وزارة التعليم العالى و البحث العلمى

Université BADJI Mokhtar – Annaba
BADJI Mokhtar – Annaba University

جامعة باجي مختار – عنابة

Faculté des Sciences
Département de Chimie

MÉMOIRE

Présenté pour l'obtention du diplôme MAGISTER

Par M^elle^. DIDI Mabrouka

Option : Chimie et environnement

THÈME

PREDICTION DE LA TOXICITE D'UNE SERIE D'AMIDES HERBICIDES

Devant le jury :

PRESIDENT :	**M^r^. D. MESSADI**	**Pr**	**UBMA**
EXAMINATEURS :	**M^me^. S. ALI-MOKHNACHE**	**Pr**	**UBMA**
	M^r^. A.TOUBAL	**MC**	**UBMA**
RAPPORTEUR :	**M^me^. N. FERTIKH**	**MC**	**UBMA**
Invitée :	**M^elle^. I.TOUHAMI**	**CC**	**CU Khenchela EPSTA**

Année 2010

Dédicace

Je dédie ce modeste travail à :

☒ Mes chers parents

☒ Mes frères et soeurs

☒ Mes nièces et mes neveux

☒ Mes amis

☒ Enfin, toute l'équipe du labo 34.

REMERCIEMENTS

Ce mémoire n'aurait pas vu le jour
sans la confiance, la patience et la générosité du
responsable de la P.G. Monsieur
le Professeur D. ***MESSADI*** *que je remercie vivement pour avoir accepté la présidence de ce jury. Je voudrais aussi le remercier pour le temps et la patience qu'il m'a accordés tout au long de ces années.*
Je tiens également à remercier :
Mme FERTIKH .N, pour la direction de ce travail ;
Mme ALI MOKHNACHE Salima, pour avoir accepté d'examiner ce travail;
Mr TOUBAL.A, pour avoir accepté d'examiner ce travail aussi.
M^{elle} TOUHAMI .I, pour avoir accepté l'invitation.
Enfin, je ne saurais oublier toute l'équipe du laboratoire 34.

ملخص:

تم تطوير نموذجين بطريقة الـ QSAR للتنبؤ بمؤشر الأستبقاء.

المعطيات الخاصة بـ50 مركب من تم تقسيمها الى مجموعتين الاولى تحتوي على40 عنصر لحساب و تجريب النموذج اما الثانية تحتوي على 10 عناصر لتصديق الخارجي للنموذج.

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

المواصفات الجزيئية النظرية تم حسابها باستعمال برمجيات النمذجة الجزيئية المتوفرة في السوق . حجم النموذج تم تحديده عن طريق دالة الـ FIT لـ Kubinyi، اما اختيار المواصفات عن طريق الخوارزمية المورثية .

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

الكلمات الدالة:

المبيدات - السمية - التركيز المميت 50 - نموذج التهجن المتعدد الخطي و نموذج التهجن اللا خطي .

Résumé:

Deux modèles QSAR ont été développés pour la prédiction de la toxicité des amides herbicides carectérisée par la dose letale 50 . Les données, concernant 50 composés amides herbicides ont été séparées en deux sous-ensembles disjoints comprenant respectivement 40 éléments pour le calcul et le test (éventuel) du modèle, et 10 éléments pour sa validation statistique externe. Deux modèles ont ainsi été créés sur le même ensemble de données: un modèle de régression multilinéaire et un modèle de réseaux de neurones artificiels.

Des descripteurs moléculaires théoriques ont été calculés en utilisant des logiciels de modélisation moléculaire du commerce. La taille du modèle à été déterminée en optimisant le FIT de KUBINYI, et la sélection des descripteurs réalisée par algorithme génétique.

Les valeurs des paramètres statistiques (R^2, Q^2, Q_{Ext}, EQMC, EQMP, EQMP (ext)) obtenues attestent de la pertinence des modèles développés, avec une supériorité établie pour les modèles de neurones artificiels.

Mots-clés:

Herbicides - Toxicité - Dose létale50 – Modèles hybrides mixtes multilinéaires et non linéaires .

ABSTRACT

Two QSAR models were developed for the prediction of the toxicty of a set of amides herbicides. A dataset of 50 compounds of amide herbicides was subdivised into two disjoined subsets containing respectively 40 compounds for calculating and (possible) testing of the model, and 10 compounds used for the external validation .Two models based on the same subset of data were thus created : a multiple linear regression model and an artificial neural network model.

Theoretical molecular descriptors were calculated using commercially available molecular modelling softwares. The model size was determined by optimizing the FIT of KUBINYI, and the selection of the descriptors realized by genetic algorithm.

Values obtained for the statistical parameters: R^2, Q^2, Q_{Ext}, SDEC, SDEP and $SDEP_{ext}$, attest relevance of the models developed, with a clear superiority for the artificial neural network models.

Key words :

Herbicides – Toxicity –Lethal dose 50 –Linear and nonlinear hybridic models.

SOMMAIRE

SOMMAIRE

CHAPITRE II : PARTIE THEORIQUE

SYMBOLES ET ABREVIATIONS

AG: Algorithme génétique (Genetic Algorithm).

AM1 : Austin Model 1.

DL: Dose létale.

EQM: Erreur quadratique moyenne.

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

EQMP: Ecart quadratique moyen de prédiction.

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

e_i : Résidu ordinaire.

$e_{i\ std}$: Résidu standardisé.

F : Statistique de Fisher.

FIT: Fonction de Kubinyi.

FIV: Facteur d'inflation de la variance.

H : Matrice de projection, ou matrice châpeau.

h_{ii} : Eléments diagonaux de la matrice chapeau.

IW : Initial Weight (Poids entrée-couche cachée).

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

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

LW : last Weight(Poids couche cachée-sortie).

n: Dimension de la population (échantillon).

n-p : Nombre de degrés de liberté.

PMC: Perception multicouches.

PRESS : Somme des carrés des erreurs de prédiction.

p_c : Probabilité de croisement.

p_M: Probabilité de mutation.

Quantitative Structure/ Activity Relationships

QSAR : (Relations Structure/ activité Quantitatives).

Q^2_{LOO} Coefficient de prédiction (leave one out).

RHF: Restricted Hartree Fock

RLM: Régression linéaire multiple.

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

RNA: Réseaux de neurones artificiels.

RSA : Relation Structure-Activité.

R^2 : Coefficient de détermination.

S : Erreur standard.

SCE : Somme des carrés des écarts.

SCT : Somme des carrés totale.

w_k : Weight at the moment k(Poids à l'instant k).

w_{k+1} : Weight at the moment k -1(Poids à l'instant k-1).

X : Matrice des valeurs observées.

X´: Matrice transposée de .

x_j **:** Variable explicative.

x_j **:** jième valeur de .

x_{max} **:** Valeur maximale.

x_{min} **:** Valeur minimale

x_{norm} **:** Valeur normalisée.

Y: Vecteur de dimension n.

$\mathbf{y_i}$**:** Valeur observée.

$\hat{y}_i$ Valeur estimée.

α : Niveau de confiance; Facteur d'apprentissage.

σ^2 **:** Variance.

δ_k : Différence entre la sortie attendue et la sortie effective à l'instant k.

LISTE DES FIGURES

INTRODUCTION GENERALE

Introduction :

L’agriculture intensive actuelle emploie 90% des pesticides utilisés, reposant sur une gamme de plus de 8000 produits commercialisés à travers le monde. Faisant fi de l’équilibre des écosystèmes, le recours massif à des produits de synthèse menace nos ressources naturelles et la santé des populations [1].

En plus des conséquences agronomiques importantes (érosion et baisse de la fertilité des sols, développement de résistances aux substances actives chez les insectes et parasites), les pesticides polluent durablement l’environnement.

Les pesticides sont retrouvés dans 91 % des mesures faites dans les cours d’eau et dans plus de la moitié des nappes souterraines. Entre 25 et 75 % des quantités de pesticides pulvérisés se disséminent dans l’atmosphère [1].

Sachant qu’un tiers de l’alimentation humaine dépendrait du succès de la pollinisation, la vigilance est impérative. L’impact néfaste des pesticides touche l’ensemble de la biodiversité : chauve-souris, amphibiens, oiseaux etc. Par ailleurs le verdict scientifique est formel des dizaines d’études épidémiologiques montrent que nos corps sont contaminés par les pesticides. Elles font le lien entre l’exposition aux pesticides et certains cancers, affaiblissement du système immunitaire, fragilité respiratoire et digestive, propension à développer des maladies neurodégénératives de type Parkinson. Les agriculteurs et leurs familles sont les premiers touchés par cette exposition. Elle serait ainsi source de baisse de la fertilité (27 fois plus de risque de problèmes de fertilité pour les femmes ayant manipulé un herbicide) [1].

Hélas ! Jusqu'à maintenant rien n'est su sur la toxicité de plus de 100 000 produits chimiques jetés dans l'environnement, dont de 1-5 % des données de la toxicité sont disponibles et il est excessivement cher d'obtenir de telles informations expérimentalement (en temps, animaux et coût) [2].

Par conséquent, les compagnies et les agences régulatrices tournent vers la prédiction de la toxicité à travers l'usage des relations quantitatives structure / activité (QSAR) [2].

Les techniques les plus courantes pour établir des modèles QSAR utilisent l'analyse de régression (régression multilinéaire : RLM ; les réseaux neurones artificiels RNA ; régression par composantes principales (ACP) et la technique de régression par les moindres carrés partiels (MCP ou PLS).

Des logiciels informatiques spécialisés permettent le calcul de plus de 3000 descripteurs moléculaires appartenant à différentes classes. Plutôt que de rechercher à expliquer la variable dépendante (propriété biologique considérée) par tous les régresseurs (descripteurs moléculaires), on peut chercher seulement un ensemble réduit de régresseurs qui donne une reconstitution aussi satisfaisante de la variable à expliquer.

Parmi les stratégies mises en œuvre pour la sélection d'un ensemble limité de variables explicatives, on peut citer : les méthodes de pas à pas, ainsi que les algorithmes génétiques.

Nous avons appliqué des méthodes hybrides: algorithme génétique/régression multilinéaire (AG/RLM), et algorithme génétique/réseaux de neurones artificiels (AG/RNA) pour modéliser la toxicité de 50 amides herbicides.

Dans ce travail, nous nous sommes intéressés à la toxicité d'une série d'amides herbicides caractérisée par la dose létale 50 (pDL_{50}) testées sur des rats mâles par voie orale.

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

* Partie théorique, où nous avons décrit le prétraitement des molécules (introduction de la liste des molécules, optimisation de leur géométrie) en vue du calcul des descripteurs moléculaires à l'aide de différents logiciels de modélisation moléculaire. Nous y avons également développé les connaissances théoriques de base utilisées tout au long de ce travail: algorithmes génétiques, régression multilinéaire, réseaux de neurones artificiels, traitement statistique pour l'évaluation de la qualité de l'ajustement; robustesse des modèles ; détection des observations aberrantes; test de randomisation et validation externe.

* Partie expérimentale, où nous présentons et discutons les modèles calculés.

ETUDE BIBLIOGRAPHIQUE

I-1 - RELATION QUANTITATIVE STRUCTURE / ACTIVITE :

Une **relation quantitative structure / activité** (en anglais : Quantitative structure-activity relationship ou **QSAR**, parfois désignée sous le nom de **relation quantitative structure / propriété** - en anglais : quantitative structure-property relationship ou **QSPR**) est le procédé par lequel une structure chimique est corrélée avec un effet bien déterminé comme l'activité biologique ou la réactivité chimique.

Ainsi par exemple l'activité biologique peut être exprimée de manière quantitative, comme la concentration de substance nécessaire pour obtenir une certaine réponse biologique.

De plus lorsque les propriétés ou structures physicochimiques sont exprimées par des chiffres, on peut proposer une relation mathématique, ou relation quantitative structure / activité, entre les deux. L'expression mathématique obtenue peut alors être utilisée comme moyen prédictif de la réponse biologique pour des structures similaires.

Construire un modèle QSAR consiste à établir une relation mathématique entre une propriété mesurable et la structure chimique qui peut-être décrite par des variables chimiques la distribution électronique ou les propriétés stériques.

Ces relations mathématiques sont des fonctions des « descripteurs » judicieusement choisis pour une propriété donnée. Ces relations sont construites à l'aide de logiciels utilisant des méthodes mathématiques plus ou moins sophistiquées

La QSAR la plus commune est de la forme : activité = f (propriétés physico-chimiques, biologiques, ou structurales) [3].

I-1-1 - La RSA et paradoxe RSA :

Le postulat de base pour les hypothèses sur des objets chimiques est que des objets similaires ont des activités similaires. Ce principe est appelé relation structure-activité (RSA, ou SAR pour structure-activity relationship en anglais). Le problème sous-jacent est donc la définition d'une petite différence sur un niveau moléculaire, chaque type d'activité, comme la réaction chimique, la biotransformation, la solubilité, l'activité de cible et d'autres encore, peuvent dépendre d'une autre différence. Un exemple concret est donné par l'article de revue sur le bioisostérisme [4].

En général, l'intérêt est plus de trouver de fortes tendances. Les hypothèses avancées reposent habituellement sur un nombre fini de données chimiques. Ainsi, le principe d'induction

devrait être respecté afin d'éviter les hypothèses sur apprises et les interprétations erronées et inutiles sur les données chimiques/structurales.
Le paradoxe SAR est le fait que toutes les molécules similaires ne montrent pas des activités similaires [4].

I-1-2 - Applications en chimie :

Une des premières applications de la QSAR concernait la prédiction des points d'ébullition.

Il est bien connu par exemple que pour une famille de composés chimiques, particulièrement en chimie organique, il existe une corrélation forte entre la structure et les propriétés observées.

On peut citer comme exemple simple la relation entre le nombre de carbones dans les alcanes et leur point d'ébullition. Il existe une tendance nette à l'augmentation de la température d'ébullition avec le nombre d'atomes de carbone, ce qui sert de moyen prédictif pour les points d'ébullition des alcanes les plus lourds [5].

I-1-3 - Utilisation:

L'utilisation de modèles QSAR pour la gestion du risque chimique s'accroissant régulièrement et étant aussi utilisé pour des visées réglementaires (en Union européenne : enregistrement, évaluation et autorisation des produits chimiques), il est crucial d'être capable d'affirmer la pertinence des prédictions. L'espace des descripteurs chimiques engendré par un ensemble spécifique de produits chimiques est appelé domaine d'application, qui permet d'indiquer lorsqu'un composé peut être pertinemment «prédit» [3].

I-2- LES PESTICIDES :

1-2-1 - Définition :

Sous le nom générique de pesticides se cachent de nombreux produits chimiques qui ont des rôles bien définis. Herbicides, insecticides, fongicides...

Les **pesticides** que l'on appelle aussi « produits phytosanitaires » dans l'agriculture sont des produits issus de l'industrie chimique reposant sur un principe actif d'origine naturelle ou synthétique. Ils sont utilisés en agriculture pour se débarrasser des nuisibles.

Lorsque les nuisibles sont des insectes ravageurs on utilise des insecticides : plusieurs familles sont commercialisées ou ont été commercialisées.

Lorsque les nuisibles sont des champignons pouvant causer des maladies on utilise des **fongicides**. Les agents actifs de synthèse utilisés aujourd'hui reposent selon les formulations sur les carbamates, les dérivés du benzène, les dérivés du phénol, les quinones, les amines, les amides, les triazoles, etc.… Certains de ces produits sont considérés comme potentiellement cancérigènes.

Lorsque les nuisibles sont des herbes indésirables concurrentes des cultures on utilise des **herbicides** : ces produits aussi appelés communément désherbants sont soit sélectifs (ils ne tuent que les mauvaises herbes) soit de portée plus radicale.

De nombreuses familles d'herbicides sont sur le marché dont notamment les phénols nitrés,les amides, les benzonitriles, les urées substituées, les triazines, les sulfonurées, les ammonium quaternaires... la liste est longue ! Certains de ces produits sont considérés également comme potentiellement cancérigènes.

Lorsque les nuisibles ne sont pas des insectes, ni des champignons, ni des mauvaises herbes, on utilise des produits ciblés comme les molluscicides contre les limaces, les rodenticides contre les rongeurs, les corvicides contre les corbeaux... là encore, la liste est très longue[1] .

1-2-2 - Les herbicides :

1-2-2-1 - Définition :

Un produit **herbicide** est un type de pesticide défini comme une substance active ou une préparation ayant la propriété de tuer les végétaux . Le terme « désherbant » est un synonyme d'herbicide.

En protection des cultures, les herbicides sont employés pour lutter contre les adventices, ou mauvaises herbes, destinées à détruire ou à limiter la croissance des végétaux, qu'ils soient herbacés ou ligneux. Ils peuvent être utilisés, selon leur mode d'action, en pré ou post-levée.
On distingue:

o Les désherbants sélectifs, les plus nombreux.

o Les débroussaillants et désherbants totaux.

o Les défanants qui détruisent la partie aérienne des végétaux. Ils sont par exemple utilisés pour la récolte mécanique de la pomme de terre ou de la betterave.

o Les anti-germes, qui empêchent le démarrage de la végétation, par exemple, les oignons ou pommes de terre destinés à l'alimentation [6].

1-2-2-2 - Types d'herbicides:

A- Herbicides synthétiques :

- ❖ Sélectifs : utilisés pour tuer des variétés végétales données tout en laissant la récolte relativement intacte.
- ❖ Non sélectifs : utilisés pour dégager des terrains vagues et tuer tous les végétaux et les matériaux avec lesquels ils entrent en contact.

B - Herbicides Naturels :

Certaines plantes produisent des herbicides organiques naturels, par exemple le Juglans noyer (genre de plante originaire des régions tempérées et chaude ; principalement de l'hémisphère nord).

Ces herbicides sont beaucoup moins efficaces et généralement plus chers que les herbicides synthétiques. Ils sont généralement combinés à des pratiques culturelles et mécaniques de contrôle des mauvaises herbes. Entre autres exemples mentionnons les épices, le vinaigre, la vapeur et les flammes [7].

1-2-2-3 - Les groupes d'herbicides :-

Il existe trois grands groupes d'herbicides : les herbicides de pré-levée, les herbicides de post-levée et les herbicides totaux.

A- Les herbicides de pré-levée:

Ils ont été les premiers diffusés en zone tropicale, notamment sur les cultures de la rotation cotonnière et les cultures industrielles (canne à sucre). Ces produits sont faciles à vulgariser, car leur spectre d'efficacité est souvent assez large et ils s'appliquent à une période bien définie juste après le semis.

Toutefois, ces herbicides sont très dépendants de l'état physique du sol : ils ne peuvent pas être appliqués sur un sol trop motteux ou couvert par un paillis épais. Leur disponibilité dans la solution du sol dépend de la texture. Le produit est adsorbé par les feuillets d'argile ou les

colloïdes de la matière organique. Inversement, en sol sableux, les risques de phytotoxicité sont accrus.

La pluie, avant ou après l'application, favorise généralement la diffusion à la surface du sol de ces herbicides à pénétration racinaire ; cependant, une pluie érosive qui survient après l'application risque d'entraîner le produit par ruissellement [8].

B - Les herbicides de post-levée:

Fréquemment employés en culture de riz ou de canne à sucre, ils sont choisis en fonction de la flore des mauvaises herbes présentes. Ces produits sont souvent spécifiques : action anti-dicotylédone (la suppression des dicotylédones herbacées et des mauvaises herbes) en culture de maïs, de riz ou de canne à sucre, action graminicide (un brûlement chimique rapide des mauvaises herbes vivaces et la réduction des parties aériennes) en culture de cotonnier ou de légumineuses. Ils sont indépendants du type de sol et de son état. La pluie diminue l'efficacité de ces herbicides à pénétration foliaire, épandus sur le feuillage, par entraînement du dépôt. Le délai nécessaire entre la pulvérisation et la pluie dépend du produit et de l'intensité de la pluie. Par ailleurs, la détermination de la date d'application est parfois difficile [8].

C - Les herbicides totaux:

Ce sont les plus répandus; ce sont des produits de post-levée des mauvaises herbes. Ils peuvent être employés à diverses périodes du cycle cultural, en traitement en plein ou en localisé si la culture n'est pas installée, en traitement dirigé en cours de culture. Le choix des produits dépend des espèces à détruire [8]:

*En cas d'infestation par des espèces vivaces comme Cynodon dactylon(une espèce de plantes herbacées de la famille des Poaceae d'origine européenne utilisée pour la confection de gazon), Imperata cylindrica(une espèce de plante herbacée de la famille des Poaceae appelée aussi paillote ou herbe sanglante, avec un feuillage persistant, vert acide à vert franc virant progressivement au rouge sang à rouge cramoisi de haut en bas, nervure médiane plus claire), ou par Cyperus esculentus (Cyperus est un genre de la famille des Cyperaceae. Il regroupe des plantes communément appelées papyrus ou souchets. Ce sont des plantes aquatiques, originaires des régions tropicales et subtropicales. Il compte 400 espèces qui, pour la plupart, poussent en terrain marécageux. Le plus connu est le célèbre papyrus, plante importante pour les Égyptiens, qui en faisaient des feuilles pour écrire), ce sont des produits systémiques comme le glyphosate ou le sulfosate qu'il faut employer [8].

