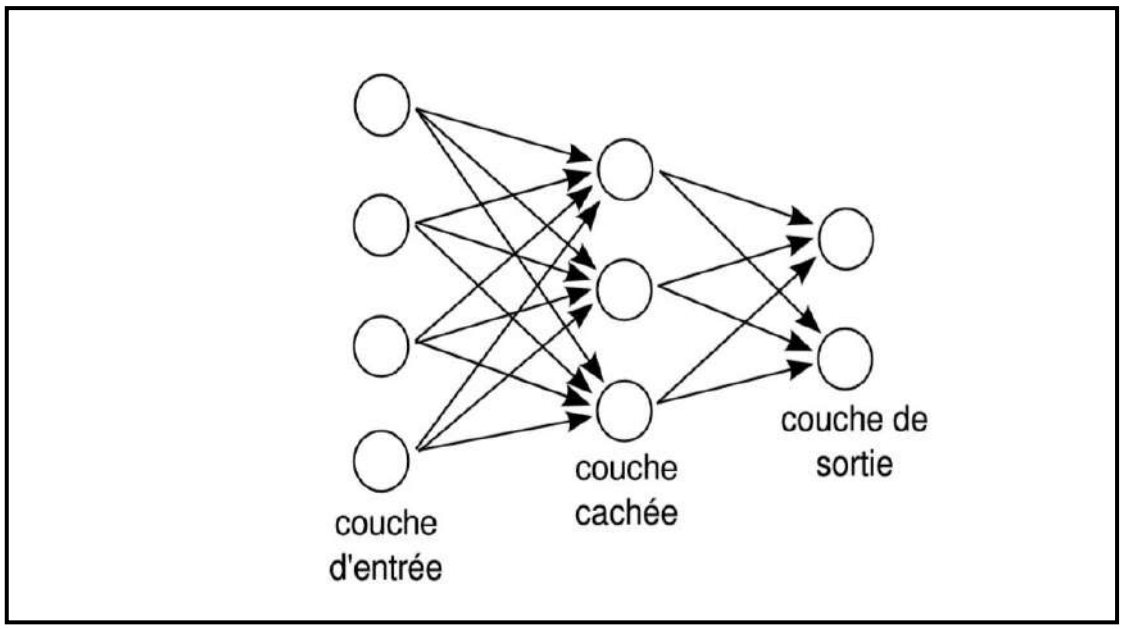


Figure II.04. Modèle du perceptron multi couches.

Lorsque tous les neurones d'une couche sont connectés aux neurones de la couche suivante, on parle alors de couches complètement connectées. Les équations de propagation décrites plus haut s'appliquent à tous les neurones. Cependant, le passage d'une couche à l'autre peut être formalisé sous forme matricielle. Soit un MLP dont le nombre de neurones sur la couche d'entrée est n0, n1 sur la couche cachée et n2 sur la couche de sortie. Les poids de la couche cachée peuvent s'écrire sous la forme d'une matrice W ∈ Rn1×n0 **(HyperCube 2003)**.

Pour un vecteur X ∈ Rn0 présenté en entrée, le vecteur potentiel V et le vecteur de sortie Y pour la couche cachée s'écrivent donc **(HyperCube 2003)** :

- $V = WX$
- $Y = \phi(V)$

Avec Φ une fonction d'activation. La sortie de la couche cachée devient ensuite l'entrée de la dernière couche.

- ***Mise à jour des poids :*** Une fois la rétro propagation du gradient effectuée pour toutes les couches, les matrices de poids sont mises à jour **(HyperCube 2003)** : $\boldsymbol{w \leftarrow w - \lambda\, dw}$, Où $\boldsymbol{\lambda}$ est le taux d'apprentissage.

II.5. Modélisation par les approches RQSP/RQSA

II.5.1. Définition de RQSP

La modélisation QSAR sur un ensemble de produits chimiques structurellement liés fait référence au développement d'une corrélation mathématique entre une réponse chimique et quantitative attributs chimiques définissant les caractéristiques des molécules analysées. Par conséquent, ces études tentent d'établir un formalisme mathématique entre le comportement d'un chimique, c'est-à-dire la réponse chimique et un ensemble d'attributs chimiques quantitatifs qui peut être extrait des structures chimiques en utilisant des moyens expérimentaux ou théoriques appropriés. Le nom de l'étude dépend de la nature de la réponse (également appelé point final) étant modélisé donnant naissance à trois classes principales, à savoir la relation quantitative structure – propriété / activité / toxicité (QSPR / QSAR / QSTR) études portant sur la modélisation des propriétés physicochimiques, l'activité biologiques et les données toxicologiques, respectivement. La nomenclature peut également être employés pour définir des paramètres plus précis tels que la structure quantitative - relation de cytotoxicité pour désigner la modélisation de la cytotoxicité des produits chimiques. Sur le d'autre part, QSPR, c'est-à-dire la modélisation quantitative de la relation structure-propriété, peut être employé pour désigner toutes les techniques connexes telles que tout type des comportements toxicologiques et physico-chimiques peuvent être considérés « Propriété » d'un produit chimique donné. Cependant, nous utiliserons le terme « QSAR » pour désigner toutes ces études. Puisqu'une relation mathématique est développée, ces études permettent la prédiction du comportement moléculaire de nouveaux produits chimiques ou même hypothétique molécules **(Cartwright 2015)**.

II.5.2. Méthodologie générale

Lors du développement des modèles QSPR, plusieurs étapes ont été utilisées **(Roubehie Fissa *et al*. 2019)** :

- Collection de bases de données : la qualité de la base de données expérimentale est la phase essentielle de tout modèle mathématique réussi.
- Calcul des descripteurs moléculaires : L'une des étapes les plus importantes de la construction du QSPR modèles est la quantification des informations de structure de l'étude des molécules.
- Prétraitement pour la réduction et la sélection des descripteurs.
- Construction du modèle de réseau de neurones artificiels.

II.5.3. Notation SMILES

SMILES est un langage de notation chimique spécialement conçu pour être utilisé par les chimistes sur ordinateur. Il est facilement accessible aux chimistes, mais suffisamment flexible pour permettre l'interprétation et génération de notation chimique indépendante du spécifique système informatique utilisé. Similaire aux produits chimiques conventionnels notation, il améliore les méthodes logicielles conventionnelles en une plus grande vitesse et une meilleure utilisation de la capacité informatique. Moléculaire les structures sont spécifiées de manière unique et précise et peuvent être utilisé avec des bases de données chimiques. Parmi plusieurs approches pour la notation chimique informatisée, la notation linéaire est populaire car elle représente la structure moléculaire par une chaîne linéaire de symboles similaires au langage naturel. La ligne Wiswesser Notation9 est le représentant le plus largement utilisé de cette méthode. Il répond aux exigences essentielles d'une approche déterministe la notation chimique, mais elle est difficile à utiliser car de nombreuses règles doit être suivi pour générer la notation correcte d'un complexe structure. Pour surmonter cette difficulté et d'autres **(Weininger 1988)**.

Le système SMILES a été conçu pour être vraiment interactif avec l'ordinateur. La simplicité de son utilisation repose sur des programmes informatiques recoder rigoureusement les données de l'utilisateur de produits chimiques. C'est le résultat d'atteindre les objectifs originaux suivants **(Weininger 1988)** :

- Le graphique d'une structure chimique devait être unique décrit, y compris, mais sans s'y limiter, le graphique moléculaire comprenant des nœuds (atomes) et des bords (liaisons).
- Une spécification de structure conviviale devait être fournie, afin que toutes les règles d'entrée puissent être apprises rapidement et naturellement.
- Un système convivial et indépendant de la machine devait être conçu pour l'interprétation et la production d'une notation unique.

Contrairement à d'autres systèmes de notation chimique, la brièveté de la notation et l'économie de l'alphabet n'étaient pas des objectifs principaux. Beaucoup des pièges dans d'autres notations de ligne peuvent être attribués à surutilisation des symboles et des règles hiérarchiques en fonction de la longueur de la notation finale. Les progrès du matériel informatique ont rendu ces restrictions obsolètes. La présente approche sépare la description non ambiguë mais générale d'un produit chimique structure (par le chimiste-utilisateur) à partir de la génération de la description structurelle unique (par l'ordinateur). Le dernier nécessite des règles et des hiérarchies qui sont intrinsèquement difficiles pour le chimiste. Cette tâche est donc reléguée à l'ordinateur algorithmes **(Weininger 1988)**.

II.5.4. Structure chimique des molécules

II.5.4.1 Atomes

Les atomes sont représentés par leurs symboles atomiques ; c'est la seule utilisation obligatoire des lettres dans SMILES. Chaque l'atome non hydrogène est spécifié indépendamment par son atome symbole placé entre crochets. La deuxième lettre de les symboles à deux caractères doit être entrés en minuscules. Éléments du « sous-ensemble organique », B, C, N, 0, P, S, F, C1, Br et I, peut être écrit sans parenthèses si le nombre de pièces jointes les hydrogènes se conforment à la plus basse valence normale cohérente avec des liens explicites. Les atomes dans les anneaux aromatiques sont spécifiés par minuscules ; par exemple, le carbone normal est représenté par la lettre C, carbone aromatique par c. Étant donné que les hydrogènes attachés sont implicites en l'absence de parenthèses, les symboles atomiques suivants sont des notations SMILES valides **(Weininger 1988)**.

- P ; phosphine (PH)
- S ; sulfure d'hydrogène (H2S)
- C1 ; chlorure d'hydrogène (HC1)

II.5.4.2. Liaisons

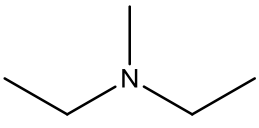
Les liaisons simples, doubles, triples et aromatiques sont :

- Représentés respectivement par les symboles **–** ; **=** ; **#** et **:**
- Les liaisons simples et aromatiques peuvent être, et sont généralement, omises.
- Par exemple : c = c éthylène (CH2 = CH2) **(Weininger 1988)**.

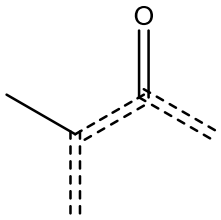
II.5.4.3. Succursales

Les branches sont spécifiées par des pièces jointes dans des parenthèses.

Par exemple :



- CCN(CC)CC Triethylamine.



- Cc(c)c(=o)c Acide isobutyrique.

- ***Structures cycliques***

Les structures cycliques sont représentées par rompre une seule liaison (ou aromatique) dans chaque cycle. Le les obligations sont numérotées dans n'importe quel ordre, désignant l'ouverture du cycle (ou anneau de fermeture) par un chiffre immédiatement après le symbole atomique à chaque fermeture d'anneau **(Weininger 1988)**.

- ***Structure non liée***

Les composés déconnectés sont écrits comme des structures individuelles séparées par une période l'ordre dans lequel les ions ou les ligands sont répertoriés est arbitraire **(Weininger 1988)**.

- ***Aromaticité***

Les structures aromatiques peuvent être distinguées en écrivant les atomes du cycle aromatique en lettres minuscules **(Weininger 1988)**.

II.5.5. Descripteurs moléculaires

Le descripteur moléculaire est le résultat final d'une procédure logique et mathématique qui transforme les informations chimiques encodées dans une représentation symbolique d'une molécule en un nombre utile ou le résultat d'une expérience standardisée. Une attention particulière est portée au terme « utile » avec sa double signification : cela signifie nombre peut donner un meilleur aperçu de l'interprétation des propriétés moléculaires et / ou est en mesure de participer à un modèle pour la prédiction d'une propriété intéressante de d'autres molécules **(Todeschini & Consonni 2000)**.

Le domaine des descripteurs moléculaires est fortement interdisciplinaire et implique une masse de différentes théories. Pour la définition des descripteurs moléculaires, une connaissance algèbre, théorie des graphes, théorie de l'information, chimie computationnelle et théories de la réactivité organique et la chimie physique sont généralement nécessaires, bien qu'à niveaux.

Les descripteurs moléculaires sont fréquemment classés par rapport à la dimensionnalité de la représentation moléculaire sur laquelle ils sont calculés, on parlera alors de descripteurs 0D, 1D, 2D, 3D ou 4D **(Dudek *et al*. 2006)**.

II.5.5.1. Descripteurs 0D

Tous les descripteurs moléculaires pour lesquels aucune information sur la structure moléculaire et les connectivités atomiques n'est nécessaire appartiennent à la classe des descripteurs 0D, les nombres d'atomes et de liaisons ainsi que la somme ou la moyenne des propriétés atomiques sont typiques de cette classe de descripteurs. Ces descripteurs peuvent toujours être facilement calculés et interprétés naturellement, ils ne demandent pas d'optimisation de la structure moléculaire, ils sont indépendants de tout problème conformationnelle. Généralement ils montrent une très forte dégénérescence, c'est-à-dire qu'ils ont des valeurs égales pour plusieurs molécules, telles que les isomères, leur contenu informationnel est faible, mais ils peuvent néanmoins jouer un rôle important dans la modélisation de plusieurs propriétés physicochimiques ou prendre part à des modèles plus complexes **(Bonachera 2011)**.

II.5.5.2. Descripteurs 1D

Ces descripteurs sont calculés à partir de la formule brute de la molécule à l'aide de la composition moléculaire, c'est-à-dire les atomes qui la constituent et ils représentent des propriétés générales tels que les pourcentages massiques des atomes, la masse molaire et le poids moléculaire **(Cramer *et al*. 1988 ; Goulon-Sigwalt-Abram 2008 ; Bonachera 2011)**.

II.5.5.3. Descripteurs 2D

Ils sont calculés à partir des formules développées des molécules étudiées, ils peuvent être de plusieurs types comme les indices constitutionnels caractérisent les différents composants de la molécule. Il s'agit par exemple du nombre de liaisons simples ou multiples, du nombre de cycles etc. Les indices topologiques peuvent être obtenus à partir de la structure 2D de la molécule et donnent des informations sur sa taille, la forme globale et les ramifications. L'indice de Wiener par exemple permet de caractériser le volume moléculaire et la ramification d'une molécule, un autre indice nommé "indice de Randic" qui est l'un des descripteurs les plus utilisés, il peut être interprété comme une mesure de l'aire de la molécule accessible au solvant **(Cramer *et al*. 1988 ; Goulon-Sigwalt-Abram 2008 ; Bonachera 2011)**.

II.5.5.4. Descripteurs 3D

Ce type de descripteurs nécessitent une conformation 3D de la molécule, ils sont évalués à partir des positions relatives de leurs atomes dans l'espace et décrivent des caractéristiques plus complexes, leurs calculs nécessitent donc de connaître le plus souvent par "modélisation moléculaire empirique" ou "ab-initio" et la géométrie 3D de la molécule. La plupart de ces descripteurs s'avèrent relativement coûteux en temps de calcul, mais apportent davantage d'informations et sont nécessaires à la modélisation des propriétés ou d'activités qui dépendent de la structure 3D **(Cramer *et al*. 1988 ; Goulon-Sigwalt-Abram 2008 ; Bonachera 2011)**.

II.5.5.5. Descripteurs 4D

Ils correspondent à la mesure des propriétés 3D, comme le potentiel électrostatique et l'hydrophobicité d'une molécule en tout point de l'espace. Ils permettent d'avoir l'information sur la structure de la cible (protéine). On pourra ainsi distinguer les descripteurs 4D qui nécessitent un alignement de la molécule guidé par l'étude des complexes ligand-cible avant d'être calculés, ces descripteurs sont obtenus par le calcul des champs d'interactions moléculaires entre une molécule et une sonde représentée par une autre molécule comme l'eau **(Cramer *et al*. 1988 ; Goulon-Sigwalt-Abram 2008 ; Bonachera 2011)**.

II.5.6. Calcule des descripteurs moléculaires

Le calcul des descripteurs a pour but d'exploiter un maximum d'informations contenues dans des structures moléculaires, ils sont traduits par une série de grandeurs en général scalaires, qui permet de quantifient leurs caractéristiques structurelles. Dans nos jours les descripteurs moléculaires ont calculé par des logiciels et plateformes en ligne spécialisés, qui sont parfois commerciales et parfois gratuites, tels que ChemDes **(Dong *et al*. 2015)**, Gaussian **(Frisch *et al*. 2004)**, ChemOffice **(ChemOffice 2012)**, ChemSketch **(ACD/Labs 2015)**, Chemaxon MarvinSketch **(Csizmadia 1999 ; MarvinSketch 2012),** QSARINS **(Gramatica *et al*. 2013)**, VolSurf **(Cruciani *et al*. 2000)** et Dragon **(Todeschini & Consonni 2009)**.

II.6. Techniques d'analyse et de validation des modèles de RNA

Les réseaux de neurones réalisent des fonctions non linéaires paramétrées. Leurs mises en œuvre nécessitent **(Si-Moussa 2010)** :

- La détermination des entrées et des sorties pertinentes, c'est à dire les grandeurs qui ont une influence significative sur le phénomène que l'on cherche à modéliser.

- La collecte des données nécessaires à l'apprentissage et à l'évaluation des performances du réseau de neurones.
- La détermination du nombre de neurones cachés nécessaires pour obtenir une approximation satisfaisante.
- La réalisation de l'apprentissage
- L'évaluation des performances du réseau de neurones à l'issue de l'apprentissage.

II.6.1. Détermination des entrées/sorties du réseau de neurones

Pour toute conception de modèle la sélection des entrées doit prendre en compte deux points essentiels **(Si-Moussa 2010)** :

Premièrement, la dimension intrinsèque du vecteur des entrées doit être aussi petite que possible, en d'autres termes, la représentation des entrées doit être la plus compacte possible, tout en conservant pour l'essentiel la même quantité d'information, et en gardant à l'esprit que les différentes entrées doivent être indépendantes.

En second lieu, toutes les informations présentées dans les entrées doivent être pertinentes pour la grandeur que l'on cherche à modéliser : elles doivent donc avoir une influence réelle sur la valeur de la sortie.

II.6.2. Choix et préparation des échantillons

Le processus d'élaboration d'un réseau de neurones commence toujours par le choix et la préparation des échantillons de données. La façon dont se présente l'échantillon conditionne le type de réseau, le nombre de cellules d'entrée, le nombre de cellules de sortie et la façon dont il faudra mener l'apprentissage, les tests et la validation **(Bishop 1995)**.

Il faut donc déterminer les grandeurs qui ont une influence significative sur le phénomène que l'on cherche à modéliser. Lorsque la grandeur que l'on veut modéliser dépend de nombreux facteurs, c'est-à-dire lorsque le modèle possède de nombreuses entrées, il n'est pas possible de réaliser un pavage régulier dans tout le domaine de variation des entrées : il faut donc trouver une méthode permettant de réaliser uniquement des expériences qui apportent une information significative pour l'apprentissage du modèle. Cet objectif peut être obtenu en mettant en œuvre un plan d'expériences. Pour les modèles linéaires, l'élaboration de plans d'expériences est bien maîtrisée, par ailleurs, ce n'est pas le cas pour les modèles non linéaires **(Bitam 2018)**.

Afin de développer une application à base de réseaux de neurones, il est nécessaire de disposer de deux bases de données, une pour effectuer l'apprentissage et l'autre pour tester le réseau obtenu et déterminer ses performances. Notons qu'il n'y a pas de règle pour déterminer ce partage d'une manière quantitative, néanmoins chaque base doit satisfaire aux contraintes de représentativité de chaque classe de données et doit généralement refléter la distribution réelle, c'est à dire la probabilité d'occurrence des diverses classes **(Bitam 2018)**.