*Si la flore n'est constituée que d'espèces annuelles comme Digitaria horizontalis (herbe annuelle, en touffe étalée, pouvant atteindre 70 cm, de couleur gris ver, poussant dans les jardins et au bord de routes), Tridax procumbens (type biologique annuelle, 30 cm de taille couleur blanc-jaune vif), etc..., les produits de contact, comme le paraquat ou le glufosinate ammonium seront suffisants [8].

1-2-2-4- Modes d'action des herbicides :

Les modes d'action des herbicides sont fondés sur :

• la perturbation de la photosynthèse,

• l'inhibition de la synthèse des lipides,

• l'inhibition de la synthèse des acides aminés,

• la perturbation de la régulation de l'auxine,

• l'inhibition de la division cellulaire à la métaphase (phytohormone de croissance végétale disponible au développement des plantes),

• l'inhibition de la synthèse des caroténoïdes (pigments protecteurs des chlorophylles),

• l'inhibition de la synthèse de l'enzyme PPO (protoporphyrinogène oxydase) synthèse des chlorophylles**,**

• la dérégulation des pH entre les différents compartiments cellulaires ou découplant**,**

• la perturbation de la croissance [6] **.**

1-2-2-5 - Principales familles d'herbicides :

Les herbicides sont des produits aux structures chimiques complexes. Bien que chaque produit ait ses propriétés particulières, les herbicides d'une même famille présentent des structures chimiques semblables et de nombreuses caractéristiques.

A- les herbicides minéraux :

Ils furent surtout utilisés au début du siècle. Les plus utilisés actuellement sont :

• Le cyanure de calcium ($Ca(CN)_2$), il rentre par les racines et pénètre la sève brute pour ensuite s'accumuler dans les feuilles .

• Le sulfate de fer ($FeSO_4$), herbicide de contact utilisé pour lutter contre les mousses et qui accélère de plus l'humification des déchets végétaux,

• Le chlorate de sodium ($NaClO_3$) qui détruit les plantes à fort enracinement. Oxydant puissant, le chlorate de soude pénètre principalement par les racines et est transporté par la sève brute vers les feuilles. Son action n'est pas sélective et peu perdurer jusqu'à six mois dans la terre. Il est détruit par le calcaire, les matières organiques et les corps réducteurs, il peut être aussi lessivé par les eaux d'infiltration. Il est peu toxique pour l'homme mais c'est un comburant (qui peut entrer dans la fabrication d'explosifs). Il peut être employé pour la dévitalisation des souches .Ce dernier produit, du fait de son danger (risque d'explosion) est de plus en plus remplacé par des substances organiques [6].

B - Les herbicides organiques :

Ils constituent la très large majorité des herbicides du marché actuel. Par commodité, on les regroupe suivant leur type de pénétration dans le végétal :

- **Le glyphosate** : est un désherbant total, c'est-à-dire un herbicide non sélectif, autrefois produit sous brevet, exclusivement par la société Monsanto à partir de 1974, sous la marque Roundup. Le brevet ayant expiré, d'autres sociétés produisent désormais du glyphosate. Le mécanisme d'action de ce pesticide est systémique. Il agit en bloquant l'enzyme EPSPS (enoyl pyruvyl shikimate 3-phosphate synthase responsable de la synthèse des acides aminés aromatiques) [6].

C - Les herbicides racinaires :

- **Les Dinitroanilines (toluidines) :**

Apparus en 1960, les dinitroanilines sont très peu solubles dans l'eau, ont une forte volatilité et sont souvent photodégradables : ce sont donc des produits à incorporer dans le sol, avant la mise en place de la culture.Ils agissent en stoppant la croissance des plantules peu après leur germination. Ils sont désignés sous le terme -impropre- "d'antigerminatif". Ce sont plus précisément des antimitotiques (perturbent la division cellulaire). Ils s'utilisent en pré-levée contre les graminées. Leur toxicité est faible et leur persistance varie selon la dose employée (quelques semaines à un an). Leur nom se termine par le vocable ˝line˝ [6].

Exemples : butraline

- **Les Urées substituées (NH_2-CO-NH_2) :**

Ce sont exclusivement des herbicides. Leur absorption est essentiellement racinaire. Véhiculés par la sève brute, ils s'accumulent dans les feuilles où ils inhibent la photosynthèse. Ils ont une très faible solubilité dans l'eau et présentent une assez longue persistance d'action dans le sol (2 à 3 mois) mais variable selon les conditions écologiques rencontrées (sol, pluie, température). Ils ont une bonne action sur les graminées et sur certaines dicotylédones. Ils sont utilisés en pré ou post-levée. Leur toxicité est quasiment nulle. Leur nom se termine par le vocable "uron " [6].

Exemples : linuron…

- **Les Triazines :**

Ce groupe présente une structure cyclique. Ils agissent en bloquant la photosynthèse. Ils pénètrent par absorption radiculaire et sont véhiculés par la sève brute. Ils sont appliqués directement sur le sol. Le maïs est une plante très tolérante à ces composés. Le sorgho est également tolérant mais le blé et le soja y sont sensibles.

Leur toxicité est faible et leur sélectivité souvent bonne. Leur solubilité dans l'eau est réduite et sont donc peu entrainés dans le sol. Leur persistance peut ainsi atteindre 6 à 12 mois pour certains [6].

Exemples : simazine…

D - Herbicides racinaires et foliaires:

- **Les Imidazolinones :**

Certains produits de cette famille sont des herbicides totaux, d'autres sont sélectifs. Étant absorbés par voies foliaire et racinaire, ils sont indépendants des conditions climatiques. Ils agissent en bloquant l'activité de l'enzyme AHAS (l'enzyme acétohydroxyacide synthase) catalyse la première étape de la biosynthèse des 3 acides aminés essentiels : la valine, la leucine et l'isoleucine. Ceci empêche la plante de croître et entraîne une sénescence prématurée. Ce mode d'action explique le peu de toxicité de ces substances à l'égard des animaux et de l'homme, vu que ces derniers ne peuvent synthétiser ces acides aminés. Utilisés sur céréales ou en désherbage total, ils sont très souples à l'emploi. Leur persistance est de plusieurs mois [6].

Exemples : imazapyr**…**

- **Les sulfonylurés :**

Ils agissent sur la même enzyme que les imidazolinones[6].

Exemples : Amidosulfuron…

- **Les Dyphényls-éthers**

Synthétisées à partir de 1964, ces molécules possèdent 2 noyaux benzènes reliés par un oxygène. Ils sont absorbés par les feuilles et les racines. Leur transport dans la plante est très limité, ils ont une action de contact. Ils ont un effet inhibiteur sur la croissance des méristèmes et sont de ce fait généralement utilisés en prélevée ou en post-levée précoce contre les graminées. Ils inhibent également la respiration. Leur solubilité dans l'eau est faible et ils persistent dans les sols de 2 à 4 mois. Leur toxicité vis-à-vis des mammifères est faible. Leur nom se termine généralement par le vocable "fène"[6].

Exemples : Aclonifen ...

E - Herbicides foliaires :

- **Les Phytohormones de Synthèse**

Connus en 1942, ils sont absorbés par le feuillage et véhiculés par la sève. Leur causticité est nulle. Il en existe 2 grands groupes :

* Le premier dérive de l'acide β- indole acétique (acide indole 3-acétiqueAIA), hormone de croissance des végétaux. Ils entraînent une croissance anormale de la plante (dicotylédone), débouchant sur la mort.

Le plus connu est le 2,4-D (acide dichloro 2,4 phénoxyacétique), très utilisé pour le désherbage sélectif des monocotylédones qui y sont peu sensibles, à la différence des dicotylédones.

* les composés dérivant des acides propionique et butyrique. Ils sont absorbés par le feuillage et s'accumulent dans les zones à divisions cellulaires intensives (méristème, bourgeon, racine) où ils provoquent une croissance anormale. Leur persistance dans les pailles interdit l'usage de ces dernières en horticulture [6].

Exemples : 2,4-D...

- **Colorants nitrés (dérivés du phénol, dinitrophénol) :**

Dérivé du benzène, ce groupe comprend des molécules toxiques pour les animaux (insecticide) et les végétaux. Ils sont de couleur jaune. Ils ont été très utilisés contre une large gamme de dicotylédones au stade plantule, pour la protection des céréales en traitement de post-levée. Ce sont des herbicides de contact à action rapide entraînant des nécroses sur les tissus qui se dessèchent et meurent. Ils agissent sur les membranes cellulaires qu'ils perméabilisent aux ions H^+, abaissant fortement le pH des cellules. Ils ne se déplacent pas dans la plante, seules les parties touchées seront affectées par l'herbicide par l'apparition de brûlures au point d'impact.

Ils sont dangereux pour l'homme et l'environnement de par leur toxicité élevée. Le DNOC (Dinitro-Ortho-Crésol interdit par l'union européenne depuis 1999) à l'état sec, présente de plus des risques d'explosion. Les colorants nitrés sont actuellement remplacés par des produits plus sélectifs [6].

Exemples : DNOC...

- **Les Carbamates** :

Conçus en 1945 pour la destruction des graminées, ces herbicides se subdivisent en 4 catégories:

- 1) Les dérivés de l'acide carbamique (NH_2-COOH) qui agissent sur la division cellulaire.

- 2) Les dérivés de l'acide thiocarbamique (NH_2-CO-SH) qui inhibent la synthèse des lipides à longue chaîne et des gibbéréllines.

- 3) Les dérivés de l'acide dithiocarbamique (NH_2-CS-SH) qui empêchent la germination.

- 4) Les biscarbamates qui empêchent la photosynthèse.

Ces herbicides ont en commun leur faible toxicité et une volatilité plus ou moins grande. Ils perturbent la division cellulaire (antimitotique) et la physiologie générale de la plante, provoquant le phénomène d'anse en panier (les feuilles ne pouvant pas se déplier).

Ils s'emploient le plus souvent en pré-levée (thiocarbamates) ou post-semis, parfois en post-levée (phenmediphame). A l'exception des composés allates, qui persistent plusieurs mois dans le sol, leur persistance est quasiment nulle [6].

Citons quelques exemples sur les 4 catégories précédentes :

1) -Dérivés de l'acide carbamique : Clorobufame …

2) - Thiocarbamates : Butilate…

3) - Dithiocarbamates : Nabame...

4) - Biscarbamates : Phenmédiphame…

- **Les Ammoniums Quaternaires (Bipyridiles)** :

Synthétisés dans les années 50, ils sont formés par l'association de 2 cycles pyridilyques. Ce sont des accepteurs d'électrons photosynthétiques, actifs sur les réactions lumineuses de la photosynthèse, provoquant l'arrêt de l'assimilation de CO_2. Ils provoquent également la dégradation des acides gras insaturés, l'ensemble de ces actions débouchant sur la mort. Ils se caractérisent par leur rapidité d'action et leur absence de sélectivité (désherbant total), à l'exception du difenzoquat. Ils pénètrent dans les organes aériens mais migrent peu.

Ce sont avant tout des produits de contact. Ils sont très solubles dans l'eau et n'ont pas d'effet par traitement de sol car ils sont fortement absorbés par les argiles où, de ce fait, ils ne se dégradent que très lentement. Ils sont très toxiques pour l'homme et les animaux du fait de l'absence d'antidote [6].

Exemples : Difenzoquat…

- **fop/dime et pinoxadène :**

Ce sont des herbicides antigraminés qui inhibent l'ACCase (enzyme acetyl-CoA carboxylase) dans les chloroplastes des monocotylédones.

De nombreuses résistances sont apparues, pour quelques unes liés à des modifications de l'enzyme cible, mais, pour la plupart dues à d'autres mécanismes [6].

Exemples : Alloxydime-Sodium…

I-3 - LA DOSE LETALE:

I-3 -1 - Définition:

La dose létale (DL) est une indication de la létalité d'une substance ou d'un type donné de radiation ou la radiorésistance. Puisque la résistance est variable d'un individu à l'autre, la dose létale représente la dose à laquelle un pourcentage donné d'une population donnée

décède. Cette dose est habituellement exprimée en unités de masse de substance par masse corporelle, c'est-à-dire en g/kg.

La dose létale est souvent utilisée pour décrire la puissance des venins chez les animaux comme pour les serpents.

Les mesures de dose létale sur des animaux ont beaucoup été utilisées dans la recherche sur les drogues, même si désormais la plupart des chercheurs préfèrent ne pas y avoir recours.

La dose létale dépend non seulement de l'espèce de l'animal, mais aussi du mode d'administration (oral, inhalation, contact, etc.). Ainsi, une substance donnée nécessite une dose plus petite en cas d'injection ou d'inhalation qu'en cas d'ingestion .L'indicateur de létalité le plus utilisé est la dose létale 50 ou (DL_{50}) [9].

I-3 -2- Dose létale 50 :

I-3 -2-1 - Définition :

La dose létale 50 ou DL_{50} (en anglais Lethal Dose 50 ou LD_{50}) ou CL_{50} (concentration létale 50) est un indicateur quantitatif de la toxicité d'une substance. Cette notion s'applique également aux irradiations.

Cet indicateur mesure la dose de substance causant la mort de 50 % d'une population animale donnée (souvent des souris ou des rats) dans des conditions d'expérimentation précises. C'est la masse de substance nécessaire pour tuer 50 % des animaux dans un lot. Elle s'exprime en milligrammes de matière active par kilogramme d'animal. Plus ce chiffre est petit, plus la substance est toxique. Cette dose n'est valable que pour une espèce précise (le plus souvent le rat) et un mode d'introduction précis dans l'organisme (ingestion, inhalation, application cutanée).

Notons toutefois que la DL_{50} peut varier, parfois fortement, en fonction du solvant utilisé ainsi qu'en fonction du sexe de l'animal. Ces chiffres ne sont pas directement extrapolables à l'homme [10].

La DL_{50} est le plus souvent exprimée pour une ingestion orale chez le rat. La DL_{50} mesurée par application dermale chez le lapin donne une information complémentaire [10].

I-3 -3 - Historique:

Le concept de dose létale 50 a été inventé par J.W. Trevan en 1927 et permet de classifier tous les produits par dangerosité à court et moyen termes.

L'OCDE (ligne directrice pour les essais 401) a fait de la DL_{50} un test officiel en 1981. En 1987, elle a réduit à 20 au lieu de 30 le nombre minimal d'animaux que doit contenir l'échantillon testé. En 2001, elle a approuvé trois nouvelles méthodes, destinées à remplacer la DL_{50} et à occasionner une moindre souffrance animale.

La ligne directrice 401 a finalement été abrogée par l'OCDE, le 17 décembre 2002 [10].

I-3 -4 - Choix de la dose létale 50 :

C'est pour des raisons de représentativité statistique qu'on utilise la valeur 50 %, plutôt que 0 %, 5 %, 95 %, ou 100 %. En effet, la courbe de Gauss est « plate » vers 50 %, ce qui fait qu'un échantillon est plus représentatif lorsqu'un seuil est franchi à 50 %.

Autour de 50 % de mortalité, de toutes petites variations de dose donneront de grandes variations dans le pourcentage de morts, ce qui a fait retenir cette valeur clé de 50% [10].

I-3 -5 - Interprétation et classes de toxicité :

La DL_{50} s'exprime en unités de masse de substance par masse corporelle, c'est-à-dire en g/kg.La DL_{50} permet de mesurer la toxicité d'une substance et d'établir des classes de toxicité en deux échelles montrées dans le tableau - **1** suivant :

Tableau 1 - Deux échelles des classes de toxicité [10] :

Classes de toxicité : Échelle de Gosselin, Smith et Hodge		Classes de toxicité : Échelle de Hodge et Sterner	
Dose orale probablement mortelle (humain)	**Classe de toxicité**	**Indice de toxicité**	**DL_{50} orale (rat)**
Moins de 5 mg/kg	Super toxique	1=extrêmement toxique	Jusqu'à 1 mg/kg
De 5 à 50 mg/kg	Extrêmement toxique	2 = hautement toxique	De 1 à 50 mg/kg
De 50 à 500 mg/kg	Très toxique	3 = modérément toxique	De 50 à 500 mg/kg
De 50 à 5000mg/kg	Modérément toxique	4 = légèrement toxique	De 500 à 5 000 mg/kg
De 5000 à 15000 mg/kg	Légèrement toxique	5 = Presque pas toxique	De 5000 à 15000 mg/kg

I -3 -6- Utilisation :

I -3 -6 -1 - Identification de la toxicité :

Elle sert à mesurer toute la toxicité d'une substance, mesures qui s'effectuent via des études qualitatives (non mesurables) et quantitatives (mesurables dont la DL_{50}).

La DL_{50} sert souvent de départ aux études de toxicité car elle fournit un minimum de connaissances en identifiant les symptômes de l'intoxication et la dose toxique. Il faut malgré tout la considérer avec prudence car c'est souvent une étude préliminaire (première

analyse) qui peut être influencée par plusieurs facteurs, tels l'espèce animale, le sexe, l'âge, le moment de la journée, etc.

Elle a cependant une valeur limitée, car elle ne concerne que la mortalité, d'où l'apparition de valeurs comme l'IC_{50} (la dose dite semi incapacitante, à laquelle 50 % de la population exposée est paralysée ou assommée).

Il existe d'autres méthodes d'étude de la toxicité, par exemple les tests d'irritation de la peau et de corrosion des yeux, qui font généralement partie d'un programme d'évaluation toxicologique [10].

I-3 -6-2 - Identification du pouvoir pathogène :

La DL_{50} est une donnée servant à mesurer le pouvoir pathogène d'un germe [10].

I-3 -7 - Notions voisines :

S'il s'agit d'une substance inhalée, on parle de concentration létale 50 (CL_{50} ou CLt_{50}) pour exprimer la concentration du toxique dans l'air inspiré et causant la mort de 50 % des animaux. La CL_{50} est exprimée en mg·min/m³.Sur le même modèle, on parle aussi de l'IC_{50} ou ICt_{50}. On parle aussi parfois de :

- DL_{01} : dose de substance causant la mort de 1 % de la population des animaux d'essai.
- DL_{100} : dose de substance causant la mort de 100 % de la population des animaux d'essai .
- $DL_{min.}$: dose de substance la plus faible causant la létalité .
- $DT_{min.}$: dose de substance la plus faible causant un effet toxique [10].

Remarque 1 : Nous considérons la dose létale 50 %(DL50), dont le logarithme de l'inverse,

$$pDL_{50} = \log\left(1/DL_{50}\right) \qquad (1)$$

, servira d'indicateur de toxicité dans cette étude.

Remarque 2 : La structure, nom et N° de CAS des composés soulignés de cette partie figurent dans l'annexe 2.

PARTIE THEORIQUE

Les relations quantitatives structures / activités, désignées par l'abréviation QSAR (Quantitative Structure/ activity Relationships), constituent des modèles mathématiques pour l'approximation des relations, souvent complexes, entre la structure moléculaire caractérisée par des descripteurs moléculaires et la dose létale des composés.

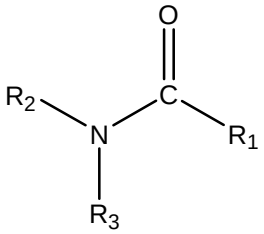
Le but fondamental des processus QSAR est d'étudier le rapport entre une variable dépendante et une ou plusieurs variables indépendantes,

Sa mise en œuvre comprend plusieurs étapes:

- La collecte des données ;
- L'optimisation de la géométrie moléculaire ;
- La génération des descripteurs moléculaires ;
- La sélection d'un sous-ensemble de descripteurs significatifs ;
- Le développement du modèle ;
- Et, finalement, l'évaluation des performances de ce modèle.

II-1 -COLLECTE DES DONNEES :

Notre travail consiste à prédire la toxicité d'une série d'amides herbicides.ayant le squelette de base suivant:



Où R_1, R_2, R_3 sont des substituants, concernent 50 composés d'amides herbicides; prélevés des articles [11, 12] et réunis dans le tableau - 2.

50 observations de ce tableau ont été scindées aléatoirement en deux ensembles disjoints de 40 éléments et 10 éléments [les 40 composés pour l'ensemble d'estimation qui sert à la construction du modèle) et les 10 restants (10 éléments =50-40), pour la validation externe]. Les éléments de l'ensemble d'estimation sont numérotés, et ceux de l'ensemble de validation externe portent les numéros avec un astérisque*, alors que ceux du test de RNA forment une série de 8 composés sont désignés par le double astérisque ** dans le tableau 2.

Tableau 2 - Nomenclature et valeurs des pDL_{50} des composés étudiés [11,12].

N°	Noms des composés	pDL_{50}
1**	2,6-di-tert-butyl-4-methylphenyl methylcarbamate	**2,1**
2	1-(Eethylamino)-1-oxopropan-2-yl phenylcarbamate	**1,67**
3	1-(Eethylamino)-1-oxopropan-2-yl phenylcarbamate	**1,67**
4*	(3-Phenylcarbamoyloxy-phenyl)-carbamic acid ethyl ester	**2,42**
5	2-Methyl-pentanoic acid (3,4-dichloro-phenyl)-amide	**1,58**
6*	2-Methyl-pentanoic acid (3-chloro-4-methyl-phenyl)-amide	**0,53**
7	2,6-dimethoxy-N-(3-(3-methylpentan-3-yl)isoxazol-5-yl)benzamide	**1,48**
8*	3,5-dichloro-N-(3-methylbuta-1,2-dienyl)benzamide	**1,59**
9**	2-(naphthalen-2-ylcarbamoyl)benzoic acid	**1,45**
10	Carbamic acid, (3-methylphenyl)-, 3-[(methoxycarbonyl)amino]phenyl ester	**1,43**
11	2-Bromo-3,3-dimethyl-N-(2-phenylpropan-2-yl)butanamide	**1,20**
12*	Methyl 2-(N-(3-Chloro-4-Fluorophenyl)benzamido)propanoate	**1,17**
13	Methyl 3,4-diChlorophenylcarbamate	**0,38**
14*	Isopropyl 2-(N-(3-Chloro-4-Fluorophenyl)benzamido)propanoate	**1,05**
15	N-(4-Chlorophenyl)-2,2-dimethylpentanamide	**1,22**
16*	[4-(Ethyl-phenyl-carbamoyloxy)-phenyl]-carbamic acid isopropyl ester	**1,07**
17**	Isopropyl 3-chlorophenylcarbamate	**1,25**
18	S-ethyl cyclohexyl(ethyl)carbamothioate	**1,22**
19*	N-isobutyl-2-oxoimidazolidine-1-carboxamide	**1,13**
20*	N-(butoxymethyl)-2-Chloro-N-(2,6-diethylphenyl)acetamide	**1,02**
21	2-Chloro-N-(2-ethyl-6-methylphenyl)-N-(1-methoxypropan-2-yl)acetamide	**0,99**
22	S-2,3,3-triChloroallyl diisopropylcarbamothioate	**0,74**
23	N-(5-Chloro-4-methyl-4,5-dihydrothiazol-2-yl)propionamide	**1,00**
24**	S-propyl dipropylcarbamothioate	**0,94**
25	2-Chloro-N-(2,6-diethylphenyl)-N-(methoxymethyl)acetamide	**0,75**
26	S-Ethyl dipropylcarbamothioate	**0,94**
27	Ethyl 2-(N-(3,4-diChlorophenyl)benzamido)propanoate	**0,63**
28	S-4-Chlorobenzyl diethylcarbamothioate	**0,70**
29**	N-(but-3-yn-2-yl)-2-Chloro-N-phenylacetamide	**0,73**
30	S-2-Chlorobenzyl diethylcarbamothioate	**0,63**

Tableau 2 : Suite et fin

N°	Noms des composés	pDL50
31	S-2-Chloroallyl diethylcarbamothioate	**0,61**
32*	N,N-diallyl-2-Chloroacetamide	**0,62**
33	2-Chloro-N-isopropyl-N-phenylacetamide	**0,53**
34	(2-Benzamidooxy)acetic acid	**1,46**
35*	N-(3,4-diChlorophenyl)-N-(dimethylcarbamoyl)benzamide	**1,17**
36*	N,N-diallyl-2,2-diChloroacetamide	**0,98**
37	2,2-diChloro-N-(3-Chloro-1,4-diOxo-1,4-dihydronaphthalen-2-yl)acetamide	**1,67**
38	Propan-2-one O-phenylcarbamoyl oxime	**0,90**
39	Isopropyl phenylcarbamate	**1,45**
40	But-3-yn-2-yl 3-Chlorophenylcarbamate	**1,03**
41*	3,4-diChlorobenzyl methylcarbamate	**0,92**
42	3,5-diChloro-N-(3-methylbuta-1,2-dienyl)benzamide	**1,51**
43	N,N-diethyl-3-methylbenzamide	**1,02**
44	Ethyl 2-(2-Chloro-N-(2,6-diethylphenyl)acetamido)acetate	**0,87**
45	N-(3,4-diChlorophenyl)propionamide	**0,81**
46*	S-propyl butyl(ethyl)carbamothioate	**0,77**
47	2-(5-Chloropent-2-ynyloxy)-N-(3-Chlorophenyl)acetamide	**0,70**
48	N,N-dimethyl-2,2-diphenylacetamide	**0,62**
49	S-ethyl diethylcarbamothioate	**0,39**
50	(E)-S-2,3-diChloroallyl diisopropylcarbamothioate	**0,16**

***** Composés de validation externe (RLM et RNA).

****** Composés tests pour RNA

II-2 : OPTIMISATION DE LA GEOMETRIE MOLECULAIRE ET GENERATION DES DESCRIPTEURS MOLECULAIRES :

Les structures des molécules ont été obtenues à l'aide du logiciel de modélisation moléculaire Hyperchem 6.03 [13], et les géométries finales à l'aide de la méthode semi empirique AM1 du même logiciel. Tous les calculs ont été menés dans le cadre du formalisme RHF sans interaction de configuration. Les structures moléculaires ont été optimisées à l'aide de l'algorithme Polak-Ribiere avec pour critère une racine du carré moyen du gradient égale à 0,001 kcal.mol^{-1}. Les géométries obtenues ont été transférées dans les logiciels DRAGON [14] utilisés pour le calcul de 1664 descripteurs appartenant à 20 classes différentes.
Les descripteurs d'un même groupe, à valeur constante (écarts types inférieurs à 0,001), et ceux hautement corrélés ($R \geq 0,95$) ont été exclus

Plutôt que de rechercher à expliquer la variable dépendante (propriété biologique considérée) par tous les régresseurs (descripteurs moléculaires), on peut chercher seulement un ensemble réduit de régresseurs qui donne une reconstitution aussi satisfaisante de la variable à expliquer. Parmi les stratégies mises en œuvre pour la sélection d'un ensemble réduit de variables explicatives, on peut citer : les méthodes de pas à pas (méthode descendante; méthode ascendante, et méthode dite stepwise), ainsi que les algorithmes évolutifs et génétiques.

En général, la comparaison se fait à l'avantage des algorithmes génétiques (AG) que nous avons appliqués dans le présent travail, et que nous rappelons succinctement.

II-3-SELECTION D'UN SOUS-ENSEMBLE DE DESCRIPTEURS SIGNIFICATIFS :

II-3-1- Principe :

Dans la terminologie des algorithmes génétiques, le vecteur binaire I **,** appelé "chromosome", est un vecteur de dimension p où chaque position (un "gène") correspond à une valeur (1 si elle figure dans le modèle, 0 sinon). Chaque chromosome représente un modèle basé sur un ensemble de variables explicatives.

On commence par définir le paramètre statistique à optimiser (par exemple maximiser Q^2 en utilisant la validation croisée par "leave-one-out"), avec la taille P de la population du

modèle (par exemple, P = 100), et le nombre maximum de variables L permises pour le modèle (par exemple, L = 10) ; le minimum de variables permises est généralement supposé égal à 1. De plus, une probabilité de croisement p_c (habituellement élevée, $p_c > 0,9$), et une probabilité de mutation p_M (habituellement faible, $p_M < 0,1$) doivent être également définies.

Après définition des principaux paramètres, la mise en œuvre de l'algorithme génétique est démarrée, son évolution comprend trois étapes principales.

II-3-2- Initialisation aléatoire du modèle :

La population est constituée au départ de modèles aléatoires avec des variables comprises entre 1 et L, puis les modèles sont ordonnés eu égard au paramètre statistique sélectionné – la qualité du modèle – (le meilleur modèle est en première position, le plus mauvais en position P) ;

II-3-3 - Etape de croisement :

A partir de la population, on sélectionne des paires de modèles (aléatoirement, ou avec une probabilité proportionnelle à leur qualité). Puis, pour chaque paire de modèles on conserve les caractéristiques communes, c'est-à-dire les variables exclues dans les 2 modèles restent exclues, et les variables intégrées dans les 2 modèles sont conservées. Pour les variables sélectionnées dans un modèle et éliminées dans l'autre, on en essaye un certain nombre au hasard que l'on compare à la probabilité de croisement p_c : si le nombre aléatoire est inférieur à la probabilité de croisement, la variable exclue est intégrée au modèle et vice versa. Finalement, le paramètre statistique du nouveau modèle est calculé : si la valeur de ce paramètre est meilleure que la plus mauvaise pour la population, le modèle est intégré à la population, à la place correspondant à son rang ; dans la cas contraire, il n'est pas pris en compte. Cette procédure est répétée pour de nombreuses paires (100 fois par exemple).

II-3- 4 - Etape de mutation :

Pour chaque modèle de la population (c'est-à-dire pour chaque chromosome) p nombres aléatoires sont éprouvés, et, un à la fois, chacun est comparé à la probabilité de mutation, p_M, définie : chaque gène demeure inchangé si le nombre aléatoire correspondant excède la probabilité de mutation, dans le cas contraire, on le change de 0 à 1 ou vice versa.

Les faibles valeurs de p_M permettent uniquement peu de mutations, conduisant à de nouveaux chromosomes peu différents des chromosomes générateurs.

Après obtention du modèle transformé, on en calcule le paramètre statistique : si cette valeur est meilleure que la plus mauvaise de la population, le modèle est intégré à la population, à la place correspondant à son rang ; dans le cas contraire il n'est pas pris en compte.

Cette procédure est répétée pour tous les chromosomes, c'est-à-dire P fois.

II -3-5 - Conditions d'arrêt :

Les étapes précédentes (croisement et mutation) sont répétées jusqu'à la rencontre d'une condition d'arrêt (par exemple un nombre maximum d'itérations défini par l'utilisateur), ou qu'il est mis fin arbitrairement au processus.

Une caractéristique importante de la sélection d'un ensemble réduit de variables par algorithme génétique est qu'on n'obtient pas nécessairement un modèle unique, mais le résultat consiste habituellement en une population de modèles acceptables ; cette caractéristique, parfois considérée comme un désavantage, fournit une opportunité pour procéder à une évaluation des relations avec la réponse à différents points de vue.

Notons que la taille du modèle est déterminée en maximisant la valeur de la fonction FIT de KUBINYI [15] pou un nombre de descripteur=8, et le FIT peut être aussi calculé selon:

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

p désignant le nombre de variables du modèle et R^2 le coefficient de détermination.
Ce critère permet de comparer entre modèles construits sur un même nombre n de données, mais avec un nombre de variable p différent.

II - 4 - DEVELOPPEMENT DES MODELES :

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

Nous avons utilisé la RLM et les réseaux de neurones artificiels (RNA). En imposant des transformations linéaires entre descripteurs moléculaires et propriétés étudiées, la RLM peut influencer négativement les capacités prédictives du modèle. Par contre, avec les réseaux de neurones il n'est nul besoin de postuler un modèle. Les réseaux de neurones ont la capacité de représenter n'importe quelle dépendance fonctionnelle qu'ils découvrent par eux-mêmes.

Ainsi, la découverte et l'exploitation des dépendances non-linéaires de haut niveau peuvent améliorer la capacité de prédiction de la variable d'intérêt.

II - 4 -1 - La régression linéaire multiple (RLM) :

Supposons qu'on ait mesuré sur n individus et p variables représentées par des vecteurs de $\Re^n : y, x_1, x_2, ..., x_k$; y est la variable dépendante ou à expliquer (propriété physique ou activité biologique d'intérêt) et les x_j les variables explicatives ou encore prédicteurs (descripteurs moléculaires). On cherche alors à reconstruire y au moyen des x_j par une formule linéaire.

On pose :

$$\boldsymbol{Y} = \beta_0 \mathbf{1} + \mathbf{X}(j)\beta(j) + \varepsilon(\mathbf{j}) \qquad (3)$$

Y est un vecteur de dimension n contenant la propriété étudiée des amides herbicides considérés, 1 est un vecteur unité, c'est-à-dire une matrice colonne formée d'éléments égaux à 1, **X**(j) indique la matrice (n×j), et $\varepsilon(j)$ correspond aux résidus qui doivent suivre une distribution Normale, posséder une espérance mathématique nulle et une matrice de dispersion I σ^2 [16]. Les estimateurs {β} sont calculés en utilisant la technique des moindres carrés ordinaires.

II - 4 -2 - Les réseaux de neurones :

Les réseaux de neurones ont été étudiés depuis les années 40 [17]. Les idées de base de cette technique viennent de la recherche cognitive, d'où vient le nom 'réseaux de neurones'.

La technique inspirait beaucoup de chercheurs à cette époque, mais beaucoup de l'intérêt disparaît après un article de Minsky et Papert [18], finalement relancée au début des années 80 après un quasi-oubli d'une vingtaine d'années. La cause de l'intérêt soudain était l'apparition de nouvelles architectures de réseaux de neurones.

II-4 - 2 -1 - Le neurone artificiel :

L'élément de base d'un réseau de neurones est, bien entendu, le neurone artificiel. Un neurone (figure 1) contient deux éléments principaux :

- Un ensemble de poids associés aux connexions du neurone, et une fonction d'activation (Figure 2).

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

.

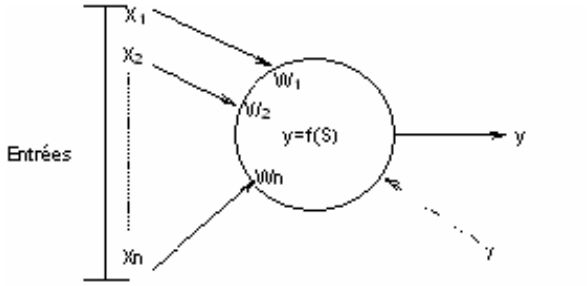
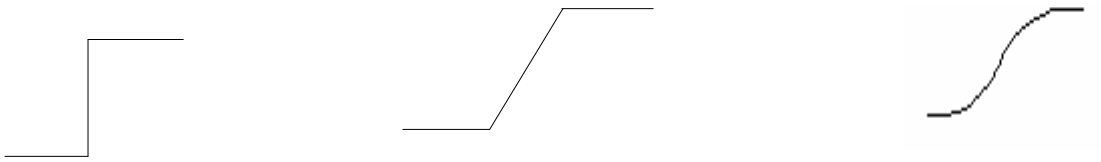


Figure – 1 le neurone artificiel générique.

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



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

Figure – 2 Fonctions d'activation.

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

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

II- 4 - 2 - 2 - Propriétés des réseaux de neurones :

Un réseau de neurones se compose de neurones qui sont interconnectés de façon à ce que la sortie d'un neurone puisse être l'entrée d'un ou plusieurs autres neurones. Ensuite il y a des entrées de l'extérieur et des sorties vers l'extérieur [19].

Rumelbart et al donnent huit composants principaux d'un réseau de neurones [19] :

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

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

Les réseaux de neurones sont souvent appelés des "boites noires" car la fonction mathématique qui est représentée devient vite trop complexe pour l'analyser et la comprendre directement. Cela est notamment le cas si le réseau développe des représentations distribuées [19], c'est-à-dire que plusieurs neurones sont plus ou moins actifs et contribuent à une décision. Une autre possibilité est d'avoir des représentations localisées, ce qui permet d'identifier le rôle de chaque neurone plus facilement. Les réseaux de neurones ont quand même une tendance à produire des représentations distribuées.

II-4 - 2 -3 - Les différents types de réseaux de neurones :

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

- Le réseau de Hopfield (et sa version incluant l'apprentissage, la machine de Boltzmann).
- Les cartes auto-organisatrices de Kohonen.
- Les réseaux à fonction radiale que l'on nomme aussi RBF (pour " Radial Basic Functions ").
- Les réseaux multicouches ou perceptron multicouches PMC

Le réseau de Hopfield [20] est un réseau avec des sorties binaires où tous les neurones sont interconnectés avec des poids symétriques. C'est-à-dire que le poids du neurone N_i au neurone N_j est égal au poids du neurone N_j au neurone N_i. Les poids sont donnés par l'utilisateur. Les poids et les états des neurones permettent de définir l'«énergie» du réseau.

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

Les cartes auto-organisatrices de Kohonen [21] sont utilisées pour faire des classifications automatiques des vecteurs d'entrées.

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

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

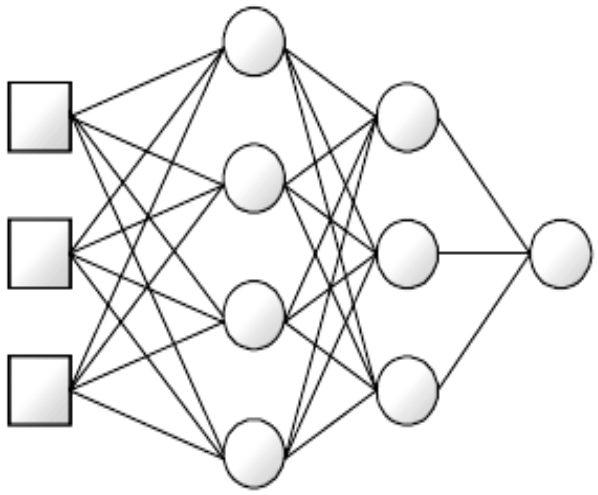
A- Les réseaux multicouches ou perception multicouches (PMC) :

Les réseaux multicouches (PMC) (figure 3) se composent des entrées, une couche de sortie et zéro ou plusieurs couches cachées [19]. Les connexions sont permises seulement d'une couche inférieure (plus proche des entrées) vers une couche supérieure (plus proche de la couche de sortie). Il est aussi interdit d'avoir des connexions entre des neurones de la même couche.

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

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

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



Les entrées Couches cachées Couche de sortie

Figure – 3 Structure générale du perceptron multicouche

Il a été prouvé qu'il existe toujours un réseau de neurones de ce type avec trois couches seulement (les entrées, couche de sortie et une couche cachée) qui peut approximer une fonction $f : [0.1]^n \Rightarrow R^n$ avec n'importe quelle précision $\varepsilon > 0$ désirée [22].

Un problème consiste à trouver combien de neurones cachés sont nécessaires pour obtenir cette précision. Un autre problème est de s'assurer a priori qu'il est possible d'apprendre cette fonction.

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

II-4 - 2 -4 – Apprentissage :

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

Il existe trois types d'apprentissages principaux .Ce sont l'apprentissage supervisé, l'apprentissage non supervisé,et l'apprentissage par tentative (graded training en anglais) [22].

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

Dans le cas de l'apprentissage non-supervisé le réseau décide lui-même quelles sont les bonnes sorties. Cette décision guidée par un but interne au réseau qui exprime une configuration idéale à atteindre par rapport aux exemples introduits. Les cartes auto-organisatrices de Kohonen sont un exemple de ce type de réseau [21].

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

Il existe plusieurs règles pour chaque type d'apprentissage. L'apprentissage supervisé est le type le plus utilisé. Pour ce type d'apprentissage la règle la plus utilisée est celle de Widrow-Hoff . D'autres règles d'apprentissage sont par exemple la règle de Hebb, la règle de perceptron, la règle de Grossberg etc …. [19, 21, 22].

A - L'apprentissage de Widrow-Hoff :

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

$$w_{k+1} = w_k - \alpha \delta_k x_k \qquad (4)$$

Où :

w_k est le poids à l'instant k ;

w_{k+1} le poids à l'instant k-1 ;

α est le facteur d'apprentissage ;

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

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

Ainsi, si δ_k et x_k sont positifs tous les deux, alors le poids doit être augmenté. L'ampleur du changement dépend avant tout de la grandeur de δ_k mais aussi de celle de x_k. Le coefficient α sert à diminuer les changements pour éviter qu'ils deviennent trop grands, ce qui peut entraîner des oscillations du poids.

Deux versions améliorées de cet apprentissage existent, la version 'par lois'et la version 'par inertie' (momentum en anglais) [22], dont l'une utilise plusieurs exemples pour calculer la moyenne des changements requis avant de modifier le poids et l'autre empêche que le changement du poids au moment k ne devienne beaucoup plus grand qu'au moment k-1.