II.6.3. Elaboration de la structure du réseau

La structure du réseau dépend étroitement du type des échantillons. Il faut d'abord choisir le type de réseau : Un perceptron standard, un réseau de Hopfield, un réseau à décalage temporel (TDNN), un réseau de Kohonen, un ARTMAP etc... Par exemple, dans le cas du perceptron multicouches, il faudra aussi bien choisir le nombre de couches cachées que le nombre de neurones dans cette couche **(Si-Moussa 2010 ; Hamadache 2014)**.

II.6.4. Apprentissage

L'apprentissage est un problème numérique d'optimisation. Il consiste à calculer les pondérations optimales des différentes liaisons, en utilisant un échantillon. La méthode la plus utilisée est la rétropropagation, qui est généralement plus économe que les autres en termes de nombres d'opérations arithmétiques à effectuer pour évaluer le gradient. Pour rendre l'optimisation plus performante, on peut utiliser des méthodes de second ordre. Le calcul est très efficace, mais lourd. Elles ont de nombreuses limitations, quant aux conditions de convergence, sur les dérivées secondes. Des corrections sont proposées pour éviter ce problème, et sont prises en compte par les méthodes dites de Quasi-Newton ou de Newton modifiée **(Si-Moussa 2010)**.

II.6.5. Validation et tests

Alors que les tests concernent la vérification des performances d'un réseau de neurones hors échantillon et sa capacité de généralisation, la validation est parfois utilisée lors de l'apprentissage. Une fois le réseau de neurones développé, des tests s'impose afin de vérifier la qualité des prévisions du modèle neuronal **(Hamadache 2014)**.

Cette dernière étape doit permettre d'estimer la qualité du réseau obtenu en lui présentant des exemples qui ne font pas partie de l'ensemble d'apprentissage. Une validation rigoureuse du modèle développé se traduit par une proportion importante de prédictions exactes sur l'ensemble de la validation. Si les performances du réseau ne sont pas satisfaisantes, il

faudra, soit modifier l'architecture du réseau, soit modifier la base d'apprentissage **(Hamadache 2014)**.

II.7. Techniques d'analyse et de validation des modèles de QSPR

Afin d'évaluer l'importance des modèles RQSA/RQSP et, par conséquent, ces capacités de prédiction des activités/propriétés d'autres (nouveaux) composés, la validation des modèles RQSA/RQSP reste une étape très sensible dans les études statistiques. Un modèle étant le résultat d'une analyse statistique, son interprétation et son application doivent se faire dans le cadre très précis du domaine couvert par l'analyse **(Chtita 2017)**.

Toute application hors de ce cadre exige beaucoup de précautions et elle est d'autant plus hasardeuse qu'on s'éloigne du cadre. Pour éviter les erreurs, tant au moment de la validation qu'au moment de l'exploitation, les limites du modèle doivent être clairement établies : le non-robuste du modèle doit être vérifié, les pouvoir de prévision interne et externe doivent être déterminés, et l'espace chimique de l'application du modèle doit être limité **(Chtita 2017)**.

II.7.1. Coefficients et tests statistiques standards

Afin de déterminer la qualité d'un modèle, différents paramètres statistiques sont employés, tels que les erreurs quadratiques moyennes (Mean Square Errors), les coefficients de corrélation qui sont régulièrement utilisés dans les études RQSA/RQSP, sont décrits en détail dans cette partie **(Chtita 2017)**.

II.7.1.1 Coefficient de corrélation (r) et de détermination (r^2)

C'est l'indicateur statistique le plus répandu est le coefficient de corrélation qui évalue la part de la variance de l'activité / la propriété cible expliquée par le modèle

$$\mathbf{r} = \sqrt{\mathbf{1} - \frac{\sum(\mathbf{y_i} - \hat{\mathbf{y}})^2}{\sum(\mathbf{y_i} - \bar{\mathbf{y}})^2}} \qquad \textbf{(II.02)}$$

II.7.1.2. Le coefficient de détermination ajusté (r^2 adj)

Ce coefficient est utilisé en régression multiple par ce qu'il tient compte du degré de liberté **(Chtita 2017)** :

$$r^2_{adj} = \sqrt{\frac{r^2(n-1)-p}{n-P-1}} \qquad \textbf{(II.03)}$$

II.7.1.3. L'erreur quadratique moyenne (MSE) et l'erreur type résiduel (S)

$$MSE = \frac{\sum|(\hat{y}_1 - y_i)^2|}{n} \quad \textbf{(II.04)}$$

$$S = \sqrt{\frac{\sum(\hat{y}_1 - y_i^2)^2}{n-p-1}} \quad \textbf{(II.05)}$$

Ou y_i et $\hat{y}_i$ sont, respectivement, les valeurs observées et calculées de la variable dépendante, n est le nombre des observations et p le nombre de variables indépendantes **(Chtita 2017)**.

II.7.1.4. Facteur d'inflation de la variance (VIF)

C'est un paramètre qui permet de détecter la colinéarité entre les descripteurs utilisés dans un modèle statistique, Il est défini par :

$$VIF = \frac{1}{1-r_i^2} \quad \textbf{(II.06)}$$

Ou r_i^2 est le coefficient de détermination de la régression de la variable Xi sur les autres variables. Avec Xi est linéairement proche des autres variables, r_i^2 est proche de **1** et le ***VIF*** est grand.

L'avantage du *VIF* par rapport à la matrice de corrélation est qu'il prend en compte des corrélations multiples **(Chtita 2017)**.

II.7.1.5. Test de Fisher (F)

L'indice de Fisher F-test est employé afin de mesurer le niveau de signifiance statistique du modèle à "x%" (le niveau usuel est 95%), c'est-à-dire la qualité du choix du jeu de paramètres **(Chtita 2017)**.

La conclusion obtenue ne doit pas nous faire penser que la corrélation a "x%" de chances d'être vraie mais seulement que la corrélation est vérifiée pour « x% » des composés pris pour référence et qu'une abstraction est faite pour les autres.

Hypothèses :

- **H_0** : les variances des échantillons sont homogènes
- **H_1** : les variances des échantillons ne sont pas homogènes

La valeur F (observé) est calculée à partir la formule suivante :

$$\boldsymbol{F(observ\acute{e})} = \frac{\sum(\hat{y}_1 - y)^2 n - P - 1}{\sum(y_i - \hat{y}_1)^2 p} \quad \textbf{(II.07)}$$

Ou $\boldsymbol{F}$ est l'indice de Fisher, $\hat{y}_i$ et $\hat{y}_i^2$ sont, respectivement, les valeurs observées et calculées de la variable dépendante ;$\overline{y}$ est la valeur moyenne des valeurs prédites, $\boldsymbol{n}$ est le nombre des observations (les molécules) et $\boldsymbol{p}$ est le nombre de variables indépendantes (les descripteurs) **(Chtita 2017)**.

Après le calcul de **F** (observé) on le compare avec le **F** théorique obtenu à partir des tables statistiques usuelles (la table de Fisher). Si **F** observé est plus grand que le F théorique : refus de l'hypothèse nulle ***H0*** et cela signifie que les variances des échantillons sont trop différentes pour être considérées comme homogènes. Autrement, si **F** observé est plus petit que le **F** théorique : acceptation de l'hypothèse nulle ***H1*** et cela signifie que les deux variances ont des valeurs suffisamment proches pour qu'on accepte l'idée qu'elles soient homogènes **(Chtita 2017)**.

II.8. Conclusion

La reconnaissance de l'incertitude dans la situation de décision et de l'existence de multiples points de vue légitimes, expertises scientifiques comprises. Ces différents points de vue méritent d'être pris en compte dans un processus itératif de compréhension, de confrontation d'analyse, et modélisation.

Dans ce chapitre, comme indique son titre "Techniques de modélisation neuronal et méthode de RQSP/RQSA" nous avons abordé à la modélisation en général, leurs objectifs, ces diverses techniques d'analyse. Dans celui-ci, on va essayer de présenter un aperçu générale la modélisation avec le neurone artificiel car ces neurones sont des éléments de base de chaque réseau de neurones artificiels. Ensuite la modélisation avec les approches de RQSP/RQSA qui ont une place primordiale dans notre champ d'étude, là où La modélisation QSAR est un outil utile pour prédire l'activité biologique, les propriétés physicochimiques et les réponses toxicologiques composés chimiques non testés, par conséquent ont généralisé le phénomène, réduit le coup, et minimiser le temps de l'expérience.

Chapitre III

Modélisation d'un paramètre pertinent des phénomènes de transport lié à la pollution atmosphérique et/ou la pollution des rejets industriels

III.1. Introduction

Le coefficient de diffusion $\mathbf{D_{AB}}$ est un facteur d'influence important dans l'ingénierie environnementale. En outre, il est l'un des principaux paramètres qui jouent un rôle important dans le phénomène de transport des particules polluantes, à la fois dans l'air et dans l'eau.

Ce chapitre est consacré sur l'ensemble des matériels et des méthodes utilisées pendant la modélisation neuronale du coefficient de diffusion $\mathbf{D_{AB}}$ dans cinq systèmes expérimentales différents, afin de développée cinq modèles de type QSPR-RNA. Nous donnerons tous d'abord les résultats préliminaires ensuite les résultats finales issue de ces modèles, l'ensemble de donnés sont soumis à des études statistiques afin de tester l'exactitude de chaque modèle, et en fin nous discuterons ces résultats pour arbitrer et valider nos modèles. Ce chapitre est devisé en trois parties principales :

- Nous abordons en première partie les matériels et méthodes utilisés, notamment les logiciels informatiques utilisées, la description de la base des données ; sa collection et sa préparation, le calcul des descripteurs moléculaire, et la procédure suivez pour sélectionner juste les descripteurs pertinents, car elle revêt une grande importance dans la détermination de la qualité des modèles qui serons obtenus. Également nous expliquerons brièvement la méthodologie utilisée pour le développement de ces modèles, et la manière de choix de leurs paramètres.
- En suite le développement des modèles de type RQSP-RNA, nous donnera l'évolution des nombres des descripteurs au cours le prétraitement, les architectures optimales des modèles obtenus, nous ferons aussi des analyse tell que le calcul des erreurs et l'analyse de sensibilité qui mesure la contribution des variables d'entrée sur les propriétés à modéliser. Finalement, la partie qui suit est consacrée à la discussion des résultats significatifs obtenus.
- Dans la dernière partie de ce chapitre, nous présenterons une conclusion qui résume brièvement le travail et se tient sur les étapes les plus importantes dans ce travail.

III.2. Matériels et méthodes

III.2.1. Outils informatiques utilisés

Divers logiciels et outils informatiques ont été utilisée dans ce travail, tels qu'ils sont illustrés dans la **Figure III.01**, et dans les parties ci-dessous :

- Nous avons utilisé le logiciel **"Microsoft Office Excel 2016"** pour organiser les bases des données et dans quelques étapes de prétraitement.
- L'extraction des notifications **SMILES** ont été faite par le logiciel **"ChemDraw Office version 12.0.2" (ChemOffice 2012)**.
- Concernant les calculs des descripteurs moléculaires et l'optimisation des molécules, on a utilisé le logiciel en ligne **"PaDEL Descriptor"** intégré sur la **platform "ChemDes" (Dong *et al*. 2015)**.
- L'apprentissage des modèles neuronales et certaines étapes de réduction des descripteurs, ont été réalisés par le logiciel **"STATISTICA Versions 10" (StatSoft 2011)**.
- Nous avons utilisé le logiciel **"MATLAB R2013b" (The MathworksInc 2013)**, pour tracer les graphes et calculé les erreurs des modèles développées.

Il convient de noter que tous les calculs d'exécution de ce travail ont été effectués à l'aide des ordinateur personnelles :

(1) ACER : doté d'un processeur Intel(R) Core(TM) i5-3120M à 2,50 GHz de CPU et de 4 Go de mémoire vive (RAM), avec Windows 8.1 Professional comme système d'exploitation.

(2) TOSHIBA : doté d'un processeur Intel(R) Core(TM) i3-3120M à 2,50 GHz de CPU et de 4 Go de mémoire vive (RAM), avec Windows 7 Professional comme système d'exploitation.

(3) DELL : doté d'un processeur Intel(R) Core(TM) i3-3120M à 2,50 GHz de CPU et de 2 Go de mémoire vive (RAM), avec Windows 7 Professional comme système d'exploitation.



Figure III.01. Capture d'écran des logiciels utilisés dans cette étude : ChemDraw3D, MATLAB, STATISTICA, Microsoft Office Excel, et PaDEL Descriptor.

III.2.2. Méthodologie générale de la modélisation

La méthodologie adoptée dans ce travail est distinguée en deux grandes catégories principales, la première est concerné à la préparation des données, car le développement d'un modèle commence par la recherche des meilleures données expérimentales en terme de qualité et de quantités, cette étape est comprend la collecte et le nettoyage de données, le calcul des descripteurs moléculaires pour les différents polluants qui permettent par la suite d'effectuer des prédictions sans avoir à synthétiser les molécules, et le prétraitement de ces derniers. La deuxième est celle de la génération du modèle, qui permet l'établissement des relations statistiquement significatives entre les propriétés ciblées dans notre étude sont les coefficients de diffusion de $\mathbf{D_{AB}}$, et les structures chimiques des composés qui sont représentées par les descripteurs moléculaires. La méthodologie générale de développement suivez dans ce travail et résumé dans la **Figure III.02**.

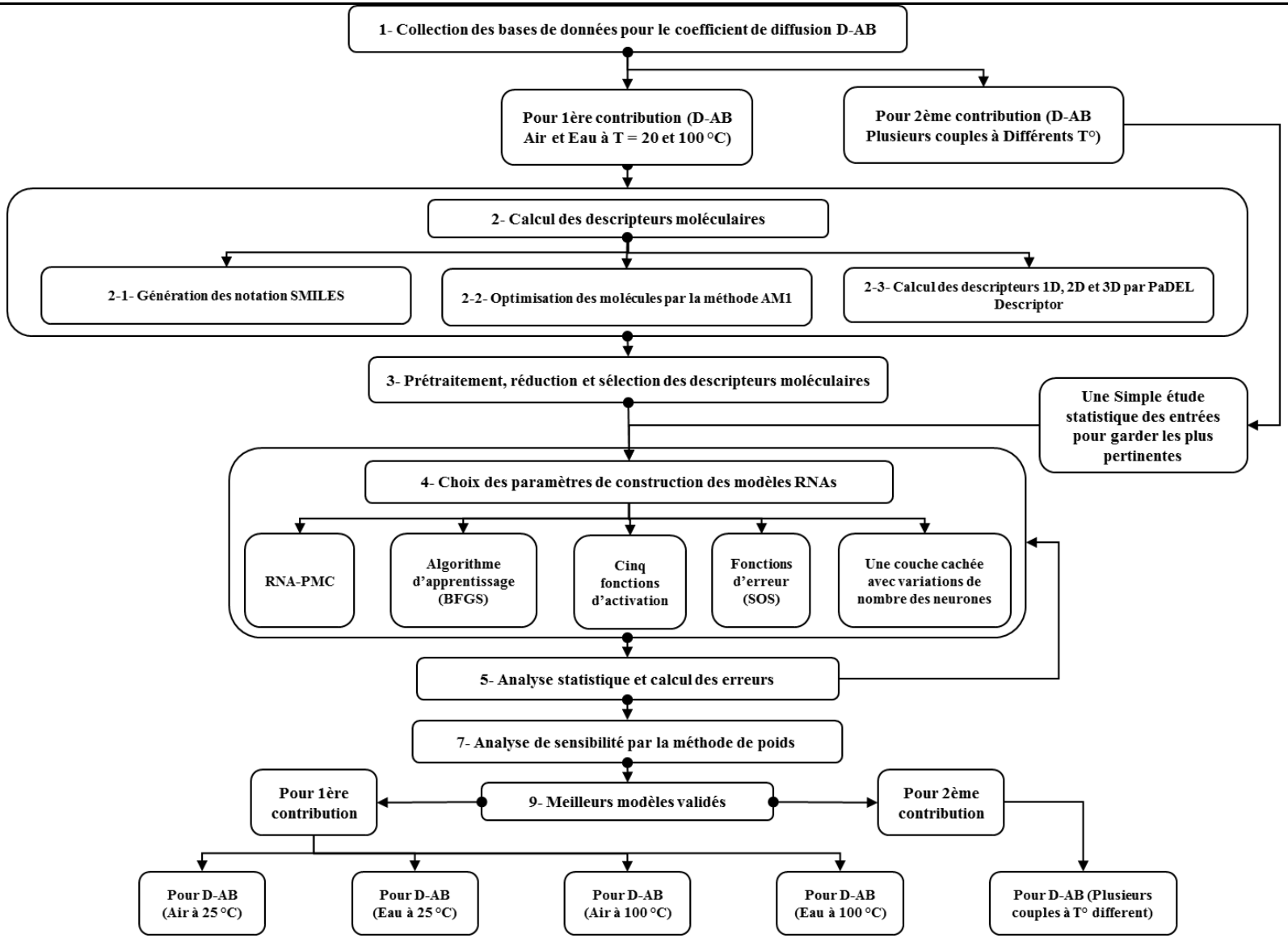


Figure III.02. Organigramme générale de la méthodologie de développement suivez dans ce travail.

III.2.2.1. Collection et préparation de la base des données

Elle est connue que la qualité de la base de données expérimentale est la pierre angulaire de tous bons modèles mathématiques. Cependant, la collection de ces données doit être bien choisir pour l'obtention des résultats satisfaisants. Selon le type et les sources de la base des données collectée dans cette étude, nous avons effectués deux contributions différentes.

La première base de données collectée à partir de la littérature comprend un nombre total de **148** molécules différents **(Yaws & Gabbula 2003)**, afin de prédire le coefficient de diffusion dans quatre systèmes différents : le coefficient de diffusion **(D_{AB})** dans l'**Air** et dans l'**Eau** exprimer en **(cm^2/s)** à deux systèmes de température **(T)** différentes exprimer en **(°C)** ; à **25°C** et à **100°C**, ces molécules sont des polluants de l'aire et de l'eau. Le **tableau III.01** donne le champ maximale et minimale des données utilisée pour chaque paramètre ainsi que leurs valeurs moyennes. De plus, une partie aléatoire de ces données est présentée dans le **tableau III.02**.

Tableau III.01. Description quantitatives et qualitatives de la première base des données.

Paramètre	Valeur minimale	Valeur moyenne	Valeur maximale	Nombre des molécules
DAB de l'eau a 25(°C)	0,000003109	8,479E-06	0,00001816	148
DAB de l'air a 25(°C)	0,1878	0,077184667	0,0294	
DAB de l'eau a 100(°C)	0,00000365	4,519E-05	0,00008997	
DAB de l'air a 100(°C)	0,0449	0,119600667	0,2787	

Tableau III.02. Partie d'une 20 molécule de la première base des données utilisées dans ce travail **(Yaws & Gabbula 2003)**.