B - L'apprentissage par rétro-propagation du gradient (Levenberg-Marquardt backpropagation) :

L'algorithme d'apprentissage par rétro-propagation du gradient (figure 4) est un algorithme itératif qui a pour objectif de trouver le poids des connexions minimisant l'écart commis par le réseau sur l'ensemble d'apprentissage. Cette minimisation par une méthode du gradient conduit à l'algorithme d'apprentissage par rétro-propagation.

La procédure d'apprentissage se décompose en deux étapes. Pour commencer, les valeurs d'entrées sont présentées au réseau, qui propage ensuite ces valeurs jusqu'à la couche de sortie et donne ainsi la réponse au réseau. A la deuxième étape les bonnes sorties correspondantes sont présentées aux neurones de la couche de sortie qui calculent l'écart, modifient leurs poids et rétro-propagent l'erreur jusqu'aux entrées pour permettre aux neurones cachés de modifier leurs poids de la même façon. Le principe de modification des poids est normalement l'apprentissage de Widrow-Hoff [23].

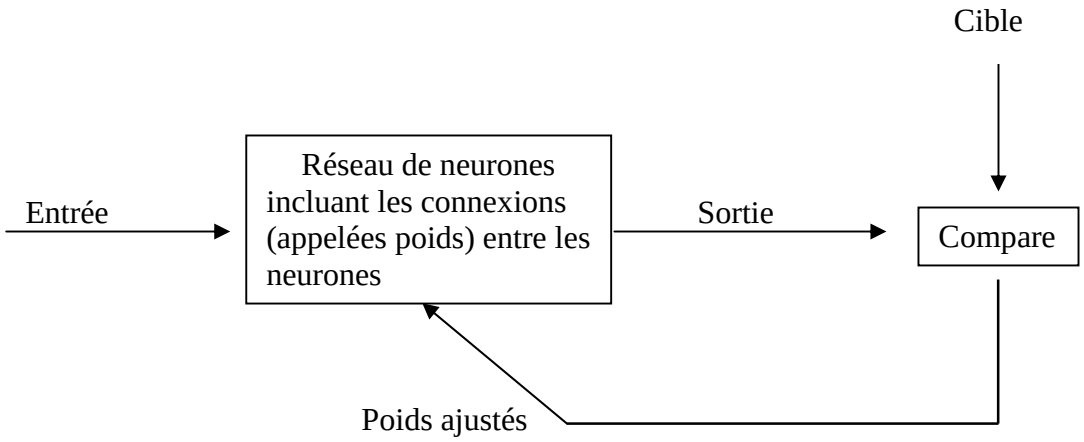


Figure – 4 Apprentissage par un algorithme de rétro-propagation

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

$$EQM = \frac{\sum_{i=1}^{n} \left(y_i - \hat{y}_i \right)^2}{n} \qquad (5)$$

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

II-4 - 2 -5 - Critères d'arrêt :

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

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

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

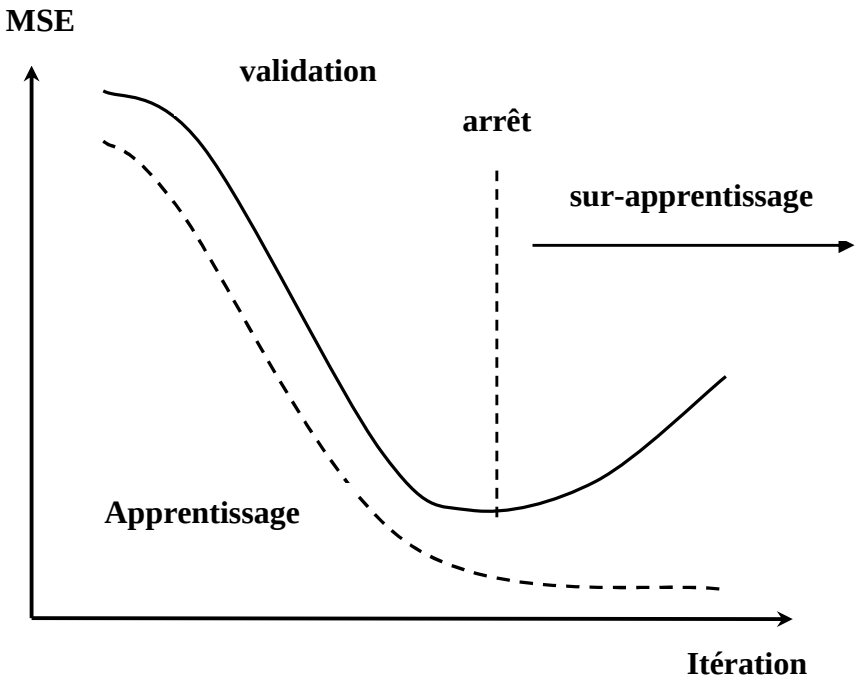


Figure – 5 Illustration de l'arrêt précoce

II-4 - 2 - 6 - Construction d'un modèle :

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

Les quatre grandes étapes de la création d'un réseau de neurones sont détaillées comme suit :

A - Construction de la base de données :

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

Afin de développer une application à base de réseaux de neurones, il est nécessaire de disposer de deux bases de données, une pour effectuer l'apprentissage et l'autre pour tester le réseau obtenu et déterminer ses performances. Pour cette raison nous avons partagé notre base des données (tableau 2) aléatoirement en trois sous-ensembles comme suit :

- Un ensemble de 32 composés pour l'apprentissage du réseau de neurones.
- Un deuxième de 10composés pour la validation externe.
- Et un troisième de 8 composés choisis aléatoirement de l'ensemble d'apprentissage pour le test.

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

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

$$x_{norm} = 2 \times \frac{\left(x_j - x_{\min}\right)}{\left(x_{\max} - x_{\min}\right)} - 1 \qquad (6)$$

où :

x_{norm} est la valeur normalisée ;

x_j est la $j^{ième}$ valeur ;

$x_{\max}$ est la valeur maximale ;

$x_{\min}$ est la valeur minimale.

B - Définition de la structure du réseau :

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

C - Nombre de couches et de neurones cachés :

Mis à part les entrées et la couche de sortie, il faut décider du nombre de couches cachées. Sans couche cachée, le réseau n'offre que de faibles possibilités d'adaptation. Néanmoins, il a été démontré qu'un Perceptron Multicouches avec une seule couche cachée pourvue d'un nombre suffisant de neurones, peut approximer n'importe quelle fonction avec la précision souhaitée [24].

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

D - Présentation de l'environnement utilisé :

Dans cette optique, le logiciel MATLAB [25], qui contient un module consacré au développement de réseaux

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

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

Les fonctions sigmoïde de transfert, tangente hyperbolique et linéaire, ont été adoptées comme fonctions d'activation pour les couches cachée et de sortie.

Nous présentons l'algorithme du réseau de neurones utilisé dans la page suivante :

Algorithme du réseau de neurones utilisé [26] **:**

```
P= [les descripteurs];
T= [la propriété étudiée];
N = 50 ;       % tous les composés
N1 = 40 ;      % Composés d'apprentissage
N2 = 10 ;      % Composés de validation
P0= (P)';      % Transposition de la matrice P
T0= (T)';      % Transposition de la matrice T
[pn,minP,maxP,tn,minT,maxT] = premnmx(P0,T0);     % Normalisation entre [-1,+1]
P1n= (pn)';
T1n= (tn)';
% Apprentissage
P1=Pn(1:N1,:)';     % Descripteurs normalisés d'apprentissage
T1=Tn(1:N1,:)';     % Propriété physique normalisée d'apprentissage
T10=T(1:N1,:)';
%  Test
[R,Q] = size(P1);
iitst = [1:5:Q ]; ];   % Choix aléatoire de 8 composés du test test.
test.P = P (:,iitst);
T20=T10 (:,iitst);
%  Validation
val.P = Pn(N1+1:N,:)';   % Descripteurs normalisés de validation
val.T = Tn(N1+1:N,:)';   % Propriété physique normalisée de validation
T30=T (N1+1:N, :)';
net = newff(minmax(P),[ S1 S2],{ TF1 TF2}, BTF);     % Création d'un réseau
 % S1 : Neurones de la couche cachée – S2 : la sortie (=1)
 % TF1, TF2 : Fonctions de transferts – BTF : Fonction de transfert de rétro-propagation
net.trainParam.epochs =800;            % Nombre d'itération
net.trainParam.goal= 0.0000001;      % Erreur désirée
net = init(net);        % Initialisation du réseau
[net,tr]=train (net, P1, T1, [], [], val);      % Entrainement du réseau
plotperf(tr)
a1n=sim(net,P1);       % Simulation du réseau pour les données d'apprentissage
[a1]=postmnmx(a1n,minT0,maxT0);%Remettre les résultats d'apprentissage à leurs valeurs
réels
E1= T10-a1; % Calcul de l'erreur
a2n=sim(net,test.P);  % Simulation du réseau pour les données du test
[a2]=postmnmx(a2n,minT0,maxT0);%Remettre les résultats du test à leurs valeurs réels
E2= T20-a2; % Calcul de l'erreur
a3n=sim(net,val.P); % Simulation du réseau pour les données de validation
[a3]=postmnmx(a3n,minT0,maxT0);%Remettre les résultats de validation à leurs valeurs
réels
E3= T30-a3;  % Calcul de l'erreur
```

II-5 - PARAMETRES D'EVALUATION DE LA QUALITE DE L'AJUSTEMENT

Deux paramètres sont couramment utilisés :

- Le coefficient de détermination multiple :

$$R^2 = 1 - \frac{SCE}{SCT} = 1 - \frac{\sum_1^n (y_i - \hat{y}_i)^2}{\sum_1^n (y_i - \bar{y})^2} \qquad (7)$$

Où $\hat{y}_i$ est la valeur estimée du paramètre étudiée de l'ensemble de calibratin, et $\bar{y}$ la moyenne des valeurs observées de cet ensemble.

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

$$\sigma_N = \sqrt{\frac{1}{n_{EXT}} \sum_{i=1}^n (y_i - \hat{y}_{(i)})^2} \qquad (8)$$

II-5 – 1 - Robustesse du modèle :

La stabilité du modèle a été explorée en utilisant la "validation croisée par omission d'une observation" (LOO : cross-validation by leave-one-out) [27]. Elle consiste à recalculer le modèle sur (n – 1) composés 'amides herbicides, le modèle obtenu servant alors à estimer la valeur du pDL_{50} du composé éliminé noté $\hat{y}_{(i)}$. On répète le procédé pour chacun des herbicides étudiés.

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

$$PRESS = \sum_1^n (y_i - \hat{y}_{(i)})^2 \qquad (9)$$

est une mesure de la dispersion de ces estimations.

On l'utilise pour définir le coefficient de prédiction :

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

Contrairement à R^2, qui augmente avec le nombre de paramètres de la régression, le facteur Q^2_{LOO} affiche une courbe avec maximum (ou avec palier), obtenu pour un certain nombre de variables explicatives, puis décroît par la suite de façon monotone. Ce fait confère une grande importance au coefficient Q^2_{LOO}. Une valeur de Q^2_{LOO} > 0,5 est, généralement, considérée comme satisfaisante, et une valeur supérieure à 0,9 est excellente [28].

II-5 -2 - Détection des observations aberrantes :

Elle a été basée sur la non – satisfaction à trois au moins (pour n'en privilégier aucun) des tests statistiques couramment utilisés pour la détection de telles observations en analyse de régression :

1°/ Les résidus ordinaires e_i, différences entre les valeurs observées (y_i) et estimées par le modèle $(\hat{y}_i)$.

2°/ Les leviers, h_{ii}, permettent de juger de l'influence d'une observation i dans la détermination de l'équation de régression.

3°/ Les résidus de prédiction standardisés e_{istd} ont été calculés avec le logiciel MOBYDIGS [29].

II -5 - 3 - Test de randomisation :

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

En plus du test de randomisation, il est intéressant [30], pour juger de la qualité du modèle, de considérer la racine de l'écart quadratique moyen EQM (RMSE : Root Mean Squared Error), calculée sur différents ensembles.

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

$$EQMC = \sqrt{\frac{\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2}{n}} \tag{11}$$

$$\sigma_N = EQMP = \sqrt{\frac{\sum_{i=1}^{n}\left(y_i - \hat{y}_{(i)}\right)^2}{n}} = \sqrt{\frac{PRESS}{n}} \tag{12}$$

II - 5 - 4- Validation statistique externe :

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

$$Q^2_{ext} = 1 - \frac{\sum_{i=1}^{n_{ext}}\left(y_i - \hat{y}_i\right)^2 \Big/ n_{ext}}{\sum_{i=1}^{ntr}\left(y_i - \bar{y}\right)^2 \Big/ n_{tr}} \tag{13}$$

Pour une grande valeur de Q^2_{LOO}, une valeur élevée de Q^2_{ext} permet de présager d'une bonne capacité prédictive du modèle. Il est intéressant de considérer, également, la racine de l'écart quadratique moyen de prédiction externe (EQMPext), calculée sur l'ensemble de validation :

$$EQMP_{ext} = \sqrt{\frac{\sum_{i=1}^{n_{ext}}\left(y_i - \hat{y}_{(i)}\right)^2}{n_{ext}}} \tag{14}$$

PARTIE EXPERIMENTALE

Nous traiterons la dose létale comme propriété étudiée avec un ensemble d'estimation et de validation. Nous favoriserons l'approche hybride algorithme génétique soit régression multilinéaire, soit réseaux de neurones :

A G / RLM et AG / RNA

III-1- MODELE HYBRIDE ALGORITHME GENETIQUE / REGRESSION LINEAIRE MULTIPLE:

III -1 -1 - Calcul du modèle:

Le graphe de la figure 6 reproduit les variations du FIT (équation (2)) en fonction du nombre de variables du modèle .Dans les études QSAR, cinq observations, à la limite au minimum, doivent être associées à chaque variable explicative. Alors on a calculé le FIT à un nombre maximum de descripteurs égal à 8.

Le choix de la taille du model est fait en maintenant une différence minimale du FIT entre 7et 8 descripteurs et le choix finale du nombre de descripteurs est égal à 7.

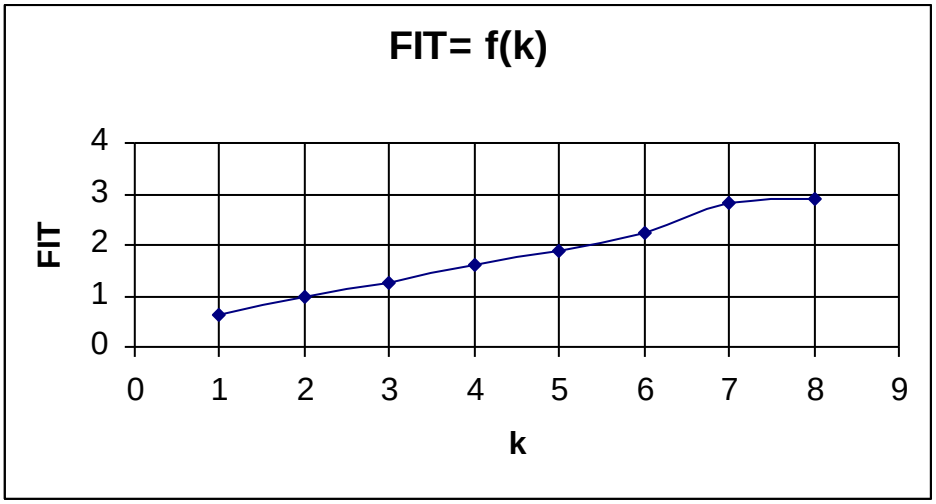


Figure 6 – Variation du FIT en fonction de nombre de descripteurs pour pDL_{50}.

Le nombre de degrés de liberté final doit être au moins égal à 10, soit, en désignant par k le nombre de descripteurs :

$$n-k-1 \geq 10 \qquad (15)$$

Cette condition est bien vérifiée pour le modèle à 7 variables.

Une indépendance globale acceptable des descripteurs sera vérifiée lorsque les facteurs d'inflation de la variance (FIV) calculés pour chacun d'eux sont inférieurs à 5.

Parmi les modèles optimaux générés, celui qui fournit la valeur maximale pour les paramètres statistiques Q^2, R^2 et Q^2_{ext} tout en vérifiant la condition : FIV≤ 5 , comporte les 7

descripteurs calculés par le logiciel DRAGON[14], dont les symboles, la classe et la signification sont réunis dans le tableau 3.

Tableau 3 - Descripteurs moléculaires intervenant dans la modélisation de pDL_{50}.

N°	Descripteur	Classe	Signification
1	MATS7m	Descripteurs d'autocorrelation2D (Bloc 6)	Autocorrélation de Moran de distance topologique7 pondéré par les masses atomiques.
2	EEig14d	Indices d'adjacence des arêtes (Bloc 7)	Ce sont des descripteurs moléculaires calculés à partir de la matrice d'adjacence des arêtes d'une molécule. C'est une matrice carrée symétrique de dimensions B×B, où B est le nombre de laisons entre les paires d'atomes non H
3 4 5	Mor28u Mor31u Mor03m	Descripteurs MoRSE -3D (Bloc 14)	Signal 28- MoRSE-3D / non pondéré Signal 31- MoRSE-3D/ non pondéré Signal 03- MoRSE-3D/ Pondéré par la masse atomique.
6	R7v	Descripteurs GETAWAY (Bloc 16)	R autocorrélation de distance topologique 7 pondéré par les volumes de Van der WAALS.
7	nHDon	Nombre des groupes fonctionnels (Bloc17)	Nombre d'atomes donneurs de laisons Hydrogènes (O et N)

L'équation de régression ainsi établie est reproduite ci-après :

pDL_{50}= - 0,013 - 0,939 MATS7m - 0,0836 EEig14d - 0,850 Mor28u - 0,545 Mor31u
- 0,129 Mor03m + 0,562 R7v + 0,428 nHDon (16)

S =0, 2021, n = 40 ; σ_N=0,181 ; R^2 = 81,66 % ; Q^2=0,7403 ; F = 20,34

III -1- 2- Analyse de régression :

Les valeurs des paramètres statistiques montrent que les sept descripteurs (**MATS7m ; EEig14d ; Mor28u ; Mor31u ; Mor03m ; R7v ; nHDon**) (tableau-3) permettent de corréler la pDL_{50} de 40 amides herbicides.

Régresseur	Coef	Er-T coef	T	P	FIV
Constante	-0,0132	0,1606	-0,08	0,935	
MATS7m	-0,9393	0,1688	-5,56	0,000	1,1
EEig14d	-0,08359	0,04461	-1,87	0,070	1,1
Mor28u	-0,8498	0,1762	-4,82	0,000	2,2
Mor31u	-0,5446	0,2320	-2,35	0,025	2,4
Mor03m	-0,12850	0,03227	-3,98	0,000	1,4
R7v	0,5625	0,3712	1,52	0,139	1,5
nHDon	0,42773	0,06168	6,93	0,000	1,8

III-1-2 -1 - Matrice de Corrélation :

	pDL	MATS7m	EEig14d	Mor28u	Mor31u	Mor03m	R7v
MATS7m	-0,336						
	0,034						
EEig14d	-0,086	-0,123					
	0,597	0,450					
Mor28u	-0,150	-0,021	0,034				
	0,357	0,895	0,834				
Mor31u	-0,338	-0,144	0,011	-0,514			
	0,033	0,375	0,945	0,001			
Mor03m	-0,002	-0,029	-0,163	-0,446	0,025		
	0,991	0,860	0,315	0,004	0,876		
R7v	0,074	-0,073	-0,007	-0,542	0,419	0,218	
	0,651	0,656	0,967	0,000	0,007	0,176	
nHDon	0,673	0,045	-0,043	0,232	-0,643	0,076	-0,284
	0,000	0,783	0,793	0,150	0,000	0,639	0,076

La valeur du cœfficient de détermination (R^2) signifie que 81,66 % de la variabilité de pDL_{50} peut être expliquée par ces sept descripteurs, alors que la racine de l'erreur quadratique moyenne de prédiction est de l'ordre de (σ_N=0,181) ; en outre ce modèle est significatif (avec une valeur du paramètre de Fisher : F=20,34).

Notons que les résidus de prédiction standardisés ainsi que les éléments diagonaux de la matrice **H** (h_{ii}), ont été calculés par le logiciel MOBYDIGS[29], et que pour l'ensemble de validation on a choisi de reporter quelques statistiques seulement figurant dans le tableau (5).

Tableau 4 - Valeurs des pDL_{50} observés et pDL_{50} prédits en exploitant la RLM, leurs différences, valeurs des leviers ainsi que les résidus de prédiction standardisés pour l'ensemble de calibration.

	Colonne1	**Colonne2**	**Colonne3**	**Colonne4**	**Colonne5**
Obs	**pDL_{50obs}**	**e_i**	**e_{istd}**	**h_{ii}**	**pDL_{50cal}**
1	2,1	-0,0909	-0,9471	0,391	2,0091
2	1,67	0,0042	0,026	0,136	1,6742
3	1,67	-0,0778	-0,498	0,157	1,5922
5	1,58	-0,4228	-2,3636	0,078	1,1572
7	1,48	-0,0194	-0,205	0,397	1,4606
9	1,45	0,0934	0,6009	0,16	1,5434
10	1,43	0,1058	0,6796	0,16	1,5358
11	1,2	0,0333	0,6102	0,582	1,2333
13	0,38	0,0803	0,7216	0,328	0,4603
15	1,22	0,0144	0,0904	0,146	1,2344
17	1,25	0,1762	1,0783	0,132	1,4262
18	1,22	-0,3845	-2,6617	0,2	0,8355
21	0,99	0,0364	0,2295	0,149	1,0264
22	0,74	-0,2015	-1,2162	0,124	0,5385
23	1	0,3129	1,8978	0,127	1,3129
24	0,94	-0,2089	-1,2487	0,118	0,7311
25	0,75	-0,0415	-0,3885	0,346	0,7085
26	0,94	-0,1932	-1,1748	0,128	0,7468
27	0,63	0,0807	0,8057	0,374	0,7107
28	0,7	0,2293	1,4541	0,152	0,9293
29	0,73	0,0453	0,2611	0,097	0,7753
30	0,63	-0,0191	-0,1092	0,091	0,6109
31	0,61	-0,0254	-0,1572	0,138	0,5846
33	0,53	0,0447	0,2971	0,179	0,5747
34	1,46	-0,0955	-0,6637	0,202	1,3645
36	0,98	-0,0017	-0,0174	0,386	0,9783
37	1,67	-0,2637	-2,1785	0,289	1,4062
38	0,9	-0,0136	-0,0916	0,186	0,8864
39	1,45	-0,0541	-0,3853	0,216	1,3959
40	1,03	-0,0174	-0,104	0,119	1,0126
41	0,92	0,03	0,1888	0,148	0,95
42	1,51	-0,1067	-0,6591	0,138	1,4033
43	1,02	0,0416	0,3361	0,278	1,0616
44	0,87	-0,0119	-0,0775	0,166	0,8581
45	0,81	0,0231	0,14	0,127	0,8331
46	0,77	-0,0907	-0,7665	0,3	0,6793
47	0,7	0,2582	1,4596	0,085	0,9582
48	0,62	0,2678	1,6353	0,131	0,8878
49	0,39	-0,1355	-0,9125	0,186	0,2545
50	0,16	0,5983	3,8021	0,154	0,7583

Tableau 5 - Valeurs des pDL_{50} observés et pDL_{50} prédits en exploitant la RLM, leurs différences, valeurs des leviers ainsi que les résidus de prédiction standardisés pour l'ensemble de validation externe.

$Obs_{pré}$	pDL_{obs}	$pDL_{50pré}$	$e_{(i)}$	h_{ii}	e_{istd}
4	2,42	1,97172	0,44828	0,226	-2,5333
6	0,53	0,98493	-0,4536	0,104	2,3716
8	1,59	1,42539	0,16461	0,126	-0,8817
12	1,17	1,01655	0,15345	0,382	-0,9758
14	1,05	0,98422	0,06578	0,369	-0,4223
16	1,07	1,10038	-0,03038	0,216	0,1623
19	1,13	1,41483	-0,28483	0,265	1,637
20	1,02	0,84231	0,17769	0,47	-1,2124
32	0,62	0,81738	-0,19738	0,378	1,2343
35	1,17	0,92348	0,24652	0,768	-2,554

III-1-3- Résultats et discussion:

L'analyse des résidus (colonne 2 tableau 4) permet, en particulier, de voir que les résidus ordinaires,en valeur absolue, sont inférieurs à 3 fois l'erreur standard ($|e_i|<3S$), soit

$3\times 0,2021 = 0,6063$.

La colonne (3) rassemble les résidus de prédictions standardisés (e_{istd}) , qui sont dans l'intervalle $\pm 3\sigma$ à l'exception de l'observation 50 (ensemble de calibration).

La colonne (4) donne les valeurs de h_{ii}, $i^{ème}$ terme diagonal de la matrice de projection : $\mathbf{H=X(X'X)^{-1}X'}$

Où **X** est la matrice des valeurs observées des variables explicatives et **X'**est sa transposée, ces valeurs sont utiles pour le calcul des résidus caractéristiques.

La valeur critique pour déterminer les points leviers correspond à :

$$h^* = \frac{3p}{n} = \frac{3\times 8}{40} = 0,6$$

Le diagramme de Williams (e_{istd} en fonction de h_{ii}) de la figure (7) fait ressortir le point (50) comme aberrant. Le point (11), de l'ensemble de calibration toujours, est presque influent (h_{ii}=0,582 proche de h*). Pour le groupe de validation le point (35) a une valeur de h_{ii} supérieur à h* =0,6.

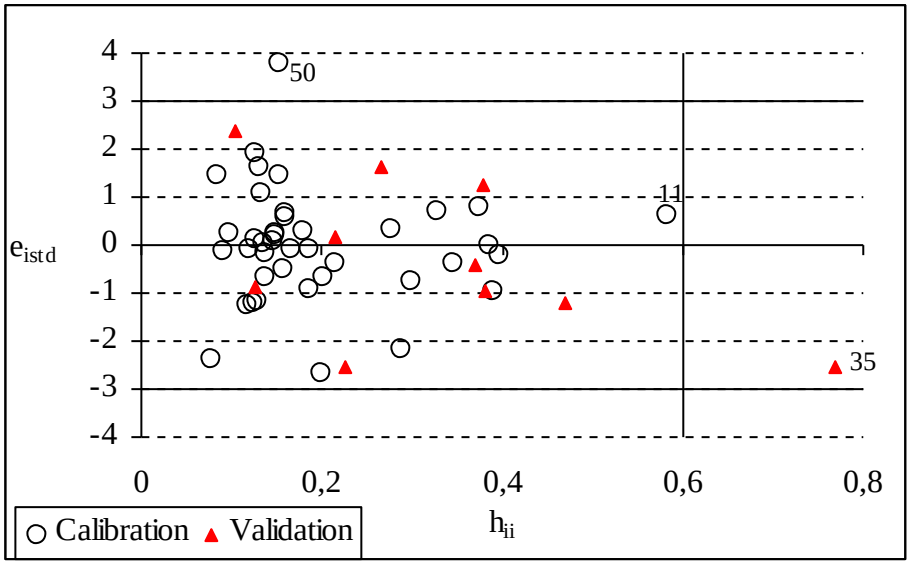


Figure – 7: Diagramme de Williams pour les deux ensembles; calibration et validation.

III -1- 4 - Vérification de la qualité de l'ajustement:

La qualité de l'ajustement a été vérifiée en procédant à une validation croisée par « Leave –one –out ». La figure (8), qui reproduit les valeurs prédites pDL_{obs} en fonction de celles calculées, fait ressortir une dispersion caractéristique d'un assez bon ajustement, d'ailleurs confirmé par la grande valeur de Q^2 (=0,7403).

$pDL_{50obs} = 0.0000010 + 1.00000\ pDL_{50\ cal}$

S =0.185420, R2= 81.7 ٪ , $R2_{ajust}$ = 81.2 ٪

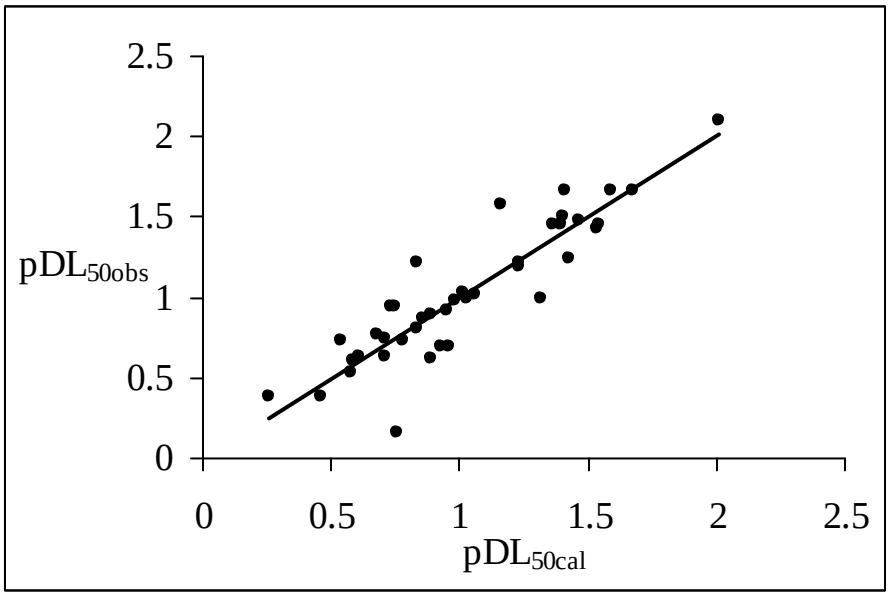


Figure 8– Graphe des valeurs calculées $pDL_{50\ obs}$ en fonction de $pDL_{50\ cal}$.

Dans le but d'établir que le modèle obtenu n'est pas dû au hasard ou à une sur-spécification, nous avons appliqué le test de randomisation.

Ainsi 100 nouveaux vecteurs de doses létales ont été générés par permutation des positions des composantes du vecteur réel:

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

et utilisés comme sources d'observations pour des modèles QSAR dans les conditions optimales établies (7 paramètres) .

La figure 9 qui représente le graphe des coefficients statistiques Q^2 et R^2 permet de comparer les résultats obtenus pour les modèles randomisés (cercles noircis) au modèle réel de départ (triangle).

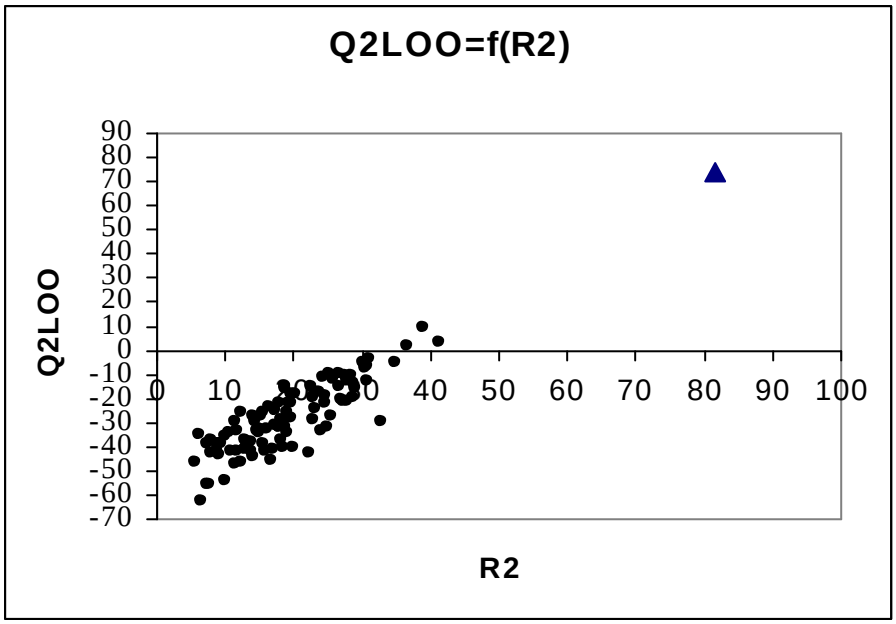


Figure 9 – Test de randomisation associé au modèle QSAR.

Les points noircis représentent les doses létales ordonnées de façon aléatoire, et le triangle (▲) correspond au modèle réel.

Il est clair que les statistiques obtenues pour les vecteurs modifiés de la dose létale sont plus petites que celles du modèle QSAR réel, et pour la majeure partie on obtient un $Q^2 < 0$ à l'exception de trois observations pour lesquelles Q^2 est inférieur à 10 ٪ et R^2 de l'ordre de 40٪ .Ceci permet d'assurer qu'une relation structure / dose létale testée a été établie.

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

III-1-5 –Validation statistique externe:

Pour savoir la capacité prédictive du notre modèle, nous avons opéré par validation externe sur l'ensemble de 10 composés choisis aléatoirement et qui ne font pas partie de l'ensemble d'essai (composés surmontés d'un astérisque * dans le tableau 2).

Une validation rigoureuse du modèle se traduit par une proportion importante de prédictions exactes données sur l'ensemble de la validation. La performance du modèle est alors mesurée par le coefficient de régression R^2.

$$pDL_{50obs} = -0.307581 + 1.29476 \; pDL_{50obs}$$

$$S=0.266717, \; R2=77.2\,\%, \; R2_{ajust}=74.4\,\%$$

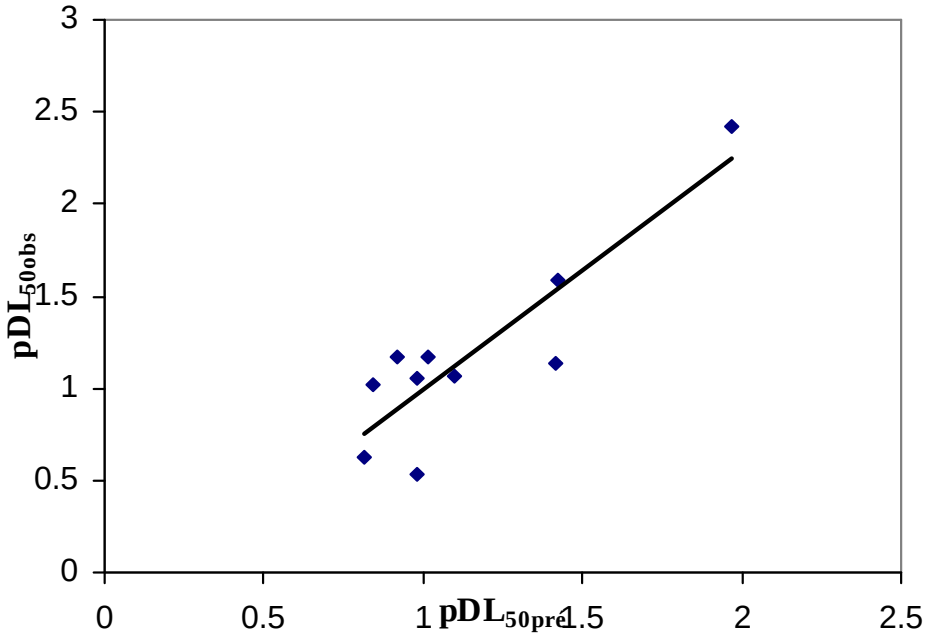


Figure 10 – Graphe des pDL_{50obs} en fonction des $pDL_{50pré}$ pour la validation externe.

Les valeurs des paramètres statistiques, réunies ci-après,

EQMC = 0,181	(40 objets)	R^2	= 81,66
EQMP = 0,215	(40 objets)	Q^2	= 74,03
EQMP (ext) = 0,260	(10 objets)	Q^2 (ext)	= 61,89

montrent tout à la fois, une mauvaise capacité prédictive (valeurs des EQMC,EQMP,$EQMP_{ext}$ élevées), On peut dire aussi que les valeurs des pDL_{50} prédites sont un peu différentes des pDL_{50} observés sauf pour certains composés tels que 14 et 16 qui possèdent des erreurs moins élevées.

III-2- MODELE HYBRIDE ALGORITHME GENETIQUE / RESEAUX DE NEURONES ARTIFICIELS :

III-2-1- choix des paramètres statistiques :

Les descripteurs choisis par algorithme génétique sont utilisés pour la configuration du réseau de neurones, qui est perfectionnée en phase d'apprentissage ; les paramètres de fonctionnement sont déterminés de façon à obtenir une bonne adéquation entre les valeurs simulées et les données d'apprentissage, combinée à une généralisation correcte de ces simulations,

III-2-2- Choix du nombre de neurones dans la couche cachée:

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

III-2-3 -Choix du nombre d'itérations et de neurones dans la couche cachée:

Le choix de ce nombre est très important, Au départ, on fixe un nombre de neurones (2 à 6), avec un nombre de neurones ne dépassent pas le nombre de descripteurs formant la taille du model et on fait varier le nombre d'itérations pour calculer à chaque fois l'erreur quadratique moyenne EQM, dite MSE en anglais.

Le nombre de neurones de la couche cachée ainsi que le nombre d'itérations et fixé par la valeur minimale de l'erreur quadratique moyenne EQM,

Le graphe EQM = f (nombre d'itérations) à balayage de nombres de neurones de la figure suivante permet de visualiser EQM_{min} qui correspond à 800 pour le nombre d'itérations et à 04 pour le nombre de neurones de la couche cachée pour une meilleure configuration du réseau,

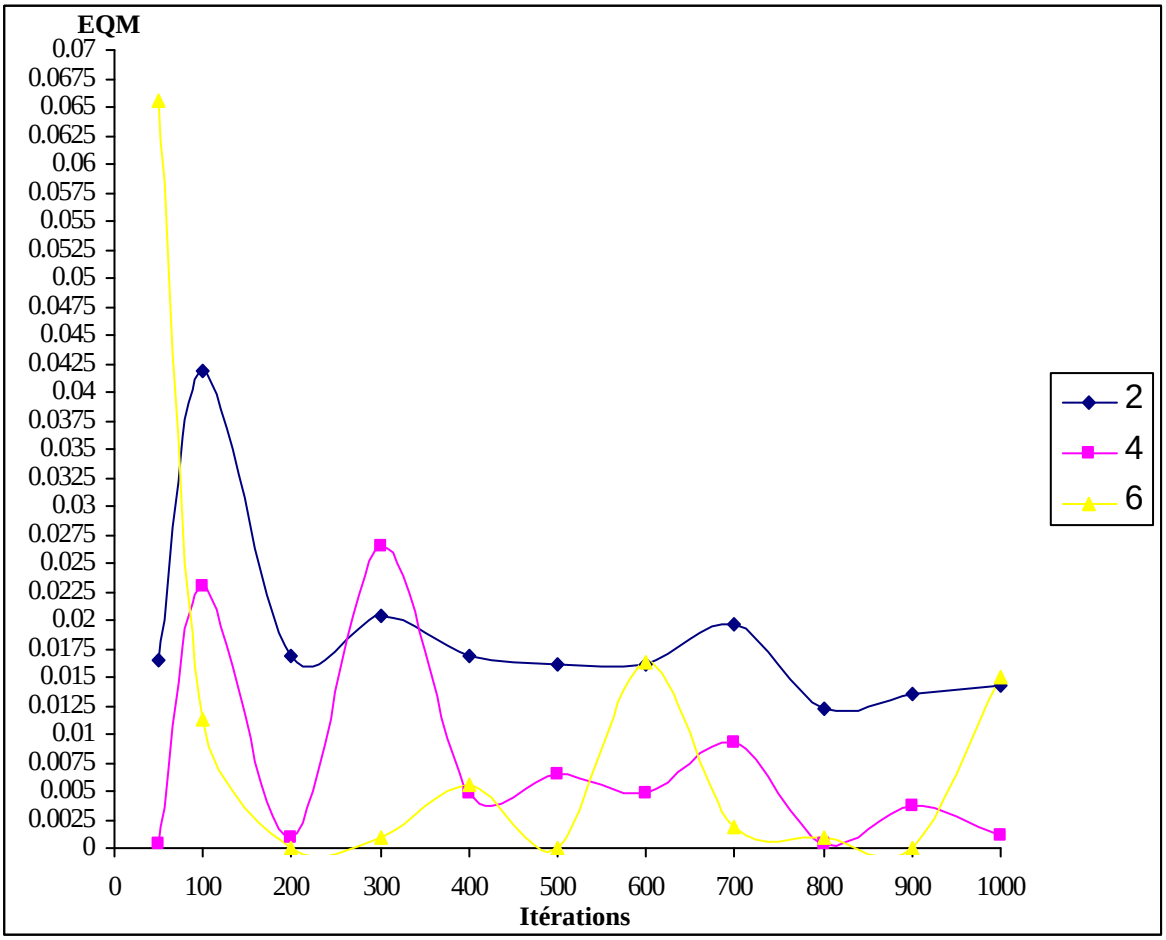


Figure 11 – Choix du nombre d'itérations, et de neurones dans la couche cachée

III-2- 4- Choix de la fonction de transfert:

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

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

III-2-5 - Choix des paramètres d'apprentissage:

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

- Indice de performance choisi: EQM (pour l'erreur quadratique moyenne),

L'apprentissage du réseau de neurones représente un fragile équilibre entre tous ces paramètres, d'où la difficulté pour l'atteindre, Une fois cet apprentissage achevé, le réseau de neurones devient un outil viable et peut être utilisé pour la simulation de nouvelles données, Le tableau 6 précise la structure optimale du réseau de neurones.