ID	IUPAC des molécules	Formules	D_{AB} Eau à T=25°C	D_{AB} Air à T=25°C	D_{AB} Eau à T=100°c	D_{AB} Air à T=100°C
01	**acenaphthene**	**$C_{12}H_{10}$**	0,000006875	0,0578	0,00003617	0,0927
09	**aniline**	**C_6H_7N**	0,0000105	0,0735	0,00004334	0,1199
14	**1-chloro-1-fluoroethane**	**C_2H_4ClF**	0,00001306	0,1151	0,00005343	0,1726

17	**chlorobenzene**	C_6H_5Cl	0,000008974	0,0747	0,00004645	0,1196
21	**cumene**	C_9H_{12}	0,000007245	0,0677	0,000038	0,1061
32	**eicosane**	$C_{20}H_{42}$	0,000003538	0,0351	0,00001939	0,0554
36	**ethylchloroformate**	$C_3H_5ClO_2$	0,000009338	0,0828	0,00004821	0,1296
27	**dichlorodifluoromethane**	CCl_2F_2	0,00001109	01054	0,00005665	0,1595
54	**hexachloro-1,3-butadiene**	C_4Cl_6	0,000006356	0,603	0,0000336	0,0929
60	**1-hexene**	C_6H_{12}	0,000008149	0,0809	0,00004243	0,1254
66	**indene**	C_9H_8	0,00000797	0,0682	0,00004155	0,109
76	**4-methylbiphenyl**	$C_{13}H_{12}$	0,000006793	0,0607	0,0000278	0,0922
81	**methylchloride**	CH_3Cl	0,0000149	0,1295	0,00007544	0,2025
105	**methylpropyl sulfide**	$C_4H_{10}S$	0,000008811	0,0826	0,00004566	0,1287
118	**1-nitrononane**	$C_9H_{19}NO_2$	0,000006148	0,0567	0,00002516	0,0861
130	**octadecanal**	$C_{18}H_{36}O$	0,00000426	0,0403	0,00001744	0,0614
137	**nonadecylcyclopentane**	$C_{24}H_{48}$	0,00000365	0,0345	0,00000365	0,0526
140	**1-pentadecylcyclopentene**	$C_{20}H_{38}$	0,00001717	0,0398	0,00001717	0,0606
143	**1,1,1,2-tetrachloropropane**	$C_3H_4Cl_4$	0,000008473	0,0745	0,00003468	0,1226
148	**trichloroethylene**	C_2HCl_3	0,000009934	0,0875	0,00005109	0,1355

La deuxième contribution est basée sur l'étude de coefficient de diffusion $\mathbf{D_{AB}}$ de plusieurs couples à différentes températures. La base des données est contenue au départ **1252** points expérimentales de différents couples des polluants **A-B**, avec neuf entrées qui ont été proposer dans l'étude de Melzi et ces co-auteurs **(Melzi *et al*. 2019)**, mais après traitement et choix des couples qui ont une pression égale à **1 atm** et une simple étude statistique des entrées, il reste envirent 727 points expérimentales de différents couples des polluants **A-B** avec sept entrées, afin de prédire leurs coefficients de diffusion $\mathbf{D_{AB}}$ exprimer en **(cm^2/s)**. Le **tableau**

III.03 donne les champs descriptifs des données utilisée pour chaque entrée et la sortie étudié. Encore, une partie aléatoire de ces données est présentée dans le **tableau III.04**.

Tableau III.03. Description quantitatives et qualitatives de la deuxième base des données.

Paramètre	Valeur minimale	Valeur moyenne	Valeur maximale	Nombre des molécules
T(°k)	77,2	395,968487	1080	727
Tc_A [K]	5,19	223,078487	761	
Tc_B [K]	5,19	274,852545	819	
Vc_A [Cm^3/mol]	41,7	128,385846	431,9	
Vc_B [Cm^3/mol]	41,7	146,35414	492,1	
M_A [g/mol]	4,003	39,1483968	235,9	
M_B [g/mol]	2,016	53,5337879	253,80894	
D_AB [Cm^2/S]	0,000045	0,66506083	8,19	

Tableau III.04. Partie d'une 20 molécule de la deuxième base des données utilisées dans ce travail (**Melzi *et al*. 2019**).

ID	Couples	T [°K]	Tc_A [K]	Tc_B [K]	Vc_A [Cm^3/mol]	Vc_B [Cm^3/mol]	M_A [g/mol]	M_B [g/mol]	D_AB [Cm^2/S]
01	**CH4 –H2**	288	190,53	33,18	99,3	64,2	16,043	2,016	0,594
05	**CH3OH–H2**	353	512,58	33,18	117,8	64,2	32,04	2,016	0,937
18	**Kr-H2**	720	209,35	33,18	91,2	64,2	83,8	2,016	3,25
26	**C3H6O-H2**	296	508,2	33,18	209	64,2	58,079	2,016	0,43
31	**He-H2**	118	5,19	33,25	57,3	64,2	4,003	2,016	0,318
37	**He-n-C4H10**	298	5,19	425,18	57,3	255	4,003	58,12	0,364
43	**He-n-C5H12**	372,6	5,19	469,7	57,3	312,3	4,003	72,15	0,422
5	**He-SF6**	255	5,19	318,19	57,3	198,4	28,014	146,05	0,388

65	**He-fluorobenzene**	429,6	5,19	560,09	57,3	269	4,003	96,1	0,566
72	**He-2-chlorobutane**	429,1	5,19	520	57,3	300	4,003	92,56	0,561
82	**He-2-Bromohexane**	427,9	5,19	612	57,3	427,5	4,003	165,07	0,47
94	**N2-n-C4H10**	378,2	126,2	425,18	90,1	255	28,014	58,123	0,159
110	**N2-1Octanol**	473	126,2	652,5	90,1	490	28,014	102,177	0,145
126	**N2-Xe**	263	126,2	430,8	90,1	122	28,014	64,065	0,104
134	**Ar-ether**	298	150,86	400	74,57	170	39,948	46,068	0,0849
149	**CHCl3 –CO2**	363,3	536,4	304,12	239	94,07	119,378	44,01	0,11
166	**CO-Kr**	820	132,92	209,4	93,1	91,2	28,01	83,8	0,879
195	**N20-H2**	273,2	431,35	33,25	82,5	64,2	44,013	2,016	0,535
204	**CS2-H 2**	263	552	33,25	161,5	64,2	76,139	2,016	0,0874
208	**SF6 -CO2**	289	318,19	154,58	198,4	73,37	146,05	31,999	0,0988

III.2.2.2. Méthode de génération des descripteurs moléculaires

L'utilisation des descripteurs moléculaires dans le développement des modèles RQSA/RQSP a permet de quantifier les structures chimiques des molécules étudier et les représenter numériquement, dans cette étude trois étapes ont été suivez pour le calcul de ces descripteurs :

1) La génération des notations SMILES qui a été obtenir à partir de logiciel ChemDraw, permet de coder les molécules et passer du langage chimique vers un langage informatique. Ultérieurement, chaque notation SMILES a été enregistrer dans un fichier **".txt"**, et désigner par leur nom IUPAC et leur N° d'ordre dans la base de donnes, comme elle indique la **Figure III.03**.
2) Avant le calcul des différents descripteurs on a passé tout d'abord par l'étape d'optimisation des molécules, la sélection des descripteurs optimaux a été effectué par l'identification des meilleurs descripteurs qui est une tâche importante car elle permet d'augmenter la précision des valeurs des descripteurs

calculées surtout ceux de 3D. De ce fait, on a choisi la méthode **AM1** (en anglais : *Austin Model 1*) qui est une méthode semi-empirique très connue pour le calcul quantique de la structure électronique moléculaire en chimie computationnelle **(Stewart 2004)**. L'optimisation se faite molécule par molécule à l'aide d'un logiciel "ChemMOP" intégré dans la plateforme "ChemDes" **(Dong *et al*. 2015)**. Les notation SMILES inclus dans les fichiers **".txt"** préparer préalablement sont utilisées comme des entrés et des fichiers **".sdf"** sont obtenue à la sortie de l'optimisation pour chaque molécule.

3) Finalement, les molécules optimisées précédemment ont été utilisées comme des entrées pour calculer des centaines de descripteurs moléculaire à l'aide le logiciel en ligne "PaDEL Descriptor" **(Dong *et al*. 2015)**. **1970** descripteurs ont été calculés de type *1D*, *2D* et *3D*, de sort que **1538** descripteurs appartiennent à des catégories ***1D_2D*** et les **432** restantes sont classées comme des descripteurs ***3D***.

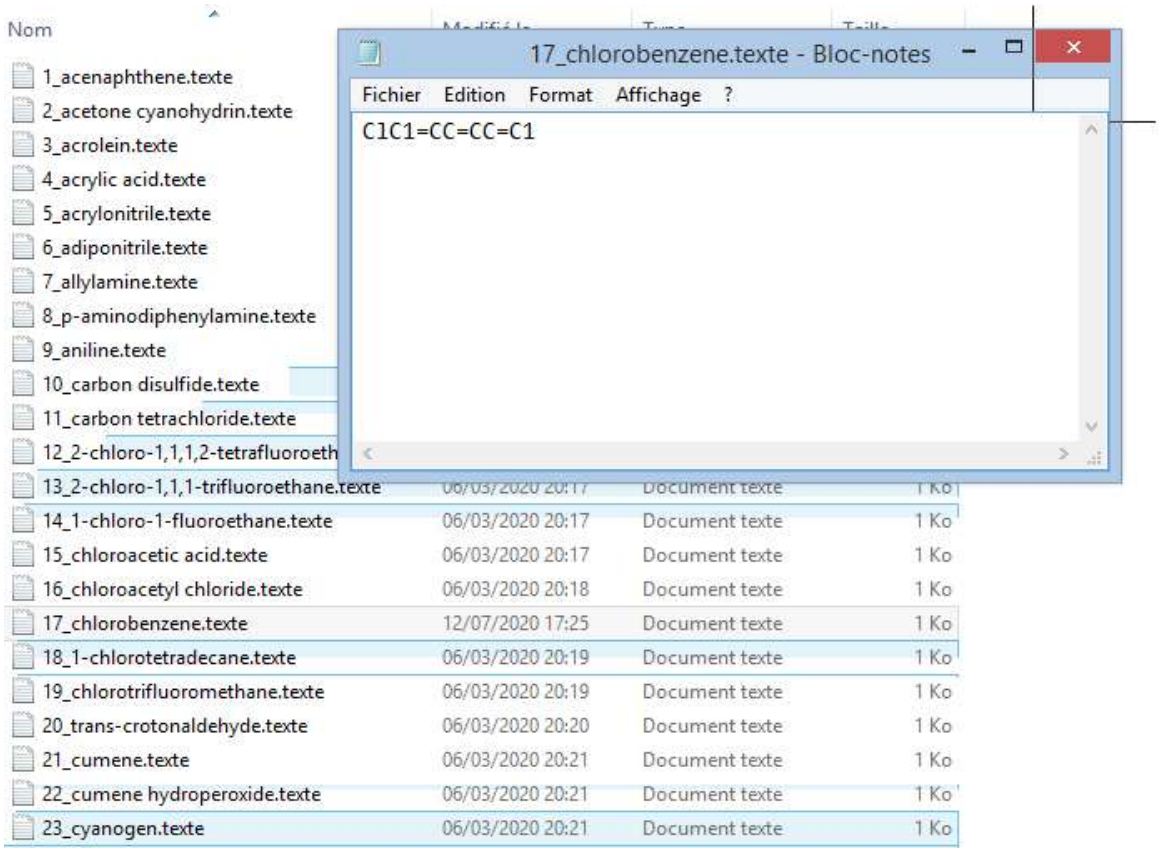


Figure III.03. Capture d'écran du fichier texte du molécule chlorobenzène.

Le **tableau III.05**, montre les caractéristiques des 1875 descripteurs moléculaires générées par l'ancien version de logiciel en ligne "PaDEL Descriptor" qui divisées en 16 catégories distinctes.

Tableau III.05. Caractéristiques des descripteurs générées par l'ancien version de logiciel en ligne "PaDEL Descriptor" **(Dong *et al*. 2015)**.

ID	Catégories des Descripteurs	Nombre	Dimensions
01	Autocorrelationdescriptors	346	2D
02	Basakdescriptors	42	2D
03	BCUT descriptors	6	2D
04	Burdendescriptors	96	2D
05	Connectivity descriptors	56	2D
06	Constitutionaldescriptors	120	1D
07	E-state descriptors	489	2D
08	Kappa descriptors	3	2D
09	Molecularpropertydescriptors	15	2D
10	Quantum chemicaldescriptors	6	2D
11	Topologicaldescriptors	265	2D
12	CPSA descriptors	29	3D
13	RDF descriptors	210	3D
14	Geometricaldescriptors	21	3D
15	WHIM descriptors	91	3D
16	3D Autocorrelationdescriptors	80	3D

NB : **(1)** Nous avons gardé les noms des catégories en anglais. **(2)** Les informations dans ce tableau sont inclus 1875 descripteurs, mais en réalité 1970 descripteurs ont été calculées par la nouvelle version de logiciel. **(3)** Pour plus d'informations sur les descripteurs visite le lien : (*http://www.scbdd.com/padel_desc/descriptors/*).

Après le calcul des descripteurs, la première base de données est enrichie, leur taille est devenue plus large, de sort qu'il tient une matrice de 148 lignes et 1970 colonnes qui représentes respectivement le nombre des molécules utilisés et les descripteurs initiale calculés. Ces derniers sont considérés comme des entré du modèle, mais il est bien clair que 1970 entrées est un nombre énorme pour des variables de n'importe type de modèle. De ce fait, elle est nécessaire de réduite ce nombre à un nombre raisonnables et de sélectionne juste les plus pertinentes.

III.2.2.3. Procédure de prétraitement et sélection des descripteurs

La tâche la plus essentiel dans cette étude est la sélection des entrées les plus adéquats possible pour chaque paramètre $\mathbf{D_{AB}}$ à étudier. Le prétraitement de la base de données vise à réduire le nombre de descripteurs non pertinents. L'idée principale est de réduire les dimensions des variables d'entrées du modèle, et elle repose sur l'élimination des entrées qui possèdent une faible influence sur la sortie à modéliser.

Après le recueil des descripteurs, la démarche statistique consiste à traiter et réduire le nombre de descripteurs, afin de les utilisent ensuite dans la construction des modèles RQSP-RNA. La méthode était basée sur les points suivants :

(1) Nettoyer tous les descripteurs avec une valeur d'erreur indiquée par une case vide après le calcule par le logiciel en ligne "PaDEL Descriptor", en utilise le logiciel Excel dans cette étape.

(2) Supprimer les descripteurs qui possèdent la même valeur répétée pour plus de 75%, en comparant les trois premiers quartiles.

(3) Calculer la déviation standard relative (RSD) pour chaque descripteur puis éliminé ceux qui ayant des valeurs inférieures à 0,05. Le logiciel "STATISTICA" a été utilisé pour réaliser les deux dernières étapes.

(4) Calculer puis supprimer tous les descripteurs avec un coefficient d'intercorrélation (R) supérieur à 0,75 pour chaque paire de descripteurs, afin de ne retenir que les descripteurs pertinents. Les calculs des coefficients (R) ont été faits par le logiciel "Matlab", la comparaison et l'élimination a été faite manuellement par le tableur Excel 2016.

(5) Finalement, une méthode de régression linéaire multiple "RLM" pas à pas (en anglais : *MLR forward step wise regression*) a été utilisé pour supprimer les descripteurs dont la valeur de probabilité est supérieure à 0,05 ; en utilisant constamment le logiciel "STATISTICA".

Pour la deuxième contribution qui s'intéresse par l'étude de $\mathbf{D_{AB}}$ de plusieurs couples à différentes températures, une simple étude statistique des entrées a été fait afin de garder que les descripteurs les plus pertinentes. De ce fait, tous les points expérimentaux qui possèdent les mêmes valeurs répétées à 100% ont été supprimées.

III.2.2.4. Choix des paramètres pour le développent des modèles de type RQSP-RNA

Après le prétraitement, et la réduction des nombres de descripteurs autant que possible, en gardant les descripteurs pertinents pour chaque paramètre, afin de les utilisées après dans la

phase de construction des modèles de RQSP-RNA. Ceci est un autre défi concernant l'existence de large paramètre de choix dans le RNA. Pour cela, diverses options ont été utilisées afin d'obtenir les meilleurs modèles de type RQSP-RNA pour chaque paramètre.

Le choix des paramètres de RNA se fait à base des études précédente **(Roubehie fissa *et al*. 2019)** et par apport le logiciel utilisé "STATISTICA", donc on va suivre ces étapes :

- L'ensemble de données dans la première contribution **($D_{AB\text{-}air_eau}$ à T° = 25** et **100 °C)** a été divisé au hasard en deux sous-ensembles : (80%) pour la phase d'apprentissage et (20%) pour la phase de test (voir la **Figure III.04**). Tandis que dans les données dans la deuxième contribution **($D_{AB\text{-}Couples\ des\ Polluants}$** à plusieurs couples et à différents **T°)** a été divisé également au hasard en trois sous-ensembles : (70%) pour la phase d'apprentissage, (15%) pour la phase de test, et (15%) pour la phase de validation (voir la **Figure III.05**).

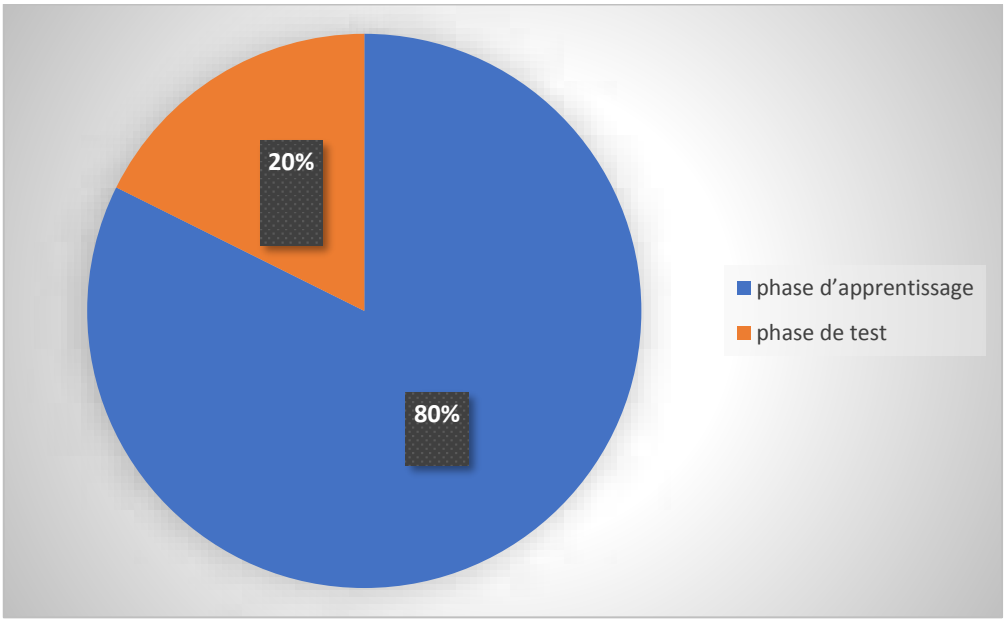


Figure III.04. Représentation graphique de division de la première base des données.