T**ableau 6 -** Structure optimale du réseau de neurones,

Nombre d'entrées	07 (les descripteurs)
Nombre de sorties	01 (pDL_{50})
Nombre de couches cachées	Une couche cachée
Nombre d'itérations	800
Nombre de neurones dans la couche cachée	04
Algorithme d'apprentissage	Rétro propagation du gradient de l'erreur
Fonctions d'apprentissage	Tangente hyperbolique (couche cachée) Linéaire (couche de sortie)

III-2-6- Résultats et discussion :

III-2-6 -1- Evaluation de la qualité de l'ajustement:

Nous avons deux paramètres utilisés pour évaluer la qualité de l'ajustement; la valeur du cœfficient de détermination R^2 = 92,25 % qui explique très bien la variabilité de pDL en fonction des descripteurs choisis; la racine de l'erreur quadratique moyenne de prédiction σ_N= 0,1174439 dont la petite valeur indique un modèle très hautement significatif, que justifie la grande valeur du paramètre de Fisher : F= 59,009.

III-2-6-2- Vérification de la qualité de l'ajustement :

La qualité de l'ajustement a été vérifiée en procédant à une validation croisée par « leave –one –out », La figure (12), reproduit les valeurs prédites pDL_{50} en fonction de celles observées, fait ressortir un bon ajustement, d'ailleurs confirmé par la valeur de Q^2 =0,8727 .

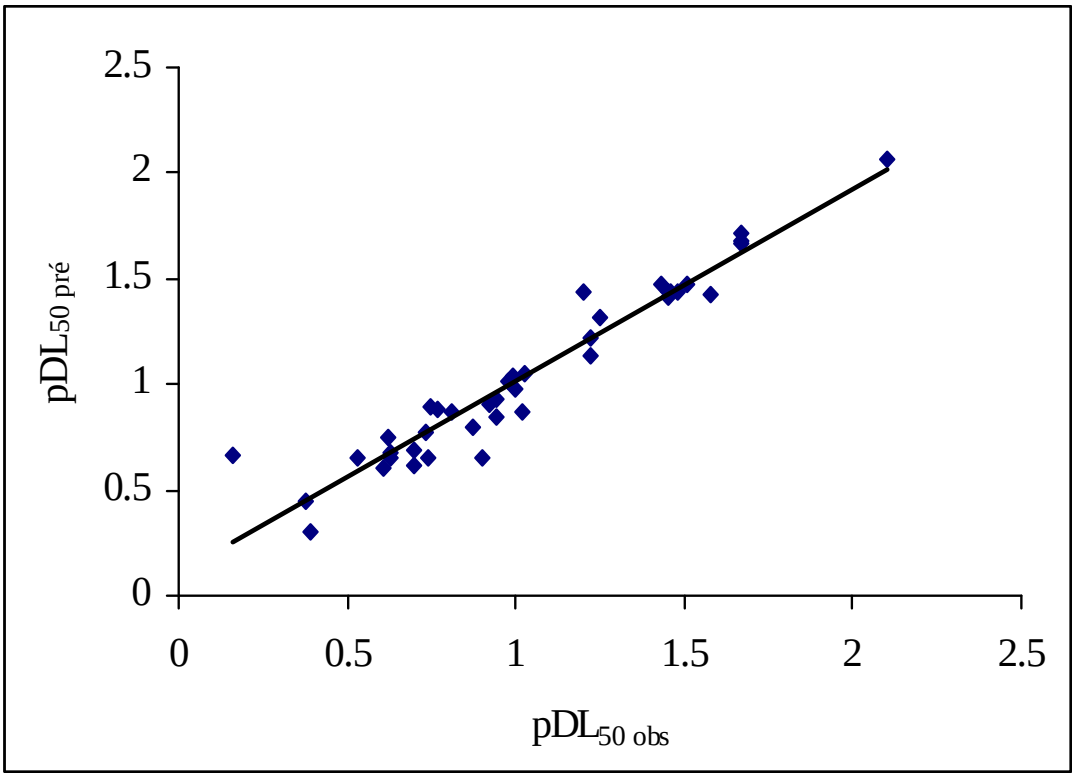


Figure 12 – Graphe des valeurs $pDL_{50pré}$ en fonction des valeurs $pDL_{50\ obs}$.

III-2- 6 -3 –Diagramme de Williams:

Le diagramme de Williams (e_{istd} en fonction de h_{ii}) de la figure (13) fait ressortir l'observation aberrante (50) et l'observation (11) ayant une valeur de h_{ii} très proche de h*(h*=0.6) pour l'ensemble d'estimation et pour l'ensemble de validation externe l'observation (16) est aberrante ($ei_{std} > 3$), l'observation (35) a une valeur $h_{ii} > h^*=0.6$.

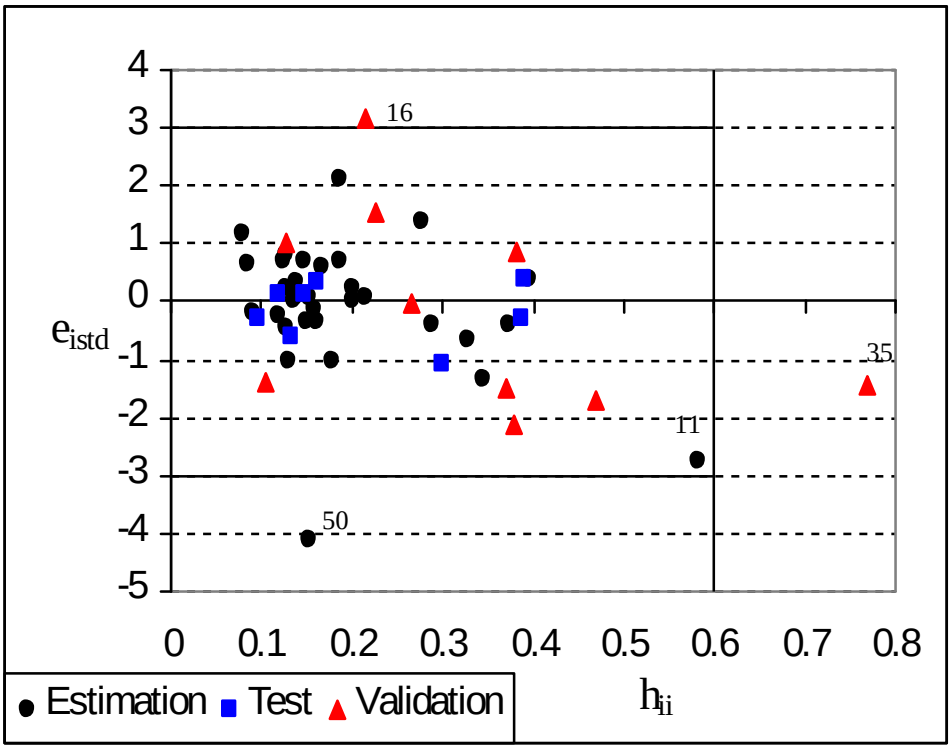


Figure 13 - Diagramme de Williams pour l'étude en RNA

Tableau 7- Les valeurs pDL_{50} observées, prédites et les erreurs pour l'ensemble de validation externe trouvées par RNA pour l'ensemble d'estimation et test .

Obs	pDL_{50osb}	$pDL_{50pré}$	e_i	e_{istd}	h_{ii}
1	2,1	2,0617	0,0383	0,37378804	0,391
2	1,67	1,6645	0,0055	0,04506518	0,136
3	1,67	1,6831	-0,0131	-0,10866578	0,157
5	1,58	1,4301	0,1499	1,1889715	0,078
7	1,48	1,4404	0,0396	0,38839337	0,397
9	1,45	1,4114	0,0386	0,32076202	0,16
10	1,43	1,4689	-0,0389	-0,32325499	0,16
11	1,2	1,4346	-0,2346	-2,76359973	0,582
13	0,38	0,4488	-0,0688	-0,6392034	0,328
15	1,22	1,1349	0,0851	0,70135177	0,146
17	1,25	1,3222	-0,0722	-0,59021822	0,132
18	1,22	1,2146	0,0054	0,0459816	0,2
21	0,99	1,033	-0,043	-0,35500866	0,149
22	0,74	0,6513	0,0887	0,72178332	0,124
23	1	0,9728	0,0272	0,22171601	0,127
24	0,94	0,9259	0,0141	0,11434577	0,118
25	0,75	0,8906	-0,1406	-1,32413334	0,346
26	0,94	0,8437	0,0963	0,78542246	0,128
27	0,63	0,6728	-0,0428	-0,41199493	0,374
28	0,7	0,6939	0,0061	0,0504507	0,152
29	0,73	0,7687	-0,0387	-0,31017184	0,097
30	0,63	0,6526	-0,0226	-0,18053515	0,091
31	0,61	0,6026	0,0074	0,06070345	0,138
33	0,53	0,6513	-0,1213	-1,01958749	0,179
34	1,46	1,4347	0,0253	0,21570209	0,202
36	0,98	1,0122	-0,0322	-0,31297306	0,386
37	1,67	1,7148	-0,0448	-0,40464901	0,289
38	0,9	0,652	0,248	2,09350861	0,186
39	1,45	1,439	0,011	0,09461716	0,216
40	1,03	1,0563	-0,0263	-0,21340425	0,119
41	0,92	0,9071	0,0129	0,10644008	0,148
42	1,51	1,4697	0,0403	0,3305877	0,138
43	1,02	0,8652	0,1548	1,38751482	0,278
44	0,87	0,8009	0,0691	0,57627569	0,166
45	0,81	0,8648	-0,0548	-0,44669255	0,127
46	0,77	0,8869	-0,1169	-1,06414489	0,3
47	0,7	0,619	0,081	0,64492579	0,085
48	0,62	0,7464	-0,1264	-1,03269592	0,131
49	0,39	0,3041	0,0859	0,7251306	0,186
50	0,16	0,6593	-0,4993	-4,13439206	0,154

L'analyse des résidus permet, de voir que le résidu ordinaire e_{50} est, en valeur absolue, supérieurs à 3 fois l'erreur standard ($|e_i|>3S$), soit 3 x 0,1313 =0,3939**.**

III-2-7-**Validation statistique externe**:

L'évaluation de la capacité de généralisation du réseau est réalisée sur la base de la validation statistique externe.

Les résultats obtenus montrent que les valeurs prédites (tableau 8) sont très proches des valeurs observées (figure14), La valeur de R^2 est égale à 92,25 %, qui confirme que le modèle neuronal décrit de façon adéquate la relation entre p D L_{50} prédites et observées.

pDL50 pré = 0.245510 + 0.795497 pDL50 obs

S = 0.166060 R-carré = 87.8 % R-carré(ajust) = 86.2 %

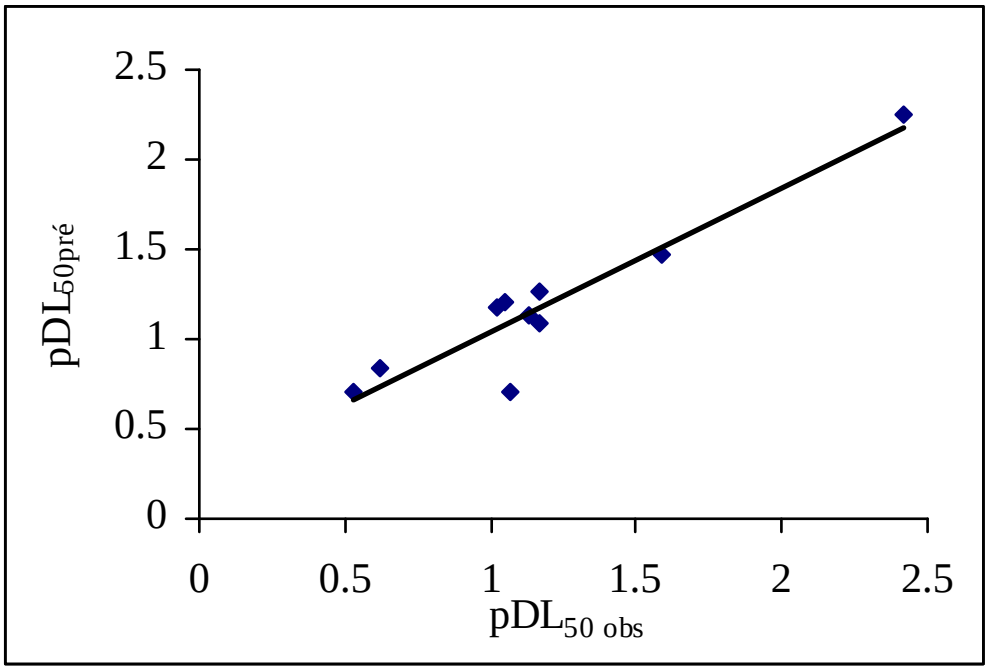


Figure 14 – Graphe des pDL_{50} prédites en fonction des pDL_{50} observées pour l'ensemble de la validation externe.

Tableau 8- Les valeurs pDL observées, prédites et les erreurs pour l'ensemble de validation externe trouvées par RNA.

Composés	pDL50 obs	pDL50 pré	e(i)	$ei_{\text{std val}}$	$h_{ii\ val}$
4	2,42	2,243	0,176	1,52528108	0,226
6	0,53	0,7053	-0,1753	-1,41040024	0,104
8	1,59	1,4658	0,1242	1,0117666	0,126
12	1,17	1,0818	0,0882	0,85445424	0,382
14	1,05	1,2037	-0,1537	-1,47357976	0,369
16	1,07	0,7031	0,3669	3,15576126	0,216
19	1,13	1,1331	-0,0031	-0,02753801	0,265
20	1,02	1,1815	-0,1615	-1,68946443	0,47
32	0,62	0,8382	-0,2182	-2,10704599	0,378
35	1,17	1,2618	-0,0918	-1,451487	0,768

Les valeurs des paramètres statistiques sont réunies ci-après :

EQMC = 0,12	(40objets)	R^2 = 92,25
EQMP = 0,15	(40 objets)	Q^2 = 87,28
EQMP (ext) = 0,1803	(10 objets)	Q^2 (ext) = 81,73

Les faibles valeurs des EQMC ,EQMP,$EQMP_{ext}$ montrent une bonne capacité prédictive du modèle et une possibilité d'extension suffisante (valeurs proches).

Tableau 9- Comparaisons des valeurs de pDL observées, prédites et les résidus trouvés par RLM et RNA pour l'ensemble de validation externe**,**

N°	$pDL_{50\ obs}$	$pDL_{50pré}$(RLM)	$pDL_{50pré}$(RNA)	e(i) (RLM)	e(i) (RNA)
4	2,42	1,97172	2,2438	0,44828	0,176
6	0,53	0,98493	0,7053	-0,4536	-0,1753
8	1,59	1,42539	1,4658	0,16461	0,1242
12	1,17	1,01655	1,0818	0,15345	0,0882
14	1,05	0,98422	1,2037	0,06578	-0,1537
16	1,07	1,10038	0,7031	-0,03038	0,3669
19	1,13	1,41483	1,1331	-0,28483	-0,0031
20	1,02	0,84231	1,1815	0,17769	-0,1615
32	0,62	0,81738	0,8382	-0,19738	-0,2182
35	1,17	0,92348	1,2618	0,24652	-0,0918

Le jugement de la qualité des modèles (RLM et RNA) a été vérifié en calculant les EQM; ces paramètres statistiques sont réunis ci-après:

Tableau 10 - Valeurs des paramètres statistiques trouvés par les deux méthodes,

Paramètres statistiques	RLM	RNA
n	40	40
$\mathbf{n_{ext}}$	10	10
S	0,2021	0,1313
σ_N	0,181	0,117
$\mathbf{R^2}$	81,66	92,85
$\mathbf{Q^2}$	74,03	88,27
$\mathbf{Q^2_{ext}}$	61.89	81,73
F	20,34	59,009
EQMC	0,181	0,12
EQMP	0,215	0,15
$\mathbf{EQMP_{ext}}$	0,260	0,1803

CONCLUSION GENERALE

CONCLUSION GENERALE

Nous avons appliqué la méthodologie QSAR pour relier La toxicité des composés amides herbicides etudiés , à des descripteurs moléculaires théoriques reflétant certaines particularités des molécules considérées.

Le mélange pris en compte comprend 50 composés substitué par des divers groupements.

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

Les 50 composés de base ont été éclatées aléatoirement en deux ensembles disjoints.

- un ensemble principal de 40 composés utilisés pour le calcul et, éventuellement, les essais du modèle ;
- un ensemble de 10 composés pour la validation externe.

La taille du modèle est fixée par la valeur optimale de la fonction FIT de KUBINYI. La sélection des variables explicatives a été réalisée par algorithme génétique, dans la version MOBYDIGS de TODESCHINI[14]en maximisant Q^2_{L00}.

Les statistiques réunies ci-après permettent de faire des comparaisons, et de tirer plusieurs conclusions comparons les résultats trouvés par RLM et RNA .

		(I)	(II)	(III)	(IV)	(V)	(VI)	(VII)	(VIII)
		R^2 (%)	**σ_N**	**Q^2(%)**	**F**	**Q^2_{ext}(%)**	**EQMP(ext)**	**Points * aberrants**	**Points ** aberrants**
pLD	(**RLM**)	81,66	0,181	74,03	20,34	61.98	0.260	50	/
	(**RNA**)	92,25	0.117	87.28	59,009	81.73	0.1803	50	16

Points de l'ensemble d'essai (*) et de l'ensemble de prédiction externe(**).

La numérotation des composés est celle du tableau 2 page

Les statistiques (R^2 ; σ_N ; F) calculées permettent de juger la qualité de notre modèle développé.

L'analyse des résidus a permis de détecter une observation aberrante précisée des ensembles d'essai dans la colonne (VII) pour les deux modèles (RLM et RNA) par contre l'ensemble de validation externe présente un seul de point aberrant uniquement pour le modèle RNA dans la colonne(VIII) .

La qualité de l'ajustement a été vérifiée en procédant à une validation croisée par "leave – one - out". Les valeurs de Q^2 obtenues (colonne (III)) sont proches de celles du

coefficient de détermination multiple correspondant R^2 (colonne (I)), ce qui fait ressortir la qualité de l'ajustement de notre modèle obtenu pour chaque méthode RLM et RNA.

Les valeurs $EQMP_{ext}$ réunies dans la colonne (VI) sont elevées dans le cas du modèle (RLM) par contre ils sont faibles dans le cas du modèle (RNA), ce qui permet,de s'assurer de la bonne capacité prédictive de ce dernier.

Le test de randomisation montre, que le modèle (RLM) obtenu n'est pas dû au hasard.

Ce travail sera étendu à d'autres familles d'herbicides , en diversifiant la structure des ensembles des données(composés du même domaine chimique).

REFERENCES BIBLIOGRAPHIQUES

[1]: www.agirpourlenvironnement.org/campagnes/c271.htm#top

[2]:J.Braz, Chem.Soc; V01, 13.No.6, 754-6762, 2002

[3]: http://fr.wikipedia.org/w/index.php?oldid=47672409

[4]: G. A. Patani, E. J. LaVoie, Bioisosterism: A Rational Approach in Drug Design. Chem. Rev., 1996, 96, 3147-3176.

[5]: D.Bonchev, D.H. Rouvray: Chemical Graph Theory: Introduction and Fundamentals. Gordon and Breach Science Publishers, 1990, ISBN 0-85626-454-7.

[6]: fr.wikipedia.org/wiki/Phytocide2009

[7]: www.fnehin.ca

[8]: malherbologie.cirad.fr/Fr/desherbage/index_desh.php?pageid=place - 21k

[9]: fr.wikipedia.org/wiki/Dose_létale - 22k –

[10]: fr.wikipedia.org/wiki/Dose_létale_50 - 39k Fev 2009

[11]: Environmental toxicology and chemistry , vol ;18,N° .5,pp. 1069-1075, 1999

[12]: SAR and QSAR in Environmental Research .1996.Vol.5.pp.269-279

[13]: Hyperchem ^{TM} Release 6.03 for windows, Molecular Modeling System (2000).

[14]: R. Todeschini, V. Consonni, M. Pavan. DRAGON, Software for the Calculation of Molecular Descriptors. Release 5.3 for windows, Milano (2005).

[15]: H.Kubinyi ,Quant. Struct. – Act. Relat , 13, (1994), 285.

[16] : : D.C. Montgomery, E.A. Peck, Introduction to linear Regression Analysis, Second Edition, Wiley-Interscience Publication, New York, 1992.