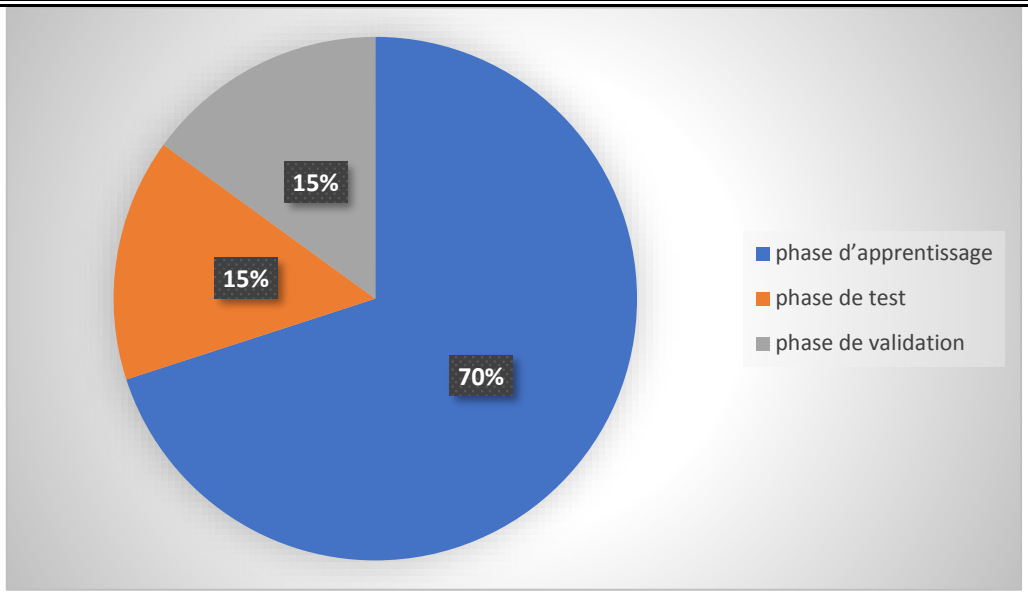


Figure III.05. Représentation graphique de division de la deuxième base des données.

- Le type de RNA utilisé dans ce travail est un perceptron multicouche (PMC), car il est l'un des réseaux de neurones les plus utilisés pour des problèmes d'approximation et de prédiction.
- L'algorithme Quasi Newton à propagation en arrière (Quasi Newton back propagation algorithm, plus précisément : Broyden-Fletcher-Goldfarb-Shanno algorithm BFGS) a été utilisé comme un algorithme d'apprentissage.
- Cinq fonctions d'activation différents sont utilisées pour chaque couche cachée et sortie. Par conséquent, à la fin de l'exécution, le logiciel donne la meilleure combinaison possible entre les deux fonctions d'activation des couches concernées **(Figure III.06)**.

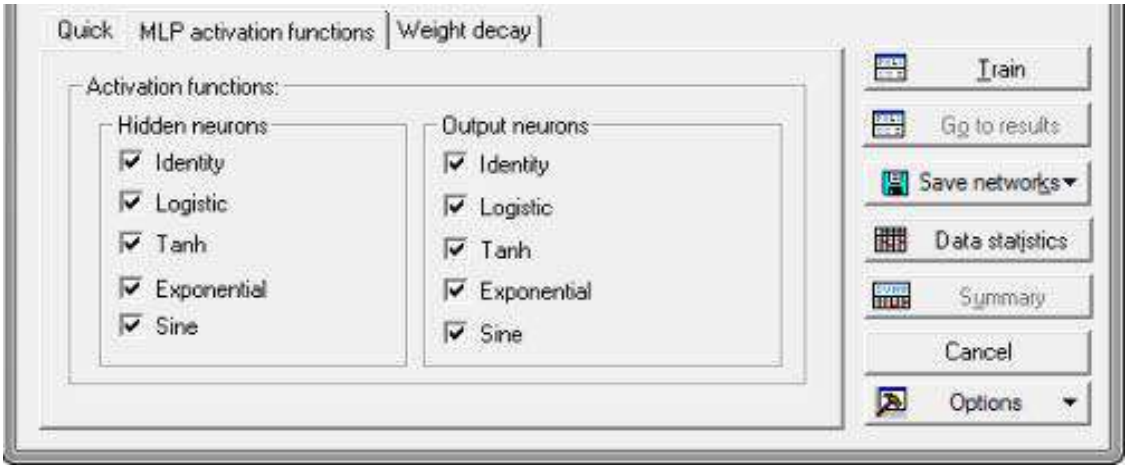


Figure III.06. Capture d'écran pour le choix des fonctions d'activation sur STATISTICA.

- La somme des carrés (Sum of Squares SOS) a été choisi en tant que fonction d'erreur.
- Une seule couche cachée est utilisée dans ce travail, avec un changement des neurones dans cette couche, afin de de garantir que les variables des modèles de RN ne dépassent pas la taille de la base des données. Selon le cas, pour les quatre modèles de la première contribution on à changer le nombre de neurones suivant cette règle **(Roubehie fissa *et al.* 2019)** : $\{(Nbr\ des\ Neurones\ \boldsymbol{Etr\acute{e}es} * Nbr\ des\ Neurones\ \boldsymbol{Cach\acute{e}es}) + (Nbr\ des\ Neurones\ \boldsymbol{Cach\acute{e}es} * Nbr\ des\ Neurones\ des\ \boldsymbol{Sorties}) \leq \boldsymbol{Taille\ de\ base\ des\ donn\acute{e}es}\}$, tandis que dans la deuxième contribution le nombre a été varié de 3 à 30 neurones.
- Plusieurs exécutions ont été effectuées pour chaque paramètre dans les deux bases des données afin d'obtenir les meilleurs modèles de type RQSP-RNA. L'exécution comporte un nombre important d'itérations égale à 10 000 itérations pour une seule exécution.

III.2.2.5. Analyses statistiques et calcules des erreurs des meilleurs modèles développées

La performance et la précision des meilleurs modèles de type RQSP-RNA obtenus dans cette étude, ont été évaluées par deux types d'erreurs qui sont : la racine d'erreur quadratique moyenne **(REQM)** et l'erreur standard de prédiction **(ESP).** Les performances des régressions ont été aussi testées à l'aide des coefficients de régression **(R)** et de détermination **(R^2)** comme des critères notamment dans l'étape de sélection des descripteurs. Ceci est fait sur l'ensemble des données d'apprentissage, de test, de validation et sur les données entières de la base des données, en plus des vecteurs de régression : les valeurs de la pente **(α),** et de l'interception **(β)** de chaque modèle qui mesurées juste pour l'ensemble des données entières.

Les définitions mathématiques des deux types d'erreurs, du coefficient de régression **(R)** et du coefficient de détermination **(R^2)** sont données par les **Eqs (III.01–III.04),** respectivement.

$$\boldsymbol{REQM = \sqrt{\sum_{i=1}^{N}\left(y_i^{exp} - y_i^{cal}\right)^2 / N}} \qquad \textbf{(III.01)}$$

$$ESP = (REQM/\hat{y}^{exp}) * 100 \quad \textbf{(III.02)}$$

$$R = \frac{\sum_i^N [(y_i^{exp} - \hat{y}^{exp})(y_i^{cal} - \hat{y}^{cal})]}{\sqrt{\sum_{i=1}^N (y_i^{exp} - \hat{y}^{exp})^2} \sqrt{\sum_{i=1}^N (y_i^{cal} - \hat{y}^{cal})^2}} \quad \textbf{(III.03)}$$

$$R^2 = 1 - \sum_{i=1}^N (y_i^{exp} - y_i^{cal})^2 / \sum_{i=1}^N (y_i^{exp} - \hat{y}^{exp})^2 \quad \textbf{(III.04)}$$

$$\hat{y}^{exp} = \sum_{i=1}^N y_i^{exp} / N \quad \textbf{(III.05)}$$

Où N est le nombre des composés dans l'ensemble des données pour chaque phase : (apprentissage, test, validation ou l'ensemble entière des données), y^{cal} est la valeur calculée (prédite) de y ; pour l'ensemble d'apprentissage de test ou de validation, y^{exp} est la valeur expérimentale (observée) de y, et $\hat{y}^{exp}$ la moyenne des valeurs expérimentales de y ; (**Eq (III.05)**).

III.2.2.6. Analyse de sensibilité des paramètres d'entrées pour chaque modèle développé

L'analyse de sensibilité est un outil important pour l'évaluation des modèles mathématiques. Elle est utilisée pour diverses raisons, notamment la validation du modèle, l'évaluation du comportement des différentes variables du modèle, et la confirmation du choix des variables du modèle via la détermination des contributions de chaque variable choisie sur le modèle. De ce fait, il est donc important d'utiliser cette analyse sur les modèles de type RQSP-RNA.

La méthode utilisée dans cette étude (méthode de Poids) ; a été organisée et programmée dans un tableur "Microsoft Excel", en suivant les étapes théoriques indiquées dans l'étude de Gevrey et ses co-auteurs (**Gevrey *et al.* 2003**). Comme l'indique de nom de la méthode "Poids", on introduit les poids de connexion de partitionnement en : **(1)** poids de coude d'entrée (w_{ij}^I) (connexion entre les neurones d'entrée et cachés), **(2)** poids de la couche cachée (w_{ij}^H) (connexion entre les neurones d'entrée et sortie). Pour trouver à la fin une signification

quantifiée de chaque descripteur d'entrée sur chaque sortie du modèle, ce qui se traduit par l'importance relative **(IR)** de chaque descripteur.

Pour mieux comprendre cette méthode nous digérera vers l'exemple utilisé dans l'étude de Goh **(Goh 1995)**, et l'étude de Gevrey et ses co-auteurs **(Gevrey *et al*. 2003)**. La procédure de calcul de cette méthode, est aussi schématisée sur l'organigramme dans la **figure III.07 (Khaouane 2013)**.

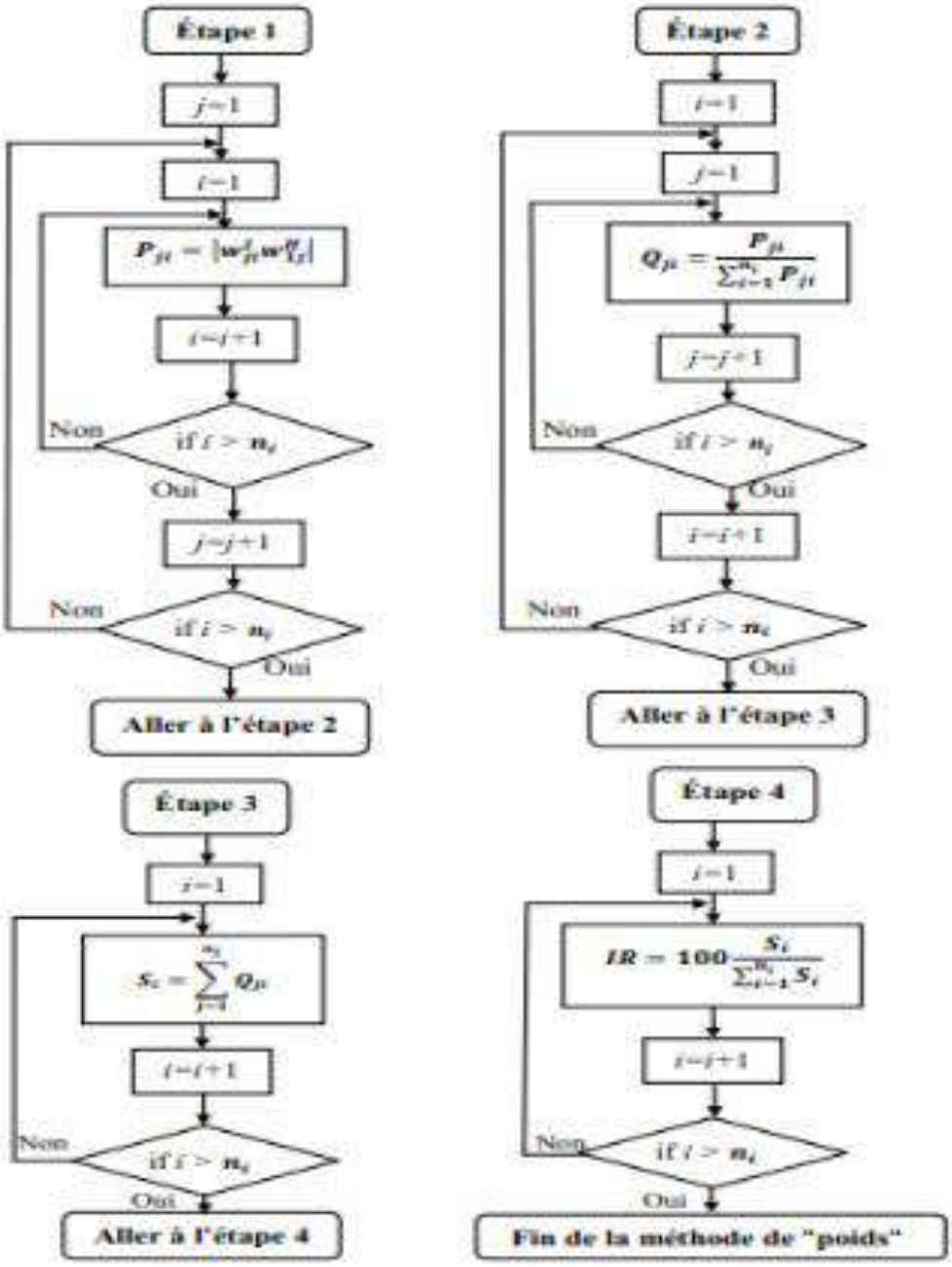


Figure III.07. Organigramme illustrant l'analyse de sensibilité par la méthode de la partition des poids (méthode de poids) **(Khaouane 2013)**.

Dans cette étude l'analyse de sensibilité par la méthode de "Poids" (*Weights method*) a été utilisée pour déterminer l'importance relative ou bien la contribution des différents descripteurs d'entrées pour évaluées leurs impacts sur les sorties **D_{AB}** des quatre modèles développés dans le cas de première contribution. Cette analyse est également importante pour confirmer notre choix des descripteurs utilisés dans le développement des modèles de type RQSP-RNA, et donc l'efficacité de la méthode de prétraitement adaptée.

III.3. Résultats et discussions

III.3.1. Évolution du nombre de descripteurs au cours de la procédure de sélection

Comme décrit dans la partie **(III.2.2.3)** de ce chapitre, la procédure de sélection utilisée dans cette étude repose sur une méthodologie statistique contenant plusieurs étapes différentes, chaque étape ayant ses propres spécifications. En conséquence, chaque étape a son propre effet sur le résultat du prétraitement et a donc un impact sur le nombre de descripteurs à choisir. L'évolution du nombre de descripteurs lors du prétraitement pour chaque paramètre étudié est donnée dans le **tableau III.06**.

Tableau III.06. Évolution du nombre de descripteurs après chaque étape de prétraitement pour les quatre **D_{AB}** étudiés de la première contribution.

Etapes de Réduction	***(0)***	***(1)***	***(2) et (3)***	***(4)***				***(5)***			
				D_{AB_Eau} (25°C)	**D_{AB_Air} (25°C)**	**D_{AB_Eau} (100°C)**	**D_{AB_Air} (100°C)**	**D_{AB_Eau} (25°C)**	**D_{AB_Air} (25°C)**	**D_{AB_Eau} (100°C)**	**D_{AB_Air} (100°C)**
Nombre des descripteurs	1970	1143	835	178	177	170	179	11	20	9	16

***NB :** Étape **(0)** indique le nombre des descripteurs initiale calculées. Étapes **(1–5)** sont indiquées dans la partie **(III.2.2.3)** de ce chapitre.*

Après la dernière étape de prétraitement, on obtient que les descripteurs pertinents pour chaque paramètre qui sont : **11** descripteurs pour **D_{AB}** de l'eau à **25°C**, **20** descripteurs pour **D_{AB}** de l'air à **25°C**, **09** descripteurs pour **D_{AB}** de l'eau à **100°C**, et **16** descripteurs pour **D_{AB}** de l'air à **100°C**. Les valeurs **P** de chaque descripteur et leurs désignations sont indiquées dans le **tableau III.07**.

Tableau III.07. Caractéristique des descripteurs retenus pour **D_{AB}** des quatre modèles développés dans la première contribution.

Paramètres	**D_{AB} Eau (25°C)**		**D_{AB} Air (25°C)**		**D_{AB} Eau (100°C)**		**D_{AB} Air (100°C)**	
	Nom_Des	**Valeur_P**	**Nom_Des**	**Valeur_P**	**Nom_Des**	**Valeur_P**	**Nom_Des**	**Valeur_P**
Entrée 01	***VAdjMat***	$1{,}40E^{-45}$	***SRW2***	$6{,}48E^{-15}$	***VAdjMa***	$0{,}00E^{+00}$	***VAdjMat***	$2{,}43E^{-29}$

Entrée 02	***AATS2i***	$3{,}32E^{-12}$	***PetitjeanNumber***	$7{,}28E^{-06}$	***AATS2e***	$2{,}00E^{-01}$	***PetitjeanNumber***	$1{,}08E^{-05}$
Entrée 03	***GATS1c***	$4{,}24E^{-05}$	***ATSC1v***	$7{,}13E^{-07}$	***JGI3***	$2{,}25E^{-04}$	***AATSC2e***	$4{,}63E^{-03}$
Entrée 04	***GATS2p***	$7{,}56E^{-07}$	***ATSC7i***	$1{,}60E^{-03}$	***nO***	$2{,}36E^{-08}$	***nBondsD2***	$5{,}14E^{-04}$
Entrée 05	***AATSC1i***	$3{,}95E^{-05}$	***Ku***	$9{,}23E^{-05}$	***RPCG***	$9{,}26E^{-07}$	***ATSC7i***	$1{,}42E^{-03}$
Entrée 06	***Du***	$6{,}35E^{-04}$	***khs.sssCH***	$2{,}02E^{-04}$	***AATSC0p***	$1{,}89E^{-05}$	***TDB2s***	$2{,}89E^{-11}$
Entrée 07	***FPSA-3***	$5{,}64E^{-09}$	***MATS2s***	$8{,}05E^{-04}$	***JGI7***	$3{,}89E^{-02}$	***khs.sssCH***	$5{,}43E^{-03}$
Entrée 08	***VE3_D***	$8{,}53E^{-03}$	***E3p***	$2{,}02E^{-07}$	***Km***	$1{,}92E^{-07}$	***GATS5m***	$1{,}76E^{-02}$
Entrée 09	***RPCS***	$2{,}39E^{-03}$	***VR3_D***	$6{,}80E^{-03}$	***RotBFrac***	$8{,}07E^{-07}$	***MATS3e***	$5{,}23E^{-03}$
Entrée 10	***JGI2***	$1{,}15E^{-03}$	***RDF45v***	$1{,}72E^{-03}$	**(-)**	**(-)**	***AATSC3p***	$1{,}12E^{-04}$
Entrée 11	***TDB2u***	$3{,}66E^{-03}$	***ATSC2i***	$4{,}33E^{-03}$	**(-)**	**(-)**	***AATSC0p***	$6{,}92E^{-05}$
Entrée 12	**(-)**	**(-)**	***AATSC3p***	$9{,}68E^{-05}$	**(-)**	**(-)**	***E2m***	$1{,}96E^{-02}$
Entrée 13	**(-)**	**(-)**	***AATSC3m***	$1{,}43E^{-03}$	**(-)**	**(-)**	***E2p***	$3{,}04E^{-02}$
Entrée 14	**(-)**	**(-)**	***AATS1i***	$3{,}54E^{-04}$	**(-)**	**(-)**	***L2s***	$3{,}90E^{-05}$
Entrée 15	**(-)**	**(-)**	***AATSC3c***	$1{,}71E^{-02}$	**(-)**	**(-)**	***MATS2p***	$3{,}86E^{-03}$
Entrée 16	**(-)**	**(-)**	***E2u***	$5{,}61E^{-03}$	**(-)**	**(-)**	***MPC6***	$1{,}35E^{-03}$
Entrée 17	**(-)**	**(-)**	***E2s***	$2{,}21E^{-03}$	**(-)**	**(-)**	**(-)**	**(-)**
Entrée 18	**(-)**	**(-)**	***nBondsD2***	$3{,}80E^{-03}$	**(-)**	**(-)**	**(-)**	**(-)**
Entrée 19	**(-)**	**(-)**	***Dv***	$5{,}37E^{-06}$	**(-)**	**(-)**	**(-)**	**(-)**
Entrée 20	**(-)**	**(-)**	***AATSC1i***	$8{,}68E^{-05}$	**(-)**	**(-)**	**(-)**	**(-)**
Totale	$\mathbf{P_{(11_Descripteurs)} < 0{,}0000}$		$\mathbf{P_{(20_Descripteurs)} < 0{,}0000}$		$\mathbf{P_{(09_Descripteurs)} < 0{,}0000}$		$\mathbf{P_{(16_Descripteurs)} < 0{,}0000}$	

D'après les résultats présentés dans le **tableau III.07**, en constate que tous les descripteurs choisis pour les quatre $\mathbf{D_{AB}}$ étudiées sont statistiquement significatifs.

III.3.2. Architectures des meilleurs modèles neuronales obtenus

Après toutes les étapes de réduction, de construction et de calcul, seuls les meilleurs modèles ont été sélectionnés pour tester leurs performances, en calculant diverses erreurs et paramètres statistiques. Les architectures des meilleurs modèles développées pour chaque paramètre avec certains paramètres statistiques primaires sont organisées dans le **tableau III.08**.

Les modèles RQSP-RNA possèdent des valeurs de SDC relativement acceptables pour tous les sous-ensembles d'apprentissage, de test et de validation, des modèles développées dans les deux contributions, en plus des valeurs de coefficient de régression **R** qui sont également montrent des performances élevées pour l'ensembles des modèles développés dans cette étude.

Tableau III.08. Architectures des meilleurs modèles obtenus pour les deux contributions.

Couches	Nombre de neurones	Fonctions d'activation	Valeur de R et SDC	
Pour $D_{AB_Couples}$ à T=Différents			R^{App}	0,9943
Entrée	7	***(Aucune)***	R^{Tes}	0,9796
Cachée	24	*Tanh*	R^{Val}	0,9921
Sortie	1	*Logistic*	SDC^{App}	0,0055
Algorithme d'App	BFGS		SDC^{Tes}	0,0120
Fonction d'erreur	SOS		SDC^{VAL}	0,0047
Pour D_{AB_Air} à T=25°C			R^{App}	0,9563
Entrée	20	***(Aucune)***	R^{Tes}	0,9790
Cachée	2	*Exponential*	(-)	(-)
Sortie	1	*Sine*	SDC^{App}	0,00003
Algorithme d'App	BFGS 45		SDC^{Tes}	0,00001
Fonction d'erreur	SOS		(-)	(-)
Pour D_{AB_Eau} à T=25°C			R^{App}	0,9535
Entrée	11	***(Aucune)***	R^{Tes}	0,9899
Cachée	2	*Exponential*	(-)	(-)
Sortie	1	*Identity*	SDC^{App}	$4,51E^{-13}$
Algorithme d'App	BFGS 29		SDC^{Tes}	$8,406E^{-14}$
Fonction d'erreur	SOS		(-)	(-)
Pour D_{AB_Air} à T=100°C			R^{App}	0,9812
Entrée	16	***(Aucune)***	R^{Tes}	0,9923
Cachée	6	*Exponential*	(-)	(-)
Sortie	1	*Exponential*	SDC^{App}	0,00003
Algorithme d'App	BFGS 147		SDC^{Tes}	0,00001
Fonction d'erreur	SOS		(-)	(-)
Pour D_{AB_Eau} à T=100°C			R^{App}	0,9534
Entrée	9	***(Aucune)***	R^{Tes}	0,9866
Cachée	10	*Tanh*	(-)	(-)
Sortie	1	*Tanh*	SDC^{App}	$1,21E^{-11}$
Algorithme d'App	BFGS 25		SDC^{Tes}	$3,04E^{-12}$
Fonction d'erreur	SOS		(-)	(-)

NB : Apprentissage **(App)**, Test **(Tes)**, Validation **(Val)**, **B**royden-**F**letcher-**G**oldfarb-**S**hanno **(BFGS)**

Comme indiqué dans le **tableau III.08**, les architectures des meilleurs modèles sont (7-24-1) pour le modèle de D_{AB} pour plusieurs couples, (20-2-1) pour le modèle de D_{AB} de l'Air à T=25°C, (11-2-1) pour le modèle de D_{AB} de l'Eau à T=25°C, (16-6-1) pour le modèle de D_{AB} de l'Air à T=100°C et (9-10-1) pour le modèle de D_{AB} de l'Eau à T=100°C. Donc, les réseaux comportent 7, 20, 16, 11 et 9 neurones entrées, 24, 2, 2, 6 et 1 neurones cachés sont apparus respectivement dans les mêmes modèles. En plus, d'une sortie **(S)** pour chaque modèle.

Pour les fonctions d'activation, nous avons trouvé en premier lieu, la fonction Exponentiel "Exponentiel" qui est semblé quatre fois. En seconde lieu nous avons trouvé la fonction tangente hyperbolique "Tanh" qui est apparu trois fois, puis nous avons trouvé les fonctions d'activation : sinus "Sine", Sigmoïde Logarithmique "Logistic" et "Identity" qui sont apparus une seul fois dans les cinq modèles développés pour cette étude. Leurs définitions mathématiques sont données dans le **tableau III.09** (Voir les **Eqs. (III.06), (III.07), (III.08), (III.09) et (III.10)**) **(StatSoft 2011 ; Beale *et al*. 2012)**.

Tableau III.09. Fonctions d'activation apparues dans cette étude **(StatSoft 2011 ; Beale *et al*. 2012)**.

Fonctions d'activation	*Exponentielle*	*Tangent Hyperbolique*	*Sinus*	*Sigmoïde Logarithmique*	*Identity*
Notation dans STATISTICA	***Exponential***	***Tanh***	***Sine***	***Logistic***	***Identity***
Notation dans MATLAB	***exp***	***tanh***	***sin***	***logsig***	***perlin***
Formules Mathématiques	$a = e^{-n}$ **(III.06)**	$a = \frac{e^{+n}-e^{-n}}{e^{+n}+e^{-n}}$ **(III.07)**	$a = sin(n)$ **(III.08)**	$a = \frac{1}{1+e^{-n}}$ **(III.09)**	$a = x(n)$ **(III.10)**

NB : (e) la fonction exponentielle ; **(n)** les variables des modèles de RNA (voir les **Eqs. (III.11) à (III.20)**)

III.3.3. Application des modèles obtenus dans la prédiction des coefficients de diffusion des polluants dans différents systèmes

Pour utiliser les modèles de type RQSP-RNA développés dans cette étude afin de prédire les cinq coefficients de diffusion $\mathbf{D_{AB}}$ dans chaque cas d'étude. Il faut tout d'abord comprendre le contenu des expressions mathématiques de chaque modèle de RQSP-RNA obtenu.

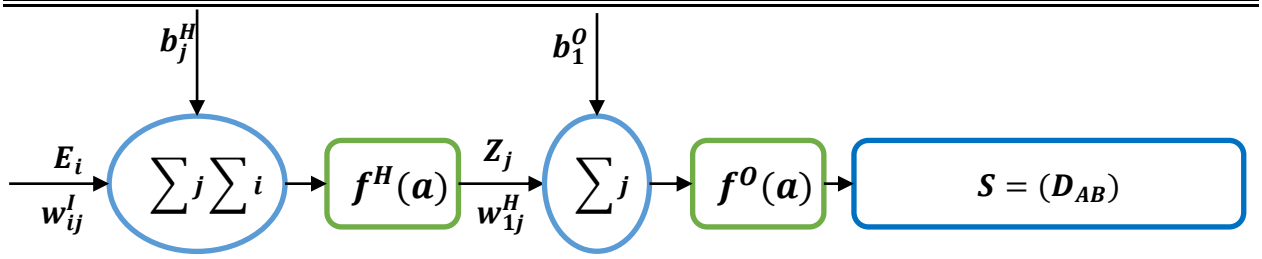


Figure III.08. Présentation schématique d'un réseau du modèle de type RQSP-RNA.

Tous les modèles sont semblables à ce qui représenté dans la **figure III.08**, la seule différence réside dans le nombre des neurones d'entrées, et celui des nombres de neurones dans la couche cachée de chaque modèle. Les expressions mathématiques des sorties (Z_j) de la couche cachée et (S) de la couche sortie de chaque modèle sont données par les équations ci-après (**III.11–III.20**) :

$$Z_j = f^H\left[\sum_{i=1}^{07} w_{ij}^I E_i + b_j^H\right] \quad \textbf{(III.11)}$$

$$S = f^O\left[\sum_{j=1}^{24} w_{1j}^H Z_j + b_1^O\right] \quad \textbf{(III.12)}$$

$$Z_j = f^H\left[\sum_{i=1}^{20} w_{ij}^I E_i + b_j^H\right] \quad \textbf{(III.13)}$$

$$S = f^O\left[\sum_{j=1}^{02} w_{1j}^H Z_j + b_1^O\right] \quad \textbf{(III.14)}$$

$$Z_j = f^H\left[\sum_{i=1}^{11} w_{ij}^I E_i + b_j^H\right] \quad \textbf{(III.15)}$$

$$S = f^O\left[\sum_{j=1}^{02} w_{1j}^H Z_j + b_1^O\right] \quad \textbf{(III.16)}$$

$$Z_j = f^H\left[\sum_{i=1}^{16} w_{ij}^I E_i + b_j^H\right] \quad \textbf{(III.17)}$$

$$S = f^{O}\left[\sum_{j=1}^{06} w_{1j}^{H} Z_j + b_1^{O}\right] \quad \textbf{(III.18)}$$

$$Z_j = f^{H}\left[\sum_{i=1}^{09} w_{ij}^{I} E_i + b_j^{H}\right] \quad \textbf{(III.19)}$$

$$S = f^{O}\left[\sum_{j=1}^{10} w_{1j}^{H} Z_j + b_1^{O}\right] \quad \textbf{(III.20)}$$

Chaque neurone caché calcule la fonction d'activation, en envoyant des résultat (Z_j); $(j = 1 \text{ à } 7\text{-}20)$ au neurone de la couche de sortie, qui donne finalement la réponse de réseau (S) de chaque modèle. Le signale de sortie de chaque neurone caché (Z_j) est calculé par les équations : **Eq. (III.11)**, **Eq. (III.13)**, **Eq. (III.15)**, **Eq. (III.17)**, et **Eq. (III.19)**, respectivement pour les modèles de : $D_{AB_Couples}$ à Différents T°, D_{AB_Air} à T=25°C, D_{AB_Eau} à T=25°C, D_{AB_Air} à T=100°C et D_{AB_Eau} à T=100°C. Les sorties de chaque réseau sont données par les équations : **Eq. (III.12)**, **Eq. (III.14)**, **Eq. (III.16)**, **Eq. (III.18)**, et **Eq. (III.20)**, respectivement pour les cinq modèles développés dans le même ordre mentionné ci-dessus.

Les poids (w_{ij}^{I}, w_{1j}^{H}) et les biais (b_j^{H}, b_1^{O}) des cinq modèles pour chaque couche (cachée et sortie) ne sont pas exposés, vu qu'il contient plusieurs valeurs numériques et nous considérons qu'il est inutile de les présentées, cependant ils sont des constates numériques compris généralement entre 1 et -1.

III.3.4. Estimation de la performance des résultats d'estimation issus des modèles développés

III.3.4.1. Comparaison graphique entre les valeurs des coefficients de diffusion expérimentales et calculées

Afin d'estimer les cinq coefficients de diffusion D_{AB}, nous utilisons les équations des modèles présenter précédemment, en introduisant les valeurs des entrées E_i qui sont les valeurs des descripteurs pertinents choisis pour chaque modèle, plus les valeurs des poids (W_{ij}^{I}, W_{1j}^{H}) et les biais (b_j^{H}, b_1^{0}) de connexions correspondantes à chaque modèle.

Les comparaisons entre les valeurs expérimentales et calculée obtenues par les modèles développés sont illustrées sur les figures : **figure III.09** pour le modèle $D_{AB_Couples}$ à Différents T°, **figure III.10** pour le modèle de D_{AB_Air} à T=25°C,la **figure III.11** pour le modèle de D_{AB_Eau} à T=25°C, la **figure III.12** pour le modèle de D_{AB_Air} à T=100°C et la **figure III.13** pour le modèle de D_{AB_Eau} à T=100°C, pour l'ensemble de base des données entières.

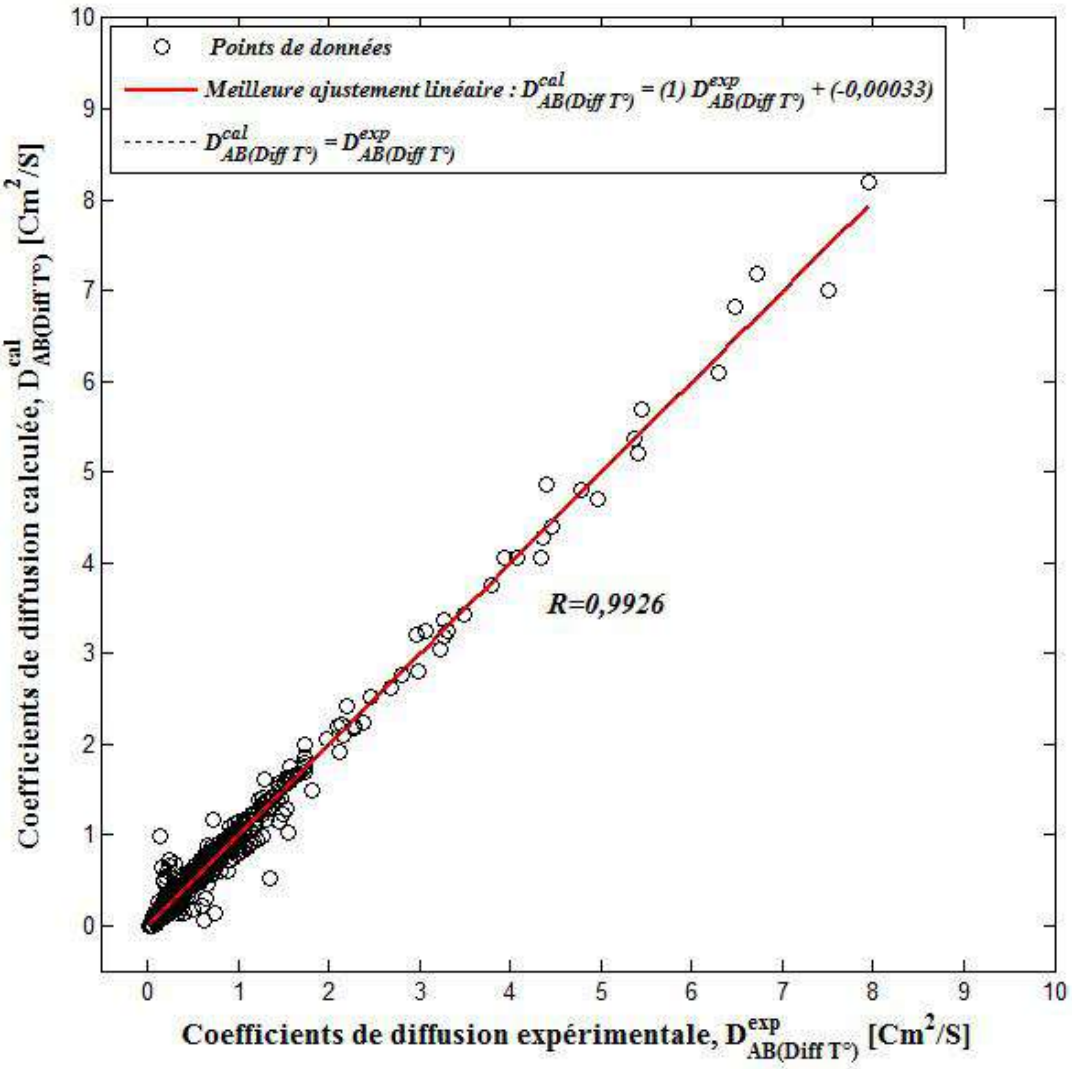


Figure III.09. Comparaison entre les valeurs expérimentales et calculées par le modèle neuronale de coefficient de diffusion $D_{AB_Couples}$ à Différents T° pour la base de données entière.