[17]: Mc Culloch-Pitts. a logical Calculus at the ideas imminent in Nervous Activity. Bulletin at math. Biophysics.1943, Vol. 5, p.115-133.

[18]: M. Minsky,S. Papert, Perceptrons. Massachusetts: MIT press, 1969.

[19]: D. E. Rumelbart, J. L. Mcclelland et al. Parallel Distributed processing Vol.1. Massachusetts: MIT press, 1988. 547 p.

[20]: J. J. Hopfield. Neural Networks and physical systems with emergent collective computational abilities. Proceedings of the National Academy of sciences. USA. 1982. Vol.79. p. 2554-58.

[21]: Kohonen Self-organization and associative memory. Bulletin: Springer-Verlag. 984.

[22]: R.Hecht-Nielson Neurocomputing.Addison-Wesly Publishing Company.1990. 433p.

[23] : F. Fogelmen-Soulié. Méthodes connexionnistes pour l'apprentissage.Actes des journées Nationales sur l'intelligence Artificielle. Paris: Teknea. 1988. p. 275-293.

[24]: K. HORNIK, Approximation capabilities of multilayer feedforward networks, Neural Networks, 4 (1991). 251-257.

[25]: Matlab Version 7.0.0.19920 (Release 14) The Language of Technical Computing The MathWorks, Inc. May 06, (2004).

[26]: N.R Draper, H. Smith, Applied Regression Analysis, Third Edition, Wiley series in Probability and Statistics, New york, 1998.

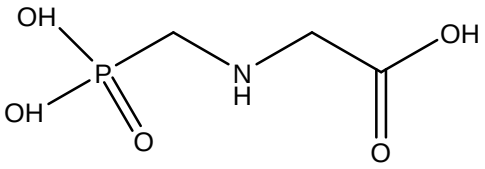
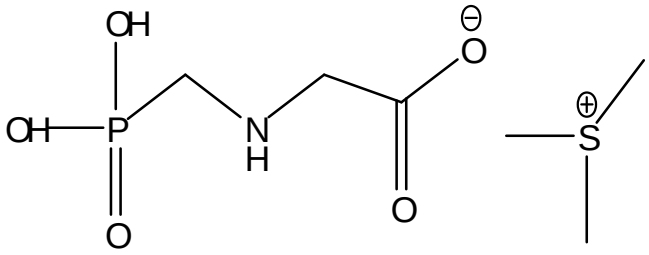
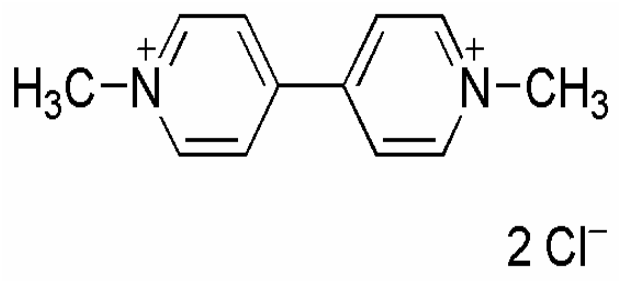
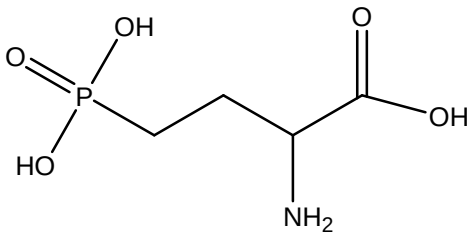
[27]:L. Eriksson, J. Jaworska, A. P. Worth, M. T. D. Cronin, R. M. Mc Dowell, P. Gramatica. Methods for Reliability and Uncertainty Assessment and for Applicability Evaluations of Classification and Regression Based QSARs. Environmental Health Perspectives 111, 1361-1375 (2003).

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

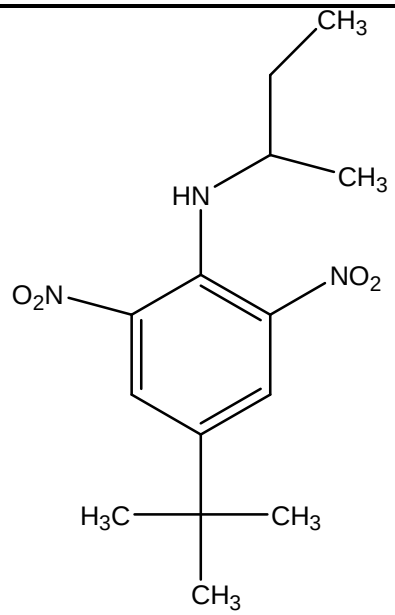
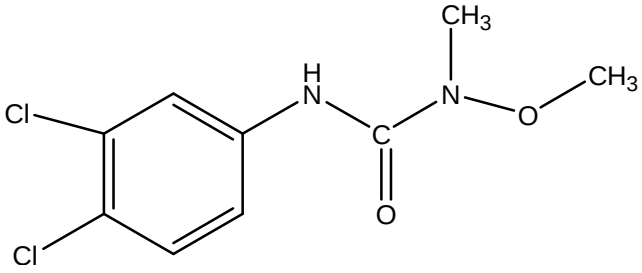
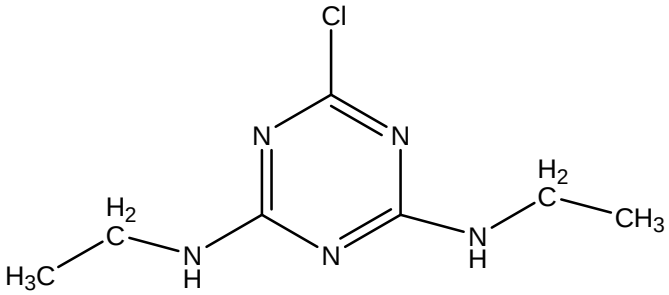
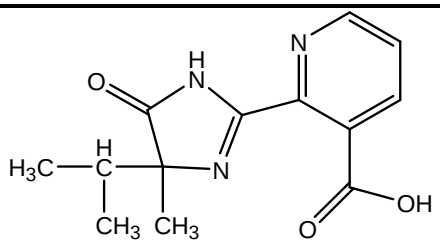
[29]: S. Weisberg, Applied linear Regression. J. Wiley, Inc., New York, 1980

Annexe2

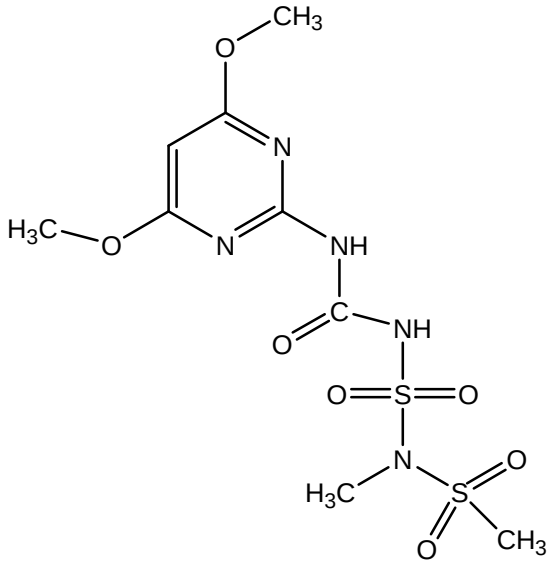
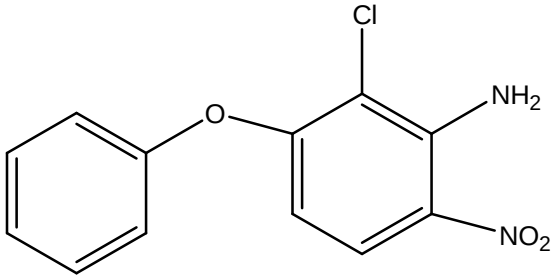
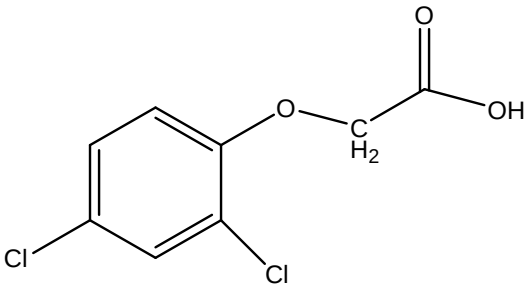
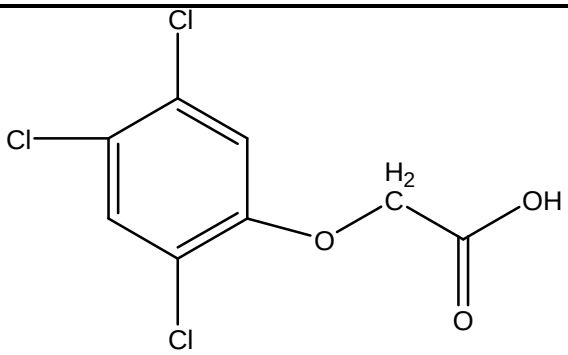
Nom usuel, structure, nom IUPAC et N° de CAS des herbicides étudiés dans l'étude bibliographique.

Nom usuel	Structure /Nom IUPAC	Famille	N° de CAS
GLYPHOSATE	 (Phosphonomethyl-amino)-acetic acid	HERBICIDE ORGANIQUE	1071-83-6
SULFASATE	 (Phosphonomethyl-amino)-acetatetrimethyl-sulfonium;	HERBICIDE ORGANIQUE	81591-81-3
PARAQUAT	 1,1'-Diméthyl-4,4'-bipyridinium	LES PYRIDINES	4685-14-7
GLUFOSINATE AMMONIUM	 2-Amino-4-phosphono-butyric acid	COMPOSE ORGANOPHOSPHORE	51276-47-2

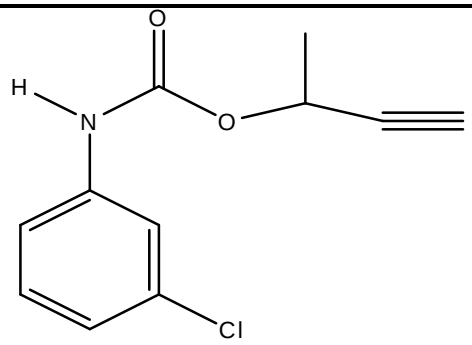
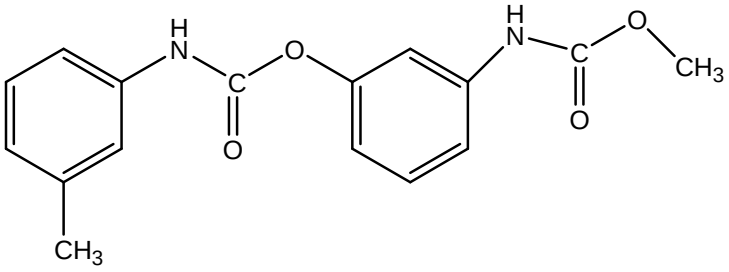
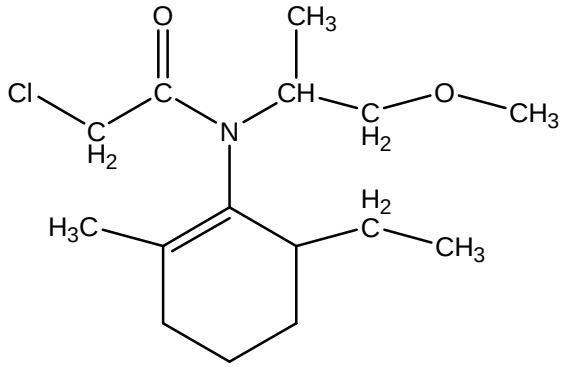
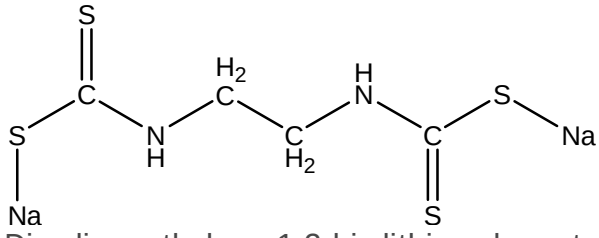
Annexe2

BUTRALINE	 *sec*-Butyl-(4-*tert*-butyl-2,6-dinitro-phenyl)-amine	LES TOLUIDINES	33629-47-9
LINURON	 1-(3,4 –DiChlorophenyl)3-methoxy-3-methyuree	UREESSUBSTITUEES	330-55-2
SIMAZINE	 6-Chloro-*N*,*N*'-diethyl-[1,3,5]triazine-2,4-diamine	LES TRIAZINES	122-34-9
IMAZAPYR	 2-(4-Isopropyl-4-methyl-5-oxo-4,5-dihydro-1*H*-imidazol-2-yl)-nicotinic acid	IMIDAZOLINONES	81334-34-1

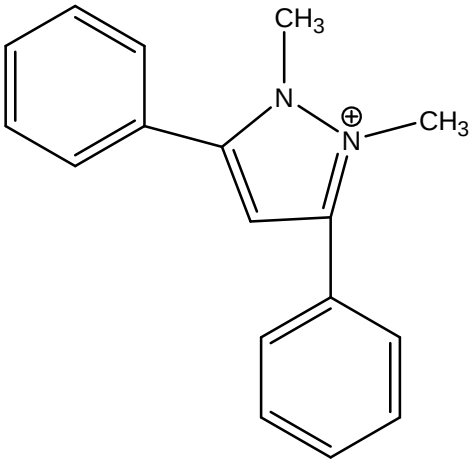
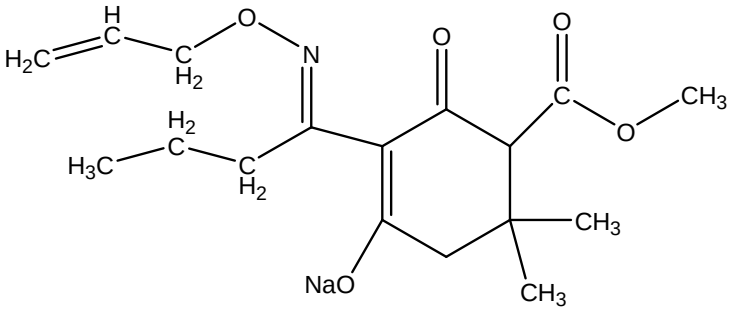
Annexe2

AMIDOSULFON		SULFONYLUREES	120923-37-7
ACLONIFEN	 2-Chloro-6-nitro-3-phenoxy-phenylamine	DIPHENYLETHERS	74070-46-5
2,4-D	 Acide dichloro 2,4 phénoxyacétique	HERBICIDES FOLIAIRES	94-75-7
DNOC	 Acide triChloro 2,4,5 phénoxyacétique	COLORANT NITRE	534-52-1

Annexe2

CLOROBUFAME	 But-3-yn-2-yl 3-Chlorophenylcarbamate	CABAMATE	1967-16-4
PHENMEDIPHAME		PHENYLCARBAMATE	13684-63-4
BUTILATE	 2-Chloro-6-ethyl-N-(2-methoxy-1-methylethyl)aceto-to	THIOCARBAMATE	2008-41-5
NABAME	 Disodium ethylene-1,2-bisdithiocarbamate	DITHIOCARBAMATE	142-59-6

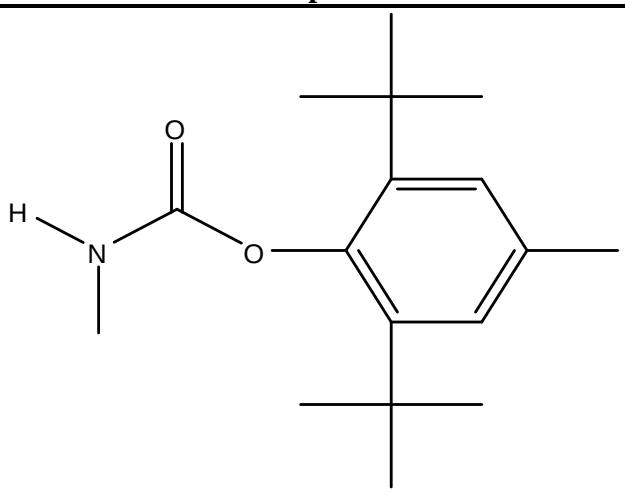
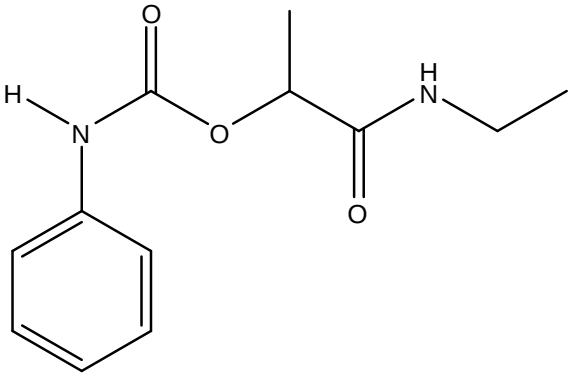
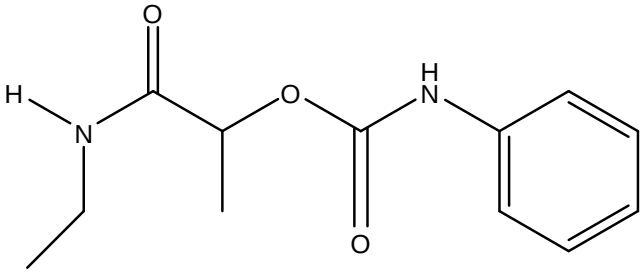
Annexe2

DIPHENZOQUAT	 1,2-dimethyl-3,5-diphenylpyrazolium	AMMONIUM QUATERENAIRE	49866-87-7
ALLOXYDIME-SODIUM		FOP/DIME ET PINOXADENE	55635-13-7

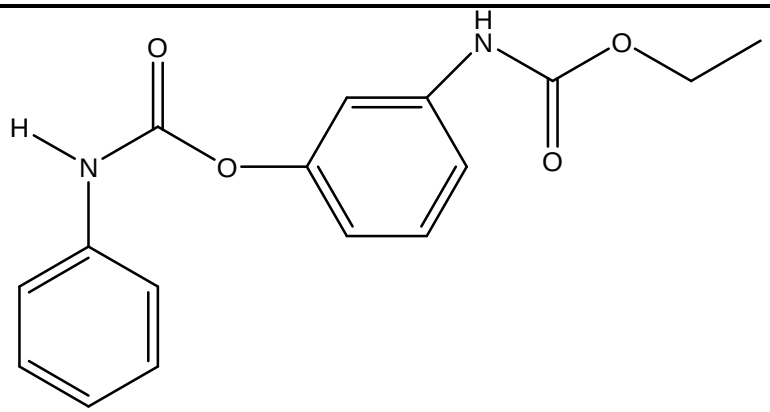
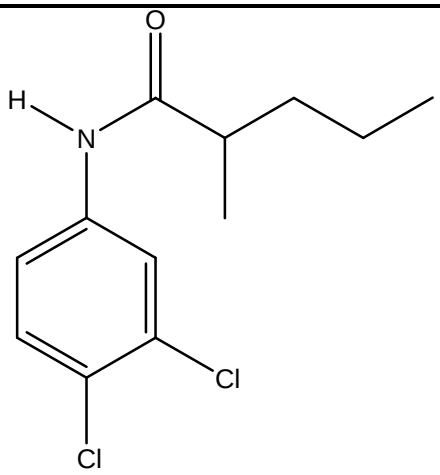
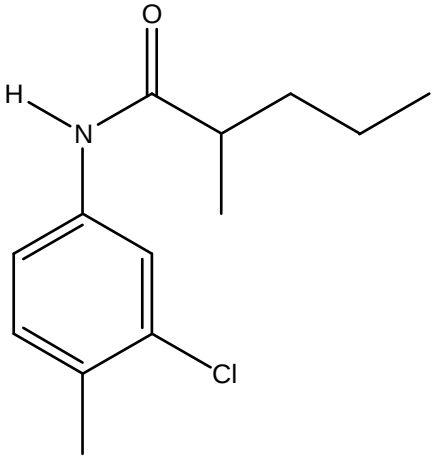
ANNEXES

Annexe 1

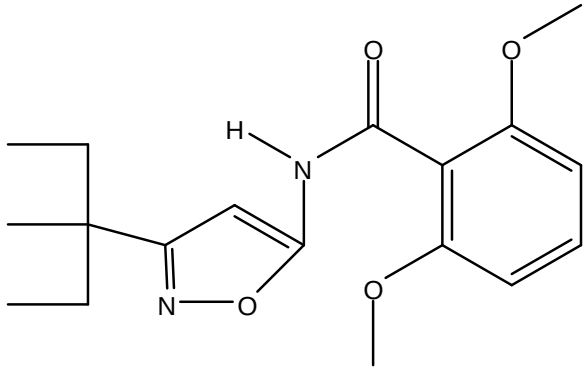
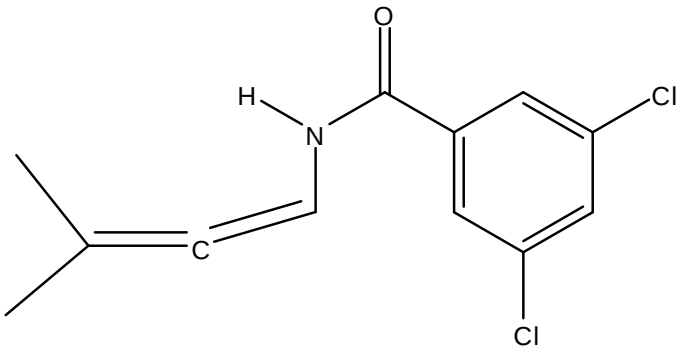
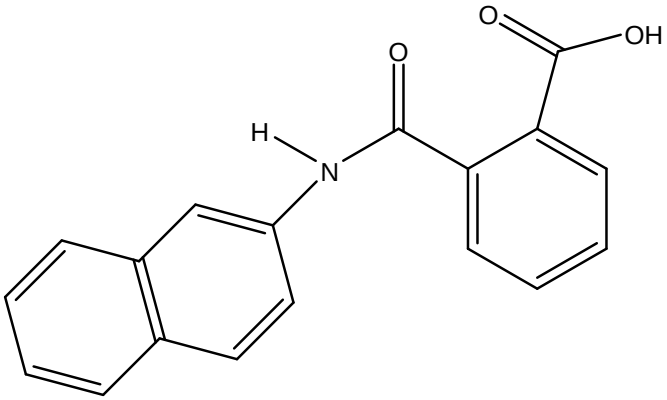
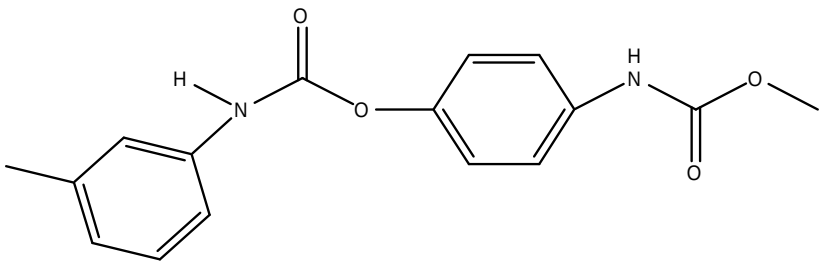
Structure, N° de CAS et toxicité des composés étudiés.