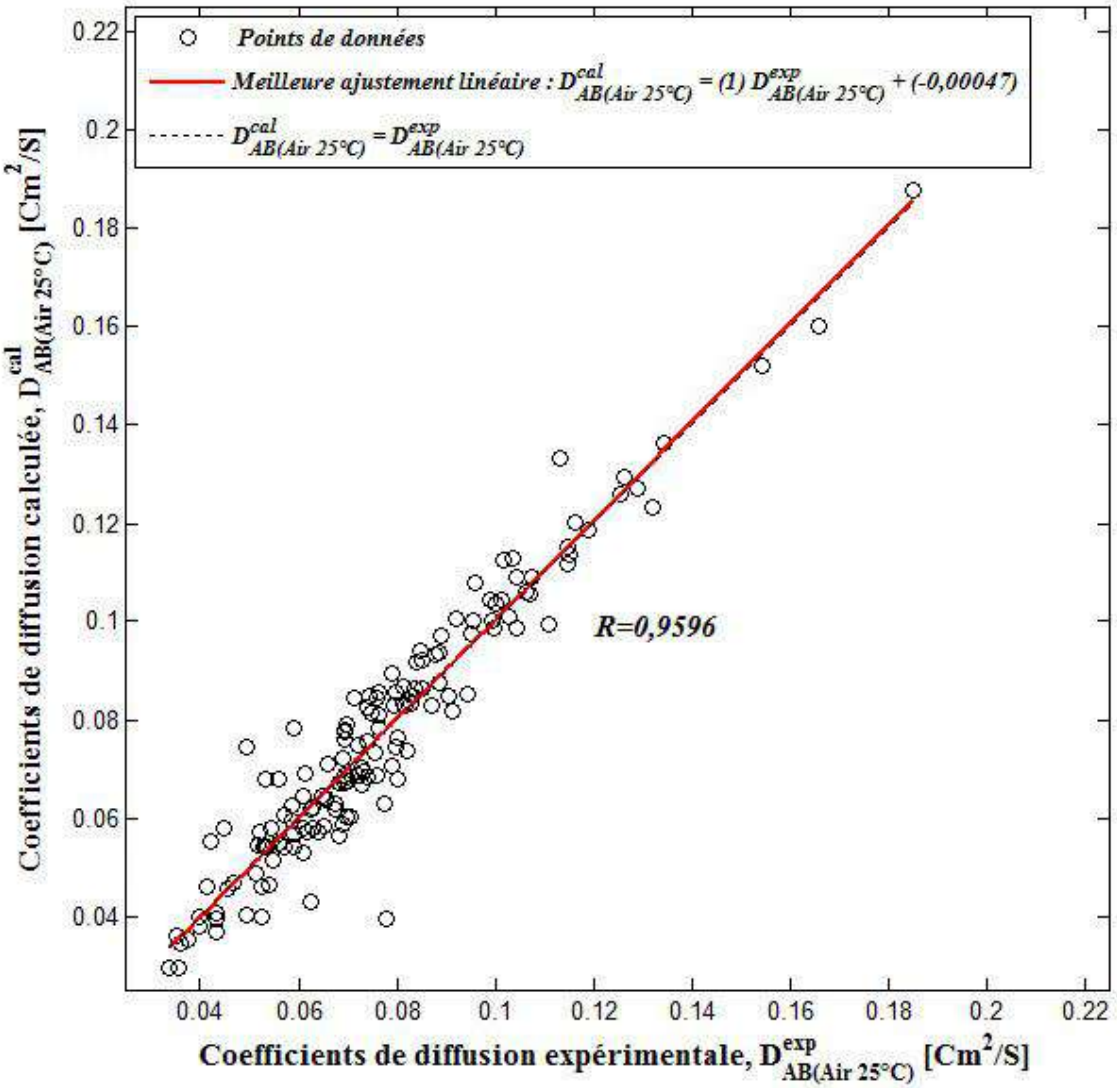


Figure III.10. Comparaison entre les valeurs expérimentales et calculées par le modèle RQSP-RNA de coefficient de diffusion D_{AB_Air} à T=25°C pour la base de données entière.

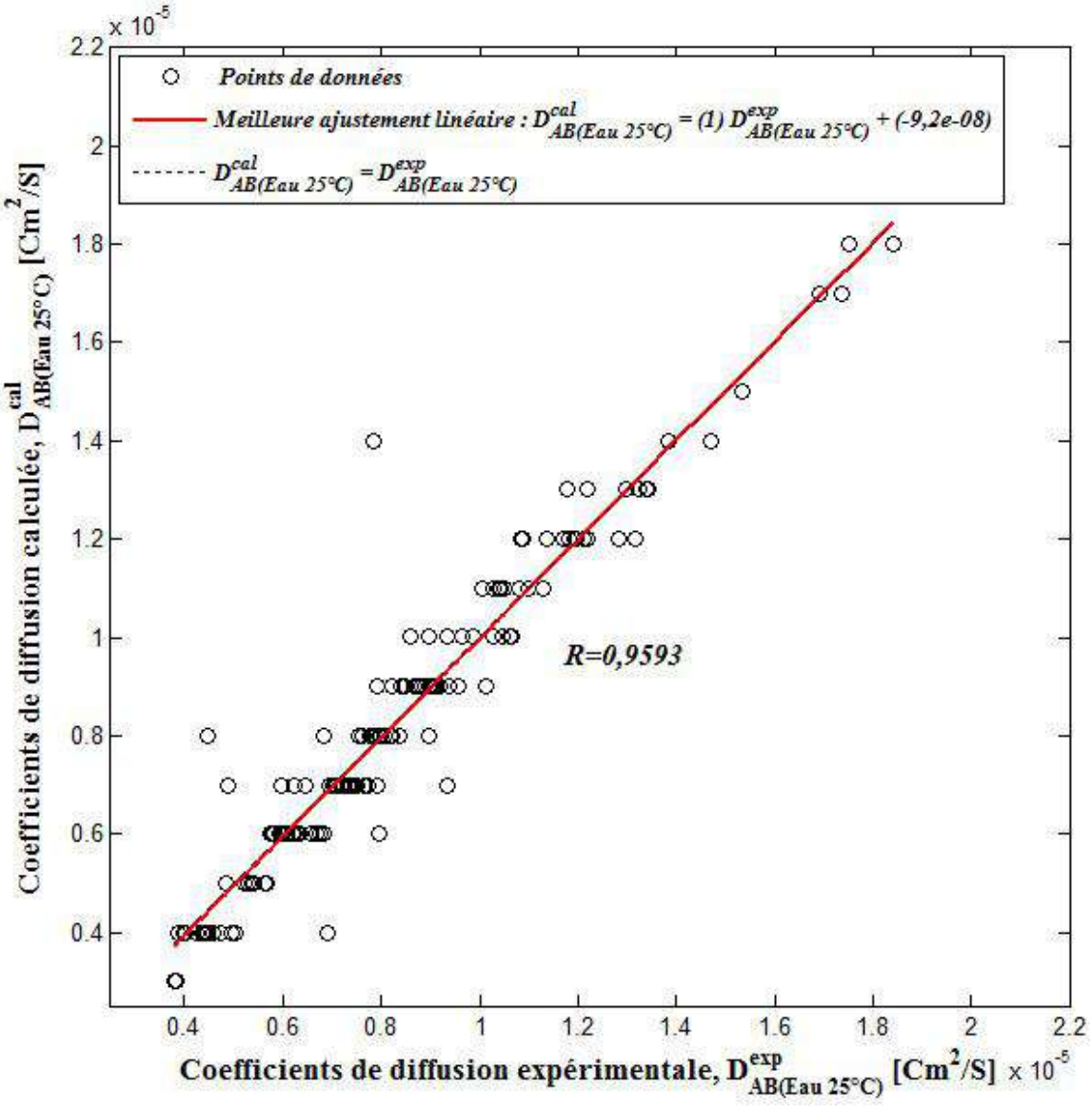


Figure III.11. Comparaison entre les valeurs expérimentales et calculées par le modèle RQSP-RNA de coefficient de diffusion D_{AB_Eau} à T=25°C pour la base de données entière.

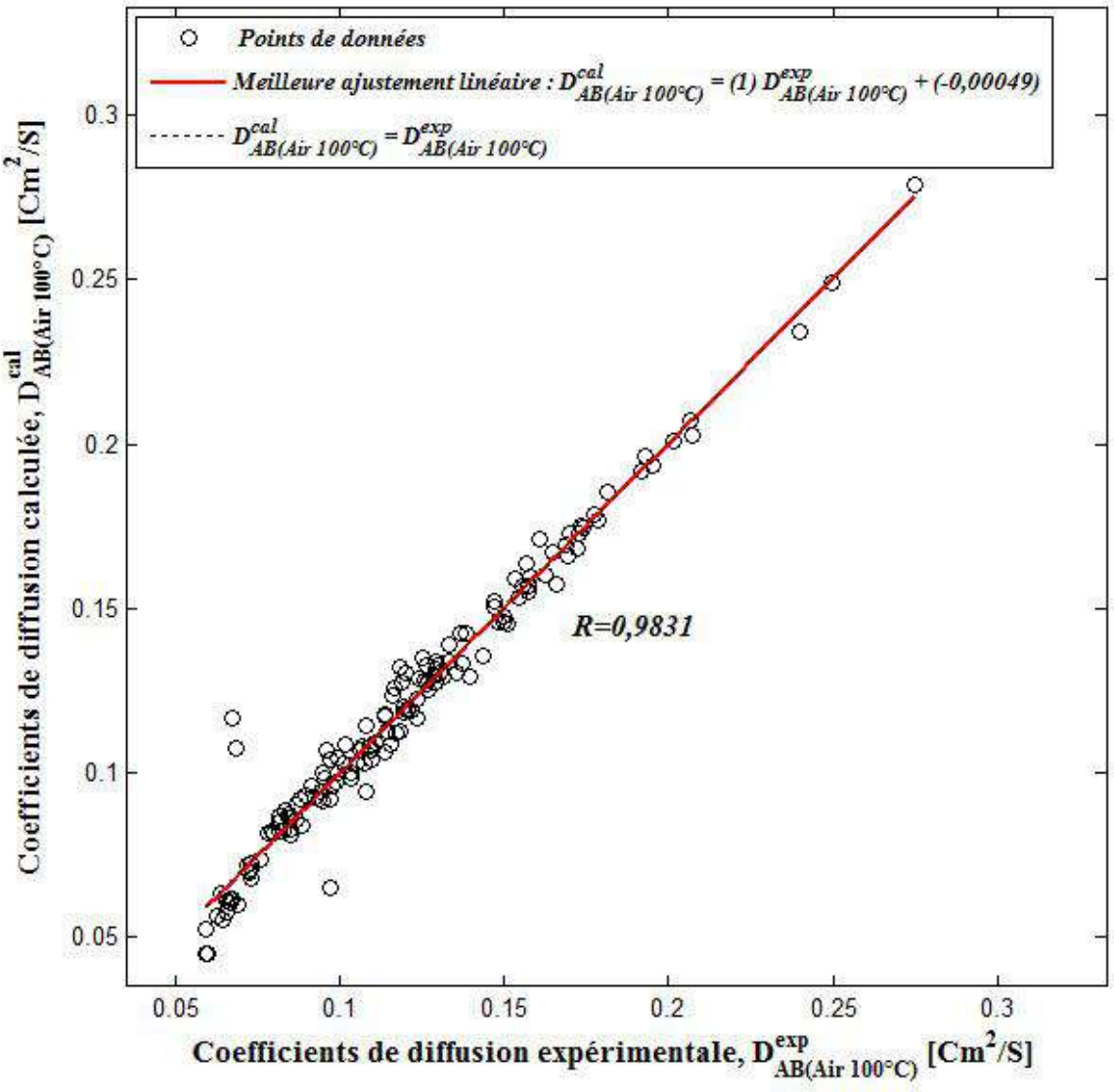


Figure III.12. Comparaison entre les valeurs expérimentales et calculées par le modèle RQSP-RNA de coefficient de diffusion D_{AB_Air} à T=100°C pour la base de données entière.

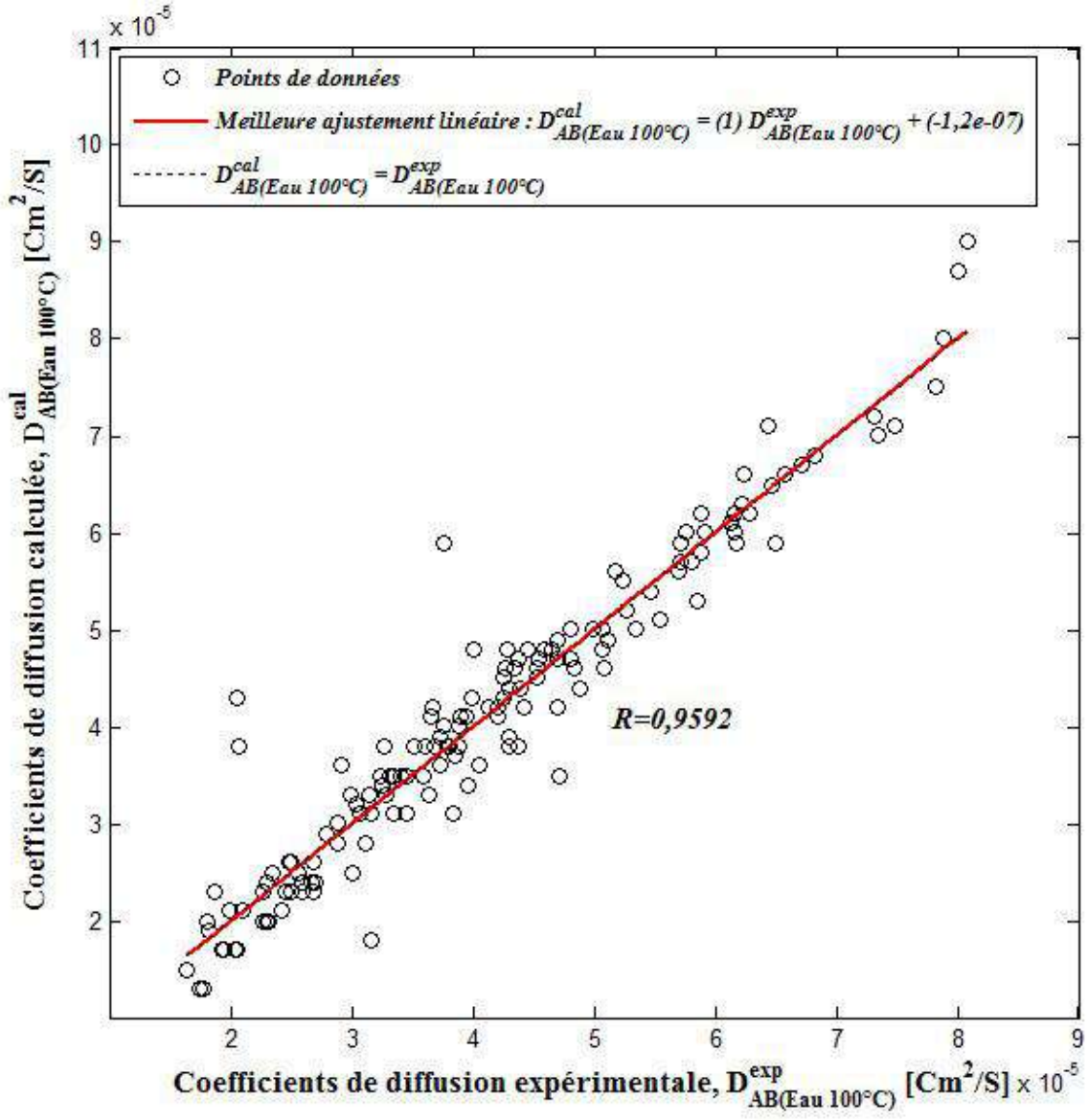


Figure III.13. Comparaison entre les valeurs expérimentales et calculées par le modèle RQSP-RNA de coefficient de diffusion D_{AB_Eau} à T=100°C pour la base de données entière.

Une concentration étroite de points de données autour de la ligne de pente unitaire indiquée par la ligne en pointillé, révèle que les modèles peuvent prédire les coefficients de diffusion $\mathbf{D_{AB}}$ des cinq systèmes étudiés avec une précision acceptable, confirmé par leurs résultats satisfaisants, qui ont une accord de précision avec les valeurs vectorielles de corrélation proches de l'idéal (i.e. α = 1 (pente), β = 0 (ordonnée à l'origine), R = 1), dans les graphiques appropriés des cinq modèles pour l'ensemble de base des données entière (voir les équations de régression désigné en haut des **figures III.09–III.13**).

III.3.4.2 Calculs des paramètres de performance et analyse statistique des cinq modèles obtenus

Les performances des meilleurs modèles sont résumées dans le **tableau III.10**, en termes de la racine d'erreur quadratique moyenne (REQM) et l'erreur standard de prédiction (ESP) correspondant à chaque modèle dans tous les sous-ensembles possibles : apprentissage, test, validation et dans la base globale. En plus des valeurs de R, R^2 et le nombre de points de données (N) dans chaque phase.

Tableau III.10. Paramètres statistiques des meilleurs modèles choisir pour les cinq coefficients de diffusion D_{AB} étudiées.

Paramètres statistiques	R (–)	R^2 (–)	REQM (cm^2/s)	ESP (%)	N (–)
Phase d'apprentissage					
$D_{AB\,Couples\ à\ \neq T°}^{App}$	0,9926	0,9853	0,1353	14,2423	509
$D_{AB\,Air\ à\ T=25°C}^{App}$	0,9596	0,921	0,0085	8,9521	119
$D_{AB\,Eau\ à\ T=25°C}^{App}$	0,9593	0,9202	0,0000009	9,2702	119
$D_{AB\,Air\ à\ T=100°C}^{App}$	0,9831	0,6964	0,0085	5,7765	119
$D_{AB\,Eau\ à\ T=100°C}^{App}$	0,9592	0,9201	0,000002	9,8538	119
Phase de test					
$D_{AB\,Couples\ à\ \neq T°}^{Test}$	0,9796	0,9597	0,1552	27,6787	109
$D_{AB\,Air\ à\ T=25°C}^{Test}$	0,9791	0,9586	0,0051	6,8527	29
$D_{AB\,Eau\ à\ T=25°C}^{Test}$	0,9899	0,98	0,0000004	5,1698	29
$D_{AB\,Air\ à\ T=100°C}^{Test}$	0,9924	0,9848	0,0046	4,0233	29
$D_{AB\,Eau\ à\ T=100°C}^{Test}$	0,9867	0,9736	0,000005	6,3441	29
Phase de validation					
$D_{AB\,Couples\ à\ \neq T°}^{Val}$	0,9921	0,9843	0,0974	14,6944	109
Phase entière					
$D_{AB\,Couples\ à\ \neq T°}^{Ent}$	0,9926	0,9853	0,1132	17,0212	727
$D_{AB\,Air\ à\ T=25°C}^{Ent}$	0,9596	0,921	0,0076	9,9835	148
$D_{AB\,Eau\ à\ T=25°C}^{Ent}$	0,9593	0,9202	0,0000008	10,3382	148
$D_{AB\,Air\ à\ T=100°C}^{Ent}$	0,9831	0,9664	0,0076	6,442	148
$D_{AB\,Eau\ à\ T=100°C}^{Ent}$	0,9592	0,9201	0,000004	10,989	148

Selon les résultats des différents coefficients statistiques et types d'erreur présentés dans le **tableau III.10**, les modèles développés étaient très intéressants, comme montrent les valeurs de R^2 qui sont supérieures à **0,9200** pour les cinq modèles et dans toutes les sous bases des données possible.

Les valeurs de REQM calculées ont aussi confirmé que nos modèles sont très précis car ils ont des valeurs compris entre 0,000009–0,1552 cm²/s pour le paramètre D_{AB} à différents couples, entre 0,046–0,085 cm²/s pour le paramètre D_{AB} de l'Air à T=25°C, entre 0,0000004–0,0000009 cm²/s pour le paramètre D_{AB} de l'Eau à T=25°C, entre 0,046–0,085 cm²/s pour le paramètre D_{AB} de l'Air à T=100°C, et entre 0,000002–0,000005 cm²/s pour le paramètre D_{AB} de l'Eau à T=100°C, ceci est pour tous les sous-ensembles possibles.

L'erreur standard de prédiction (ESP) est en fait de même type que celle de la REQM mais exprimée en pourcentage (%), ce qui donne une vision plus claire pour mieux juger les modèles. Les valeurs des ESP sont inférieures à 28% pour les cinq modèles de D_{AB} dans tous les sous-ensembles possibles. En fait qu'ils ont des valeurs de ESP comprise entre 14,2423-27,6787% pour D_{AB} à différents couples, entre 6,8527-9,9835% pour D_{AB} de l'Air à T=25°C, 5,1698-10,3382% pour D_{AB} de l'Eau à T=25°C, entre 4,0233-6,442% pour D_{AB} de l'Air à T=100°C, et entre 6,3441-10,988% pour D_{AB} de l'Eau à T=100°C.

Néanmoins, nous remarquons que toutes les valeurs d'erreurs calculées dans cette étude, très proche les uns aux autres pour le même paramètre et dans tous les sous-ensembles possibles : d'apprentissage, de test, de validation et entière. Démontrant également l'équilibre entre le sous-apprentissage (sous-ajustement) et le sur-apprentissage (sur-ajustement) des modèles neuronaux actuels, ce qui a confirmé encore la stabilité de nos modèles.

III.3.4.3 Analyse de sensibilité des paramètres d'entrées des cinq modèles développés par la méthode de poids

Comme nous avons mentionné précédemment dans le paragraphe **III.2.2.6**, nous allons utiliser la méthode d'analyse de sensibilité au poids, pour apprécier la contribution les paramètres d'entrées sur chaque modèle, en déterminant l'impact de chaque descripteur d'entrée individuellement dans le cas de la première contribution et l'impact des paramètres d'entrées dans le cas de la deuxième contribution, sur la sortie de chaque modèle développé.

Les contributions des entrées choisir sur les modèles développés, ont été établie et illustrés sur les **figure III.14**, **figure III.15, figure III.16, figure III.17** et la **figure III.18** respectivement pour les modèles de $D_{AB_Couples}$ à Différents T°, D_{AB_Air} à T=25°C, D_{AB_Eau} à T=25°C, D_{AB_Air} à T=100°C et le modèle de D_{AB_Eau} à T=100°C.

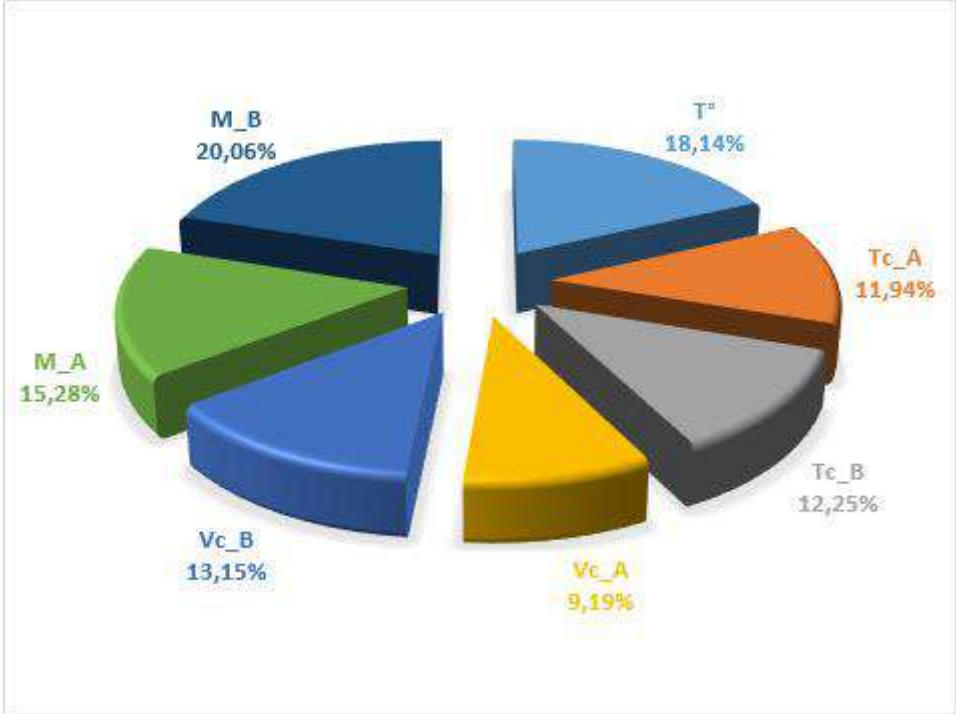


Figure III.14. Contributions relatives des sept (7) entrées pour le modèle neuronal de coefficient de diffusion $D_{AB_Couples}$ à Différents T°.

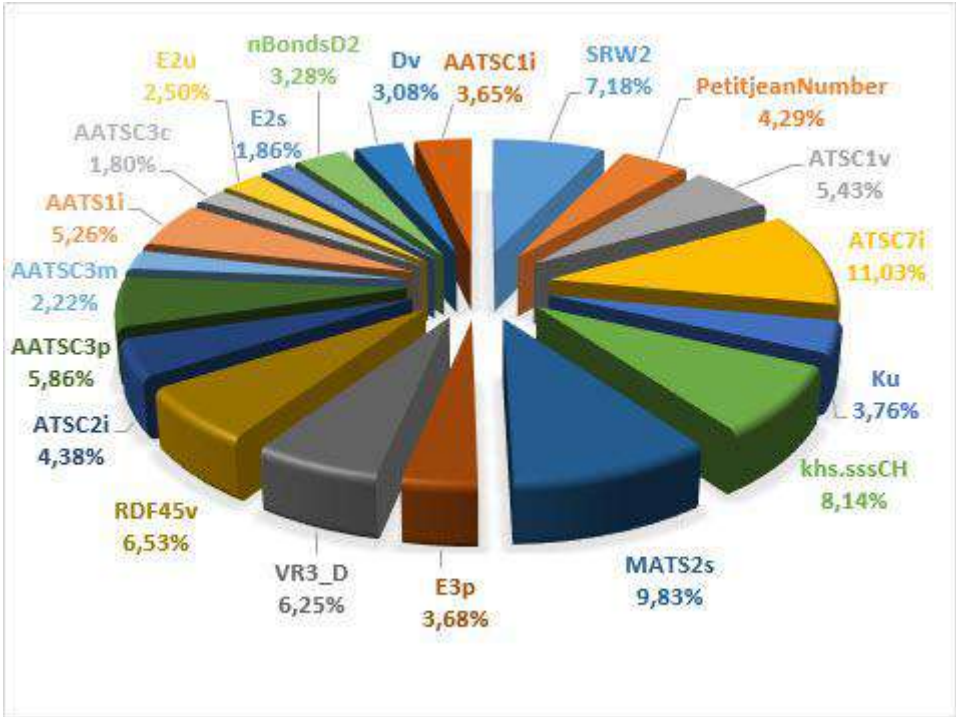


Figure III.15. Contributions relatives des vingt (20) descripteurs pertinents d'entrées pour le modèle de type RQSP-RNA de D_{AB_Air} à T=25°C.

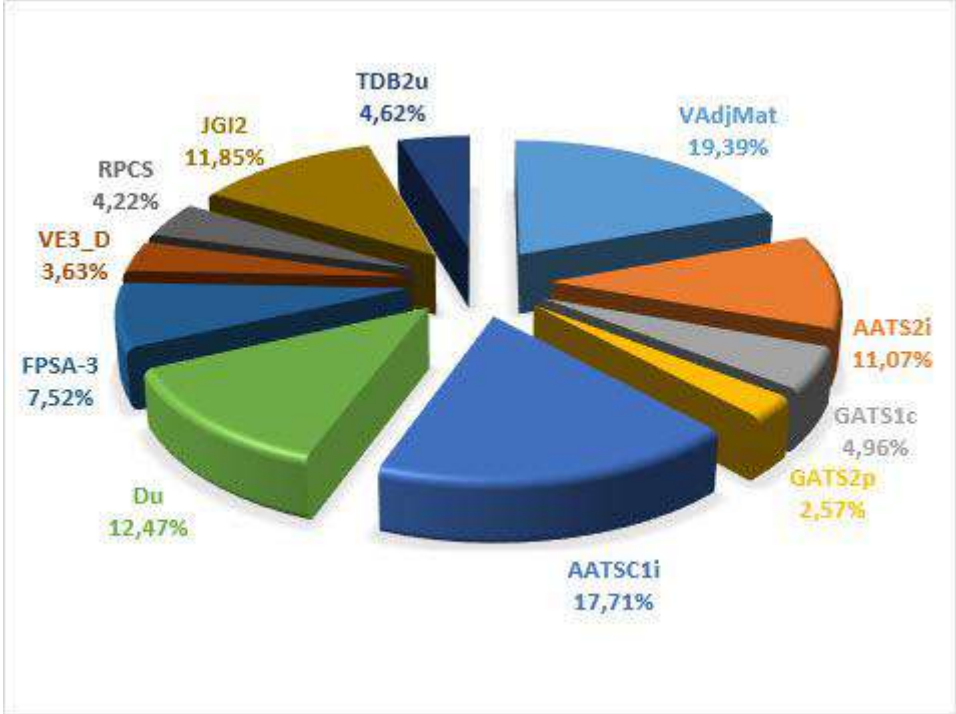


Figure III.16. Contributions relatives des onze (11) descripteurs pertinents d'entrées pour le modèle de type RQSP-RNA de D_{AB_Eau} à T=25°C.

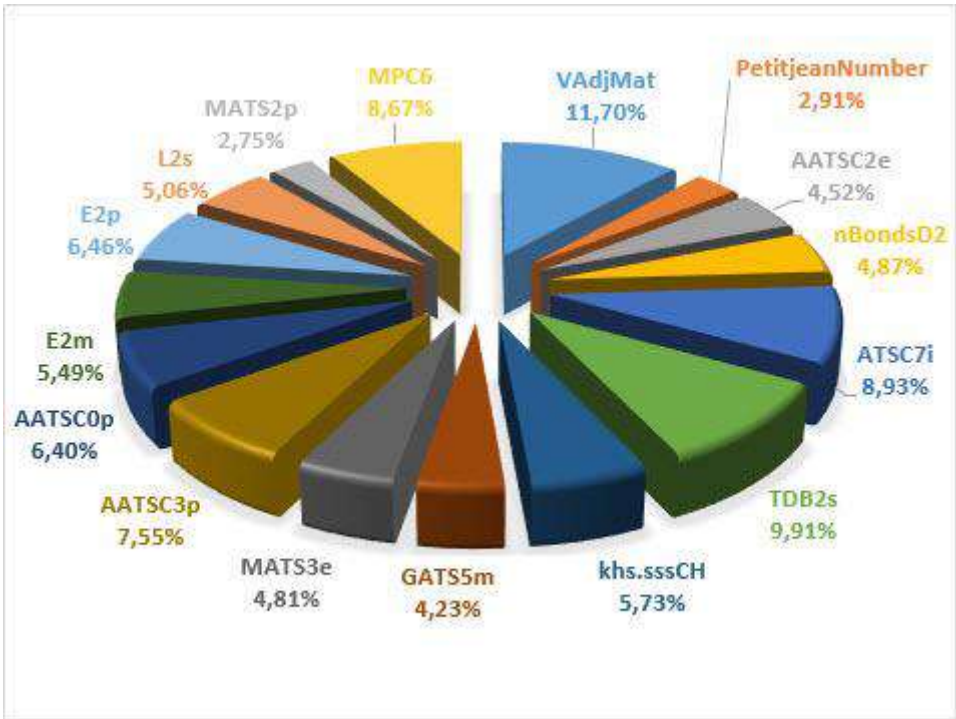


Figure III.17. Contributions relatives des seize (16) descripteurs pertinents d'entrées pour le modèle de type RQSP-RNA de D_{AB_Air} à T=100°C.

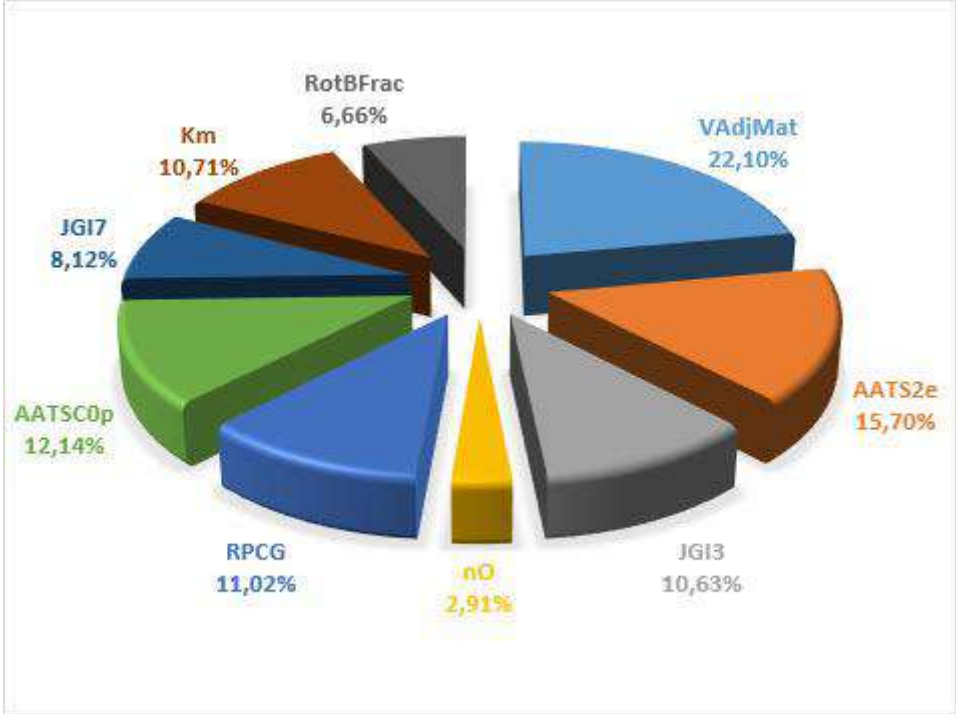


Figure III.18. Contributions relatives des neuf (09) descripteurs pertinents d'entrées pour le modèle de type RQSP-RNA de D_{AB_Eau} à T=100°C.

A partir les résultats de l'analyse de sensibilité qui sont traduit par l'importance relative (IR) de chaque descripteur présenté dans les figures précédentes, et les valeurs des contributions moyennes (VCM) de chaque modèle qui sont : 14,285% pour le modèle de $D_{AB_Couples}$ à Différents T°, 5% pour le modèle de D_{AB_Air} à T=25°C, 9,09% pour le modèle de D_{AB_Eau} à T=25°C, 6,25% pour le modèle de D_{AB_Air} à T=100°C et enfin 11,11% pour le modèle de D_{AB_Eau} à T=100°C. A l'exception de certains descripteurs ou paramètres d'entrées qui ont de faibles contributions, la grande majorité des descripteurs dans le cas des quatre modèles de la première contribution ou les paramètres d'entrées dans le cas de deuxième contribution, possèdent une moyenne ou une forte contribution.

Il convient de noter qu'il n'y a pas une grande différence entre les descripteurs ou les paramètres d'entrées, en particulier ceux qui ont de forte ou de moyenne contribution par apport aux (VCM) de chaque modèle. Ce qui signifie que tous les descripteurs sélectionnés étaient importants à des degrés divers dans le développement des modèles de RQSP-RNA.

III.4. Conclusion

Nous avons présenté dans ce chapitre l'ensemble des travaux de modélisation des coefficients de diffusion pour différents systèmes d'étude de la pollution : $D_{AB_Couples}$ à Différents

T°, D_{AB_Air} à T=25°C, D_{AB_Eau} à T=25°C, D_{AB_Air} à T=100°C et D_{AB_Eau} à T=100°C0. Ce chapitre est organisé en deux parie principale.

Dans la première partie nous avons expliqué l'ensemble des matériels et des méthodes utilisés lors du développement des modèles de type RQSP-RNA en décrivant les différentes étapes. Nous avons cité tous les logiciels et les outils informatiques utilisés pour chaque étape du développement. Tous ces étapes ont été résumé dans un organigramme, ensuit nous avons expliqué en détails chaque méthode utilisée pour le développement des modèles, commençant par les étapes de prétraitement, ou bien précédemment par la procédure de collection et traitement des bases des données, tout aussi pour l'optimisation des structures chimique et le codage SMILES, jusqu'à le choix des paramètres des modèles avant l'exécution.

Dans la deuxième partie de ce chapitre, on a discuté l'ensemble des résultats obtenus par les modèles développés, afin de prédire les cinq coefficients de diffusion qui sont principalement importants pour l'étude de la pollution atmosphérique et dans l'eau. Nous avons discuté plusieurs résultats, telles que : l'évaluation de nombre de descripteurs au cours des étapes de prétraitement pour les quatre modèle de la première contribution, les architectures et les résultats préliminaire des meilleures modèles obtenus, l'estimation des erreurs et le calcul des coefficients statistiques pour chaque modèle ; tel que le calcul des REQM et ESP, l'analyse de la sensibilité des descripteurs et des paramètres d'entrées sur les cinq sorties pour chaque modèle, puis la validation des modèles finaux de type RQSP-RNA dans le cas de première contribution.

Les résultats d'analyse de la performance des meilleurs modèles développées ont montré leurs précisions dans l'estimation des coefficients de diffusion dans différents cas d'étude, confirmé par les valeurs élevées de ses coefficients de détermination (R^2) et les faibles valeurs des deux erreurs calculées (REQM et ESP).

Conclusion générale et perspectives

La connaissance des propriétés physique et chimiques telles que les propriétés de transports jouent un rôle fondamental dans l'ingénierie environnementale, vu son importance dans la plupart des phénomènes et procédés liés à ce domaine. Cependant, le calcul des propriétés du transport dépend du milieu ou du système de transport, tel que les milieux aqueux, atmosphériques, naturels ou forcés, de l'état de ce dernier tel que l'environnement pollué ou non-pollué, en plus des propriétés physicochimiques entourées par les polluants. Certaines propriétés comme les coefficients de diffusion ne peuvent pas être facilement mesurées, cela est dû principalement aux tests de laboratoire coûteux et longs.

Dans ce projet de fin d'étude, nous voulons aborder ce problème qui trait un sujet très important dans le domaine scientifique de recherche lié avec le milieu industriel. L'objectif à porter sur la modélisation de coefficient de diffusion D_{AB} des différents polluants, notre contribution était sur le développement des modèles de type RQSP-RNA afin de prédire des coefficients D_{AB} dans cinq systèmes d'études différents : $D_{AB_Couples}$ à Différents T°, D_{AB_Air} à T=25°C, D_{AB_Eau} à T=25°C, D_{AB_Air} à T=100°C et D_{AB_Eau} à T=100°C, seulement à partir de leurs structures chimiques pour les quatre dernier modèles et en utilisant quelques propriétés liés à notre sortie pour la première modèle, une nouvelle technique sous l'égide de l'intelligence computationnelle ou des sciences cognitives a été utilisé pour développées nos cinq modèles, qui a connu par les réseaux de neurones artificiels perceptron multicouches (RNA-PMC).

La méthodologie du développement suivez dans notre étude comportent plusieurs étapes basées sur les précédentes études dans la littérature. L'ensemble de données expérimentales qui égale à 148 molécules a été divisé en deux sous-ensembles 80% pour l'apprentissage et 20% pour le test pour la première contribution et 727 molécules a été divisé en trois sous-ensembles 70% pour l'apprentissage, 15% pour le test et 15% pour la validation pour la deuxième contribution. Un nombre important des descripteurs moléculaire ont été calculées en ligne à l'aide d'une plateforme spécialisée, ensuite plusieurs étapes basées sur des méthodes statistiques y compris la méthode de RLM ont été suivies pour réduire ce nombre important de descripteurs, afin de ne conserver que les descripteurs pertinents. Les meilleurs modèles ont été obtenus par l'utilisation de la stratégie d'essais et d'erreurs avec une algorithme d'apprentissage de type "BFGS", les architectures des meilleurs modèles obtenus sont : "07-24-1" pour $D_{AB_Couples}$ à Différents T°, "20-2-1" pour D_{AB_Air} à T=25°C, "11-2-1" pour D_{AB_Eau} à T=25°C, "16-6-1" pour D_{AB_Air} à T=100°C et "09-10-1" pour D_{AB_Eau} à T=100°C. Ultérieurement, chaque modèle a été analysé et testé par des méthodes et des paramètres statistiques telles que : le

coefficient de détermination (R^2), la Racine d'Erreur Quadratique Moyen (REQM) et l'analyse de sensibilité par la méthode de poids.

Les résultats était très encourageants confirmées par les faibles valeurs des REQM ; qui a été compris entre 0,000009–0,1552 cm²/s pour le paramètre $D_{AB_Couples}$ à Différents T°, entre 0,046–0,085 cm²/s pour le paramètre D_{AB_Air} à T=25°C, entre 0,0000004–0,0000009 cm²/s pour le paramètre D_{AB_Eau} à T=25°C, entre 0,046–0,085 cm²/s pour le paramètre D_{AB_Air} à T=100°C et entre 0,000002–0,000005 cm²/s pour le paramètre D_{AB_Eau} à T=100°C. L'analyse de sensibilité a été aussi réalisée pour déterminer l'influence des descripteurs d'entrées dans le cas des modèles de la première contribution, et d'influence des paramètres d'entrées dans le cas du modèle de la deuxième contribution sur chaque paramètre de sortie D_{AB}, les résultats ont confirmé que la majorité des entrées ont de haut ou de moyen contribuion, ce qui confirme encore la qualité des descripteurs ou de paramètre d'entrées choisir dans chaque modèle.

Comme la majorité des objectives fondamentales de la modélisation ; le premier modèle neuronal et les quatre restantes modèles de type RQSP/RQSA sont conçus pour prédire différents coefficients de diffusion dans différents systèmes d'étude afin de réduire le temps, l'argent et les efforts utilisés à des fins expérimentales.

En vertu de ce travail, on peut tirer quelques perspectives qui peut être des informations utiles pour les prochaines recherches dans ce domaine :

- Il est possible d'appliquer le concept des RQSP/RQSA pour étudier d'autres paramètres dans d'autres domaines tell que les paramètres pharmacocinétiques dans le domaine de la modélisation des médicaments.
- Plusieurs logiciels et outils informatiques commerciales et open source sont conçue pour faciliter la tâche de développement des modèles de ce type, tel que le logiciel Dragon 7 commerciale et E-Dragon open source.
- Nous pouvons aussi appliquer l'idée de ce travail à d'autres techniques récentes de modélisation telles que les Machines de Vecteur de Support (SVM).
- On peut utiliser d'autres logiciels comme Matlab dans l'apprentissage des RNAs, à cause de leur stabilité et il fournit plusieurs choix telles que le type d'algorithme d'apprentissage et le nombre des couches cachées utilisées.

Références bibliographiques

ACD/Labs. (2015). Release, Advanced Chemistry Development, Inc., Toronto, ON, Canada.

Al Barakeh, Z. (2012). *Suivi de pollution atmosphérique par système multi-capteurs–méthode mixte de classification et de détermination d'un indice de pollution.* Thèse de doctorat en génie des procédés. École Nationale Supérieure des Mines de Saint-Etienne (EMSE). France.

Andrews, R., Diederich, J., & Tickle, A. B. (1995). Survey and critique of techniques for extracting rules from trained artificial neural networks. *Knowledge-based systems*, *8*(6), 373-389.

Arqués, P. (1998). La pollution de l'air, causes, conséquences, solutions, édition Edisud-ca calade, 207p.

ASQA (2010). **A**ssociation de la **S**urveillance de la **Q**ualité de l'**A**ir : Les principaux polluants Airparif, 2, 1.

Beale, M. H., Hagan, M. T., & Demuth, H. B. (2012). Neural network toolbox™ user's guide. In *R2012a, The MathWorks, Inc., 3 Apple Hill Drive Natick, MA 01760-2098, www. mathworks. com*.

BEM Recycling (2020). Site web d'une entreprise innovante spécialisée dans l'environnement, la valorisation des déchets d'équipements électriques et électroniques, lien : http://bemrecycling.com/la-pollution-industrielle-impacts-sur-lhomme-et-lenvironnement/, Consulté le 15/05/2020.

Benson, N. (2008). Encyclopedia of Global Warming and Climate Change. Ed.S. Philander. Vol 3. Thousand Oaks. CA: Sage Publications Inc .3 :813-817.

Bianchi, A. M., Fautrelle, Y., & Etay, J. (2004). *Transferts thermiques*. PPUR presses polytechniques.

Bishop, C. M. (1995). *Neural networks for pattern recognition*. Oxford university press.

Bitam, S. (2018). "*Modélisation de l'activité thérapeutique à partir de la structure moléculaire*". Thèse de Doctorat en génie des procédés, *Laboratoire de Biomatériaux et Phénomènes de Transport (LBMPT)*, Département de génie des procédés et environnement, Université Yahia Fares de Médéa. Médéa, Algérie.

Bjijou, W. (2016). Traitement des rejets de nature chimique et biologique par la technique de Biosorption.

Bonachera, F. (2011). *Les triplets pharmacophoriques flous: développement et applications*. Thèse de Doctorat Université de Lille1. France.

Brégains, J. C. (2007). *Análisis, síntesis y control de diagramas de radiación generados por distribuciones continuas y discretas utilizando algoritmos estocásticos y redes neuronales artificiales. Aplicaciones*. Univ Santiago de Compostela.

Cartwright, H. (2015). Artificial Neural Networks. *Methods in Molecular Biology*, *1260*.

Cengel, Y. A., & Cimbala, J. M. (2017). *Mécanique des fluides*. De Boeck supérieur.

Charpin, D., Pairon, J. C., Annesi-Maesano, I., Caillaud, D., De Blay, F., Dixsaut, G., ... & Delaval, P. (2016). La pollution atmosphérique et ses effets sur la santé respiratoire. Document d'experts du groupe pathologies pulmonaires professionnelles environnementales et iatrogéniques (PAPPEI) de la Société de pneumologie de langue française (SPLF)[Outdoor pollution and its effects on lung health. Expert document from the groupe pathologies pulmonaires professionnelles environnementales et iatrogéniques (PAPPEI) of the Société de pneumologie de langue française (SPLF)]. *Revue des Maladies Respiratoires*, *33*(6), 484-508.

Charton, M. (2002). *Advances in Quantative Structure-Property Relationships*. Elsevier.

ChemOffice (2012). PerkinElmer Informatics. CambridgeSoft. 100 CambridgePark Drive, Cambridge, MA 02140. www.cambridgesoft.com.

Chtita, S. (2017). *Modélisation de molécules organiques hétérocycliques biologiquement actives par des méthodes QSAR/QSPR. Recherche de nouveaux médicaments*. Thèse de doctorat en Chimie. Faculté des sciences, Université Moulay Ismaïl. Meknès, Maroc.

Constantinou, L., & Gani, R. (1994). New group contribution method for estimating properties of pure compounds. *AIChE Journal*, *40*(10), 1697-1710.

Cramer, R. D., Patterson, D. E., & Bunce, J. D. (1988). Comparative molecular field analysis (CoMFA). 1. Effect of shape on binding of steroids to carrier proteins. *Journal of the American Chemical Society*, *110*(18), 5959-5967.

Cruciani, G., Pastor, M., & Guba, W. (2000). VolSurf: a new tool for the pharmacokinetic optimization of lead compounds. *European Journal of Pharmaceutical Sciences*, *11*, S29-S39.

Csizmadia, P. (1999). MarvinSketch and MarvinView: molecule applets for the World Wide Web.

Davalo, E., Naïm, P., & Clément, F. M. (1989). *Des réseaux de neurones*. Paris : Eyrolles.

De la Fuente Aparicio, M. J., & Cano, T. C. (1999). *Aplicaciones de las redes de neuronas en supervisión, diagnosis y control de procesos*. Equinoccio.

Debacq, M. (2019). Phénomènes de transferts de matière en génie des procédés.

Díez, R. P., Gómez, A. G., & de Abajo Martínez, N. (2001). *Introducción a la inteligencia artificial: sistemas expertos, redes neuronales artificiales y computación evolutiva*. Universidad de Oviedo.

Dirac, P. A. M. (1928). The quantum theory of the electron. *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character*, *117*(778), 610-624.

Dong, J., Cao, D. S., Miao, H. Y., Liu, S., Deng, B. C., Yun, Y. H., ... & Chen, A. F. (2015). ChemDes: an integrated web-based platform for molecular descriptor and fingerprint computation. *Journal of cheminformatics*, *7*(1), 60.

Douali, L., Villemin, D., & Cherqaoui, D. (2003). Neural networks: accurate nonlinear QSAR model for HEPT derivatives. *Journal of chemical information and computer sciences*, *43*(4), 1200-1207.

Dudek, A. Z., Arodz, T., & Gálvez, J. (2006). Computational methods in developing quantitative structure-activity relationships (QSAR): a review. *Combinatorial chemistry & high throughput screening*, *9*(3), 213-228.

Etienne, M., Le Page, C., & Bousquet, F. (2015). Le collectif ComMod: modélisation et accompagnement.

Fishwick, P. A., & Zeigler, B. P. (1992). A multimodel methodology for qualitative model engineering. *ACM Transactions on Modeling and Computer Simulation (TOMACS)*, *2*(1), 52-81.

Frisch, M. J. E. A., Trucks, G. W., Schlegel, H. B., Scuseria, G. E., Robb, M. A., Cheeseman, J. R., ... & Nakatsuji, H. (2009). gaussian 09, Revision d. 01, Gaussian. *Inc., Wallingford CT*, *201*.

Frisch, M. J., Trucks, G. W., Schlegel, H. B., Scuseria, G. E., Robb, M. A., Cheeseman, J. R., ... & Millam, J. M. (2004). Gaussian 03, Revision C. 02. Wallingford, CT: Gaussian. *Inc.*

Gevrey, M., Dimopoulos, I., & Lek, S. (2003). Review and comparison of methods to study the contribution of variables in artificial neural network models. *Ecological modelling*, *160*(3), 249-264.

Gingras, J. C. et C. (2011). Les changements Climatiques : Quels en sont les causes et les Impacts ? Agriculture et Climat, 4.

Goulon-Sigwalt-Abram, A. (2008). *Une nouvelle méthode d'apprentissage de données structurées: applications à l'aide à la découverte de médicaments*. Thèse de doctorat, Université Pierre et Marie Curie-Paris VI. France.

Goverse, T. (2013). UNEP Year Book 2013: Emerging issues in our global environment.

Goyena, R., & Fallis, A. (2019). Pengaruh Work life Balance terhadap Kepuasan Kerja. *Journal. Chemical. Information. Modelling*, *53*, 1689-1699.

Gramatica, P., Chirico, N., Papa, E., Cassani, S., & Kovarich, S. (2013). QSARINS: A new software for the development, analysis, and validation of QSAR MLR models. *Journal of Computational Chemistry*, *34*(24), 2121-2132.

Hamad, B., & Khahla S. (2014). Contribution à la Prédiction de Coefficient de partage octanol /eau par la technique QSPR (Quantitive Structure Proprety Relationship) mémoire de Master, Département de chimie, Université d'El-Oued. Algérie.

Hamadache, M. (2014). "*Modélisation en vue de la prédiction ou de la corrélation de l'activité toxicologique à partir de la structure moléculaire*". Thèse de Doctorat en génie des procédés, *Laboratoire de Biomatériaux et Phénomènes de Transport (LBMPT)*, Département de génie des procédés et environnement, Université Yahia Fares de Médéa. Médéa, Algérie.

Hohenberg, P., & Kohn, W. (1964). Inhomogeneous electron gas physical review 136. *B864*.

Hopfield, J. J. (1982). Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the national academy of sciences*, *79*(8), 2554-2558.

HyperCube, I. (2003). HyperChem; Release 7.5, Standard Version ed.; HyperCube. *Inc.: Gainesville, FL.*

Jensen, F. (2007). Electron correlation methods. *Jensen F. Introduction to computational chemistry. 2nd ed. New Jersey: John Wiley & Sons*, 133-144.

Khaouane, L. (2013). "*Etude et Modélisation de la Biosynthèse des Antibiotiques à partir de Différentes Souches Productrices – Cas de Pleuromutiline*". Thèse de Doctorat en génie des procédés, *Laboratoire de Biomatériaux et Phénomènes de Transport (LBMPT)*, Département de génie des procédés et environnement, Université Yahia Fares de Médéa. Médéa, Algérie.

Kohn, W. (1999). Nobel Lecture: Electronic structure of matter—wave functions and density functionals. *Reviews of Modern Physics*, *71*(5), 1253.

Kohn, W., & Sham, L. J. (1965). Self-consistent equations including exchange and correlation effects. *Physical review*, *140*(4A), A1133.

Landry, G., Seco, J., Gaudreault, M., & Verhaegen, F. (2013). Deriving effective atomic numbers from DECT based on a parameterization of the ratio of high and low linear attenuation coefficients. *Physics in Medicine & Biology*, *58*(19), 6851.

Lejeune, M. (2004). *Statistique: La théorie et ses applications*. Springer Science & Business Media.

Lemière, B., Seguin, J. J., Le Guern, C., Guyonnet, D., Baranger, P., Saada, A., ... & Hubé, D. (2001). Guide sur le comportement des polluants dans les sols et les nappes. *BRGM éditions*, *300*, 132p.

López, R. F., Fernandez, J. M., & Fernández, J. M. F. (2008). *Las redes neuronales artificiales*. Netbiblo, S.L., Espagne, pp 12-14, 17, 21.

Mamoudou, B. I., & Ténéman, K. O. N. A. T. E. (2018). Détermination du niveau de la pollution de l'air par le plomb d'origine routière dans la ville de Khemis Miliana.

MarvinSketch. (2012). V. 5.11. 4. ChemAxon Ltd., Budapest, Hungary, (http://www. chemaxon. com).

Mayer, H. (1999). Air pollution in cities. *Atmospheric environment*, *33*(24-25), 4029-4037.

Melzi, N., Khaouane, L., Ammi, Y., Hanini, S., Laidi, M., & Zentou, H. (2019). Comparative Study of Predicting the Molecular Diffusion Coefficient for Polar and Non-polar Binary Gas Using Neural Networks and Multiple Linear Regressions. *Kemija u industriji: Časopis kemičara i kemijskih inženjera Hrvatske*, *68*(11-12), 573-582.

Moilleron, R., Bergé, A., Deshayes, S., Zedek, S., Eudes, V., & Bressy, A. (2017). Domestic micropollutants, a diffuse industrial pollution–Part 1: from greywater to receiving water.

Moody, J., & Darken, C. J. (1989). Fast learning in networks of locally-tuned processing units. *Neural computation*, *1*(2), 281-294.

Mulliken, R. S. (1928). The assignment of quantum numbers for electrons in molecules. II. Correlation of molecular and atomic electron states. *Physical review*, *32*(5), 761.

Park, C., & Allaby, M. (2007). A dictionary of environment and conservation. *Agenda*, *21*, 12.

Penot, F. (1980). Transfert de chaleur par convection naturelle dans une cavité rectangulaire isotherme ouverte sur une face. *Revue de Physique Appliquée*, *15*(2), 207-212.

Peric, B., Sierra, J., Martí, E., Cruañas, R., & Garau, M. A. (2015). Quantitative structure–activity relationship (QSAR) prediction of (eco) toxicity of short aliphatic protic ionic liquids. *Ecotoxicology and Environmental Safety*, *115*, 257-262.

Puzyn, T., Leszczynski, J., & Cronin, M. T. (Eds.). (2010). *Recent advances in QSAR studies: methods and applications* (Vol. 8). Springer Science & Business Media.

Rahal, F. (2015). *Modélisation et simulation de la pollution atmosphérique. Le cas de la région d'Alger*. Thèse de Doctorat, Université des Sciences et de la Technologie d'Oran Mohamed Boudiaf, Oran.

Rainelli, P. (1996). Pollution des sols. *Problèmes économiques. Étude et Gestion des Sols*, 3-4.

Ravier, S. S., Rigaut, M., Carbonnet, J., Roques, M., Normale, E., Lyon, S. De, Rigaut, M., Ravier, S. S., & Rigaut, M. (2000). (1000). Mécanique des fluides. 1-14. /Users/benjaminroche/Documents/Biblio/Ravier2000.pdf.

Riazi, M. R. (2005). *Characterization and properties of petroleum fractions* (Vol. 50). ASTM international.

Roubehie Fissa, M., Lahiouel, Y., Khaouane, L., & Hanini, S. (2019). QSPR estimation models of normal boiling point and relative liquid density of pure hydrocarbons using MLR and MLP-ANN methods. *Journal of Molecular Graphics and Modelling*, *87*, 109-120.

Rumelhart, D. E., Hinton, G. E., & McClelland, J. L. (1986). A general framework for parallel distributed processing. *Parallel distributed processing: Explorations in the microstructure of cognition*, *1*(45-76), 26.

Schrodinger, E. (1926). Quantisierung als Eigenwertproblem; von Erwin Schrodinger. *Annalen der Psysik*, 361-377.

Si-Moussa, C. (2010). "*Modélisation des équilibres de phases à pressions élevées par les réseaux de neurones artificiels*". Thèse de doctorat. École nationale supérieure polytechnique. Alger. Algérie.

StatSoft, I. (2011). STATISTICA for Windows (data analysis software system), Version 10.0. StatSoft Inc, Tulsa, OK.

Stewart, J. J. (2004). Optimization of parameters for semiempirical methods IV: extension of MNDO, AM1, and PM3 to more main group elements. *Journal of Molecular Modeling*, *10*(2), 155-164.

Stewart, J. J. P. (2009). MOPAC2009, version 10.060 W. *Stewart Computational Chemistry: Colorado Springs, CO*.

Taine, J., & Petit, J. P. (2003). Transferts thermiques. *Sciences Sup. Paris, dunod edition*.

Talbi, M. L. (2011). Analyse et traitement du signal électro cardiographique (ECG). Thèse de Doctorat, Université Mentouri de Constantine. Constantine, Algérie.

The MathWorks, Inc. USA (2013), Matlab software, Version 2013b.

Thomas, L. H. (1927). The calculation of atomic fields. In *Mathematical Proceedings of the Cambridge Philosophical Society* (Vol. 23, No. 5, pp. 542-548). Cambridge University Press.

Todeschini, R., & Consonni, V. (2009). *Molecular descriptors for chemoinformatics: volume I: alphabetical listing/volume II: appendices, references* (Vol. 41). John Wiley & Sons.

US Environmental Protection Agency. (2006). Technical Support Document for the Proposed PM NAAQS Rule: Response Surface Modeling.

Varenne, F., Silberstein, M., & Dutreuil, S. (2018). *Modéliser et simuler: épistémologies et pratiques de la modélisation et de la simulation* (Vol. 2). Editions Matériologiques.

Wassim, G. (2016). Cours de pollution et nuisances, section : LFSNA3. *Université de GABES*, *2017*.

Weininger, D. (1988). SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *Journal of chemical information and computer sciences*, *28*(1), 31-36.

Yaws, C. L., & Gabbula, C. (2003). *Yaws" Handbook of Thermodynamic and Physical Properties of Chemical Compounds*. Knovel.