N°	Composés	N° DE CAS	pDL_{50}
1	 2,6-di-*tert*-butyl-4-methylphenyl methylcarbamate	1918-11-2	2,10
2	 1-(ethylamino)-1-oxopropan-2-yl phenylcarbamate	16118-49-3	1,67
3	 1-(ethylamino)-1-oxopropan-2-yl phenylcarbamate	16118-49-3	1,67

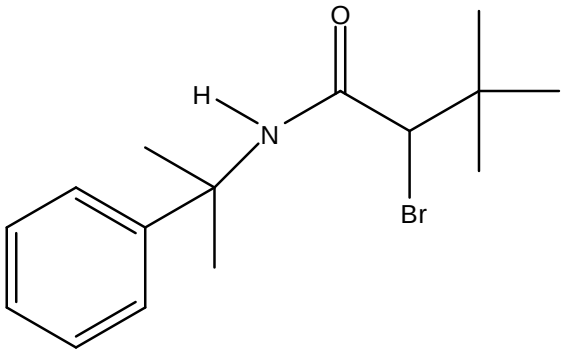
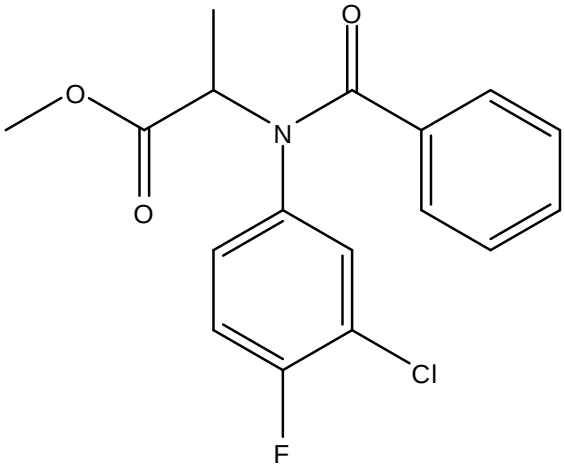
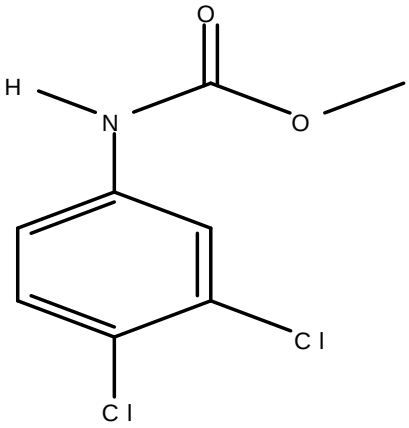
Annexe 1

4	 (3-Phenylcarbamoyloxy-phenyl)-carbamic acid ethyl ester	13684-56-5	2,42
5	 2-Methyl-pentanoic acid (3,4-dichloro-phenyl)-amide	/	1,58
6	 2-Methyl-pentanoic acid (3-chloro-4-methyl-phenyl)-amide	2307-68-8	0,53

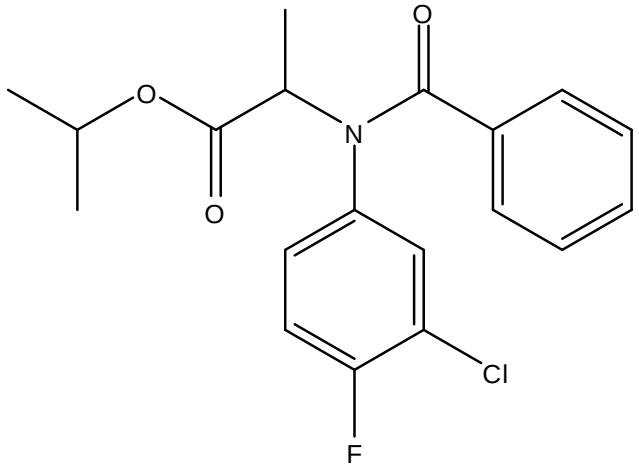
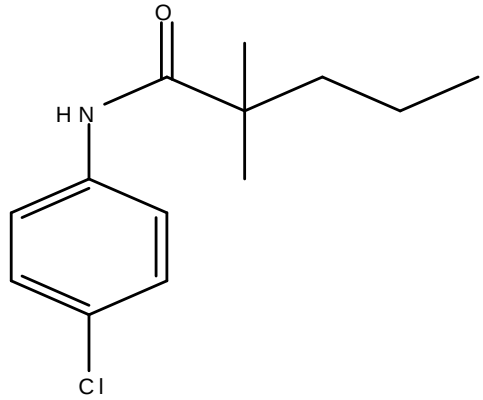
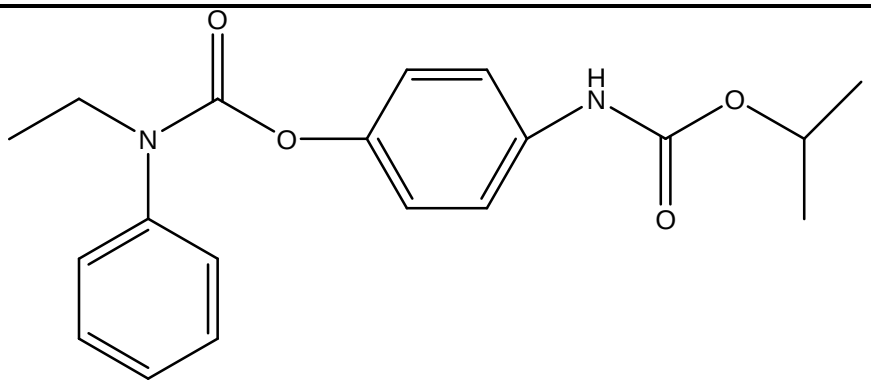
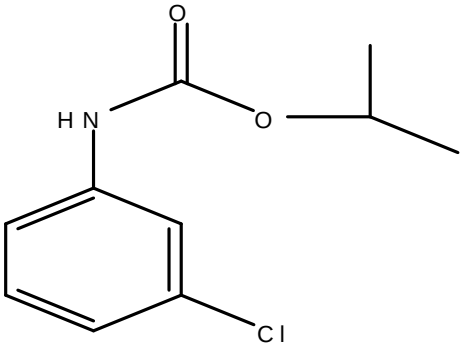
Annexe 1

7	 2,6-dimethoxy-*N*-(3-(3-methylpentan-3-yl)isoxazol-5-yl)benzamide	82558-50-7	1,48
8	 3,5-dichloro-*N*-(3-methylbuta-1,2-dienyl)benzamide	/	1,59
9	 2-(naphthalen-2-ylcarbamoyl)benzoic acid	132-66-1	1,45
10	 Carbamic acid, (3-methylphenyl)-, 3-[(methoxycarbonyl)amino]phenyl] ester	13684-63-4	1,43

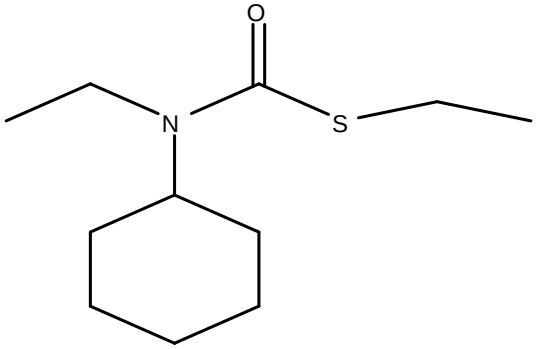
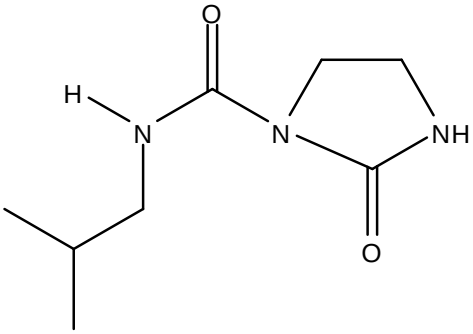
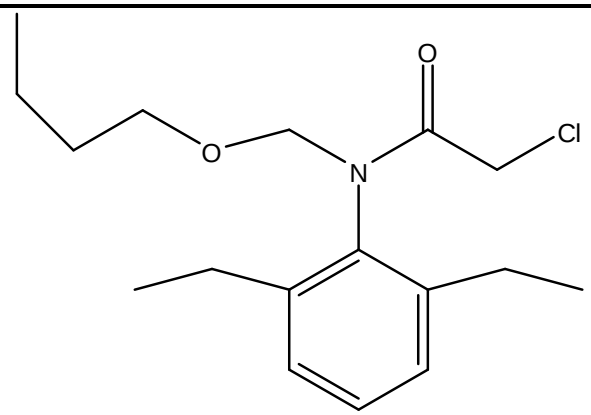
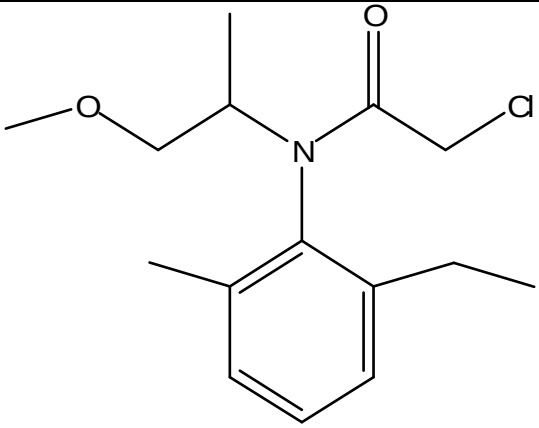
Annexe 1

11	 2-bromo-3,3-dimethyl-*N*-(2-phenylpropan-2-yl)butanamide	74712-19-9	1,2
12	 methyl 2-(*N*-(3-chloro-4-fluorophenyl)benzamido)propanoate	57973-66-7	1,17
13	 methyl 3,4-dichlorophenylcarbamate	1918-18-9	0,38

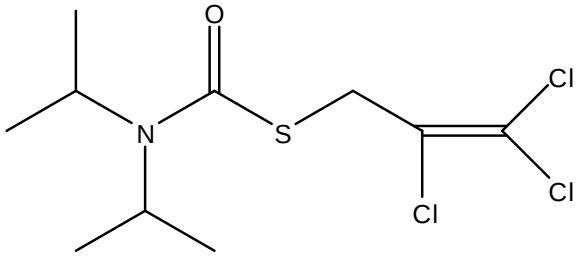
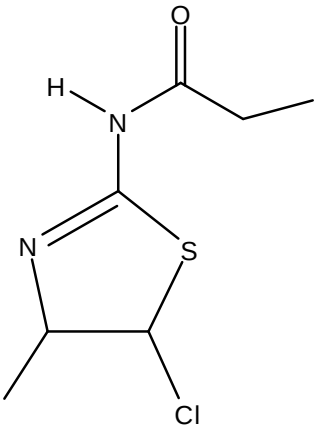
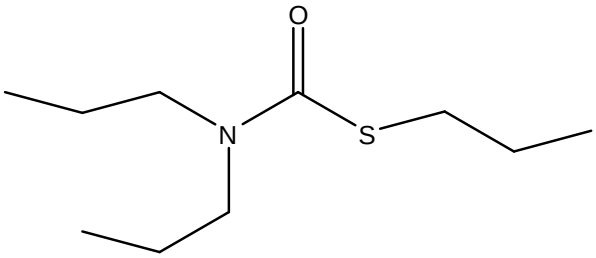
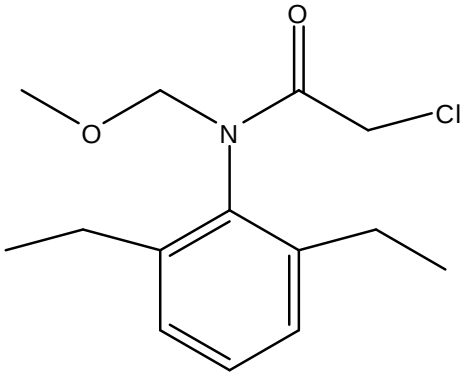
Annexe 1

14	 isopropyl 2-(*N*-(3-chloro-4-fluorophenyl)benzamido)propanoate	63782-90-1	1,05
15	 *N*-(4-chlorophenyl)-2,2-dimethylpentanamide	7287-36-7	1,22
16	 [4-(Ethyl-phenyl-carbamoyloxy)-phenyl]-carbamic acid isopropyl ester	57375-63-0	1,07
17	 isopropyl 3-chlorophenylcarbamate	101-21-3	1,25

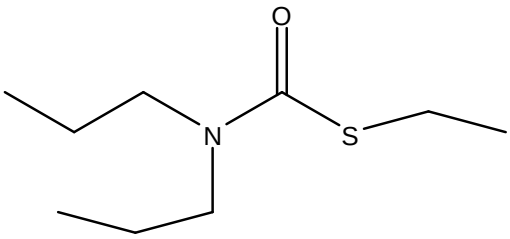
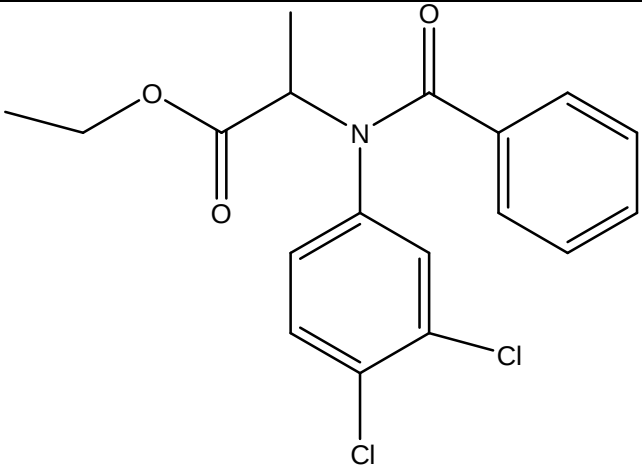
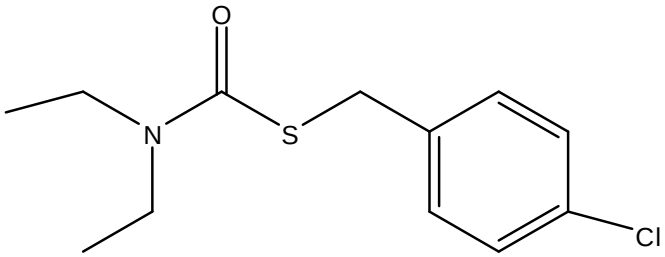
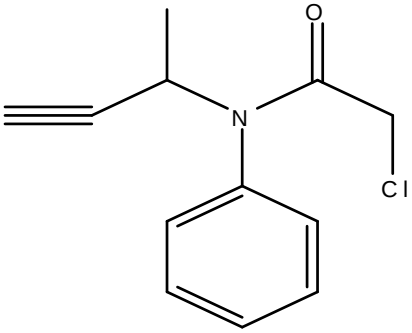
Annexe 1

18	 *S*-ethyl cyclohexyl(ethyl)carbamothioate	1134-23-2	1,22
19	 *N*-isobutyl-2-oxoimidazolidine-1-carboxamide	30979-48-7	1,13
20	 *N*-(butoxymethyl)-2-chloro-*N*-(2,6-diethylphenyl)acetamide	23184-66-9	1,02
21	 2-chloro-*N*-(2-ethyl-6-methylphenyl)-*N*-(1-methoxypropan-2-yl)acetamide	51218-45-2	0,99

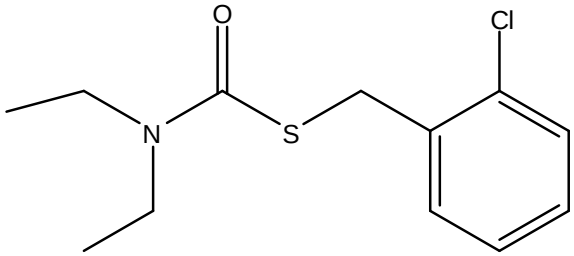
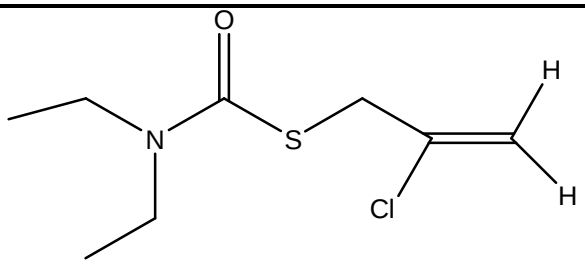
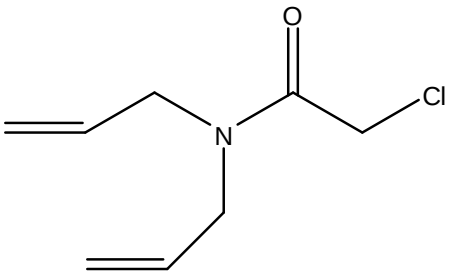
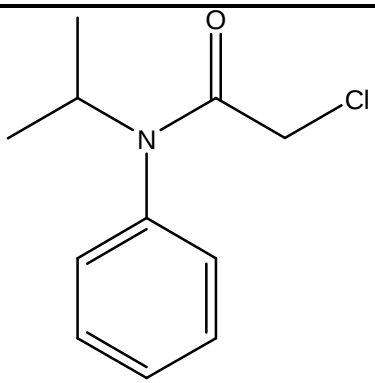
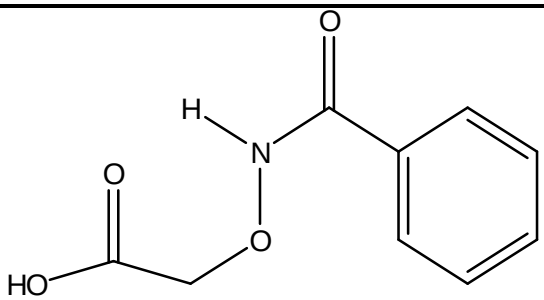
Annexe 1

22	 *S*-2,3,3-trichloroallyl diisopropylcarbamothioate	2303-17-5	0,74
23	 *N*-(5-chloro-4-methyl-4,5-dihydrothiazol-2-yl)propionamide	/	1,00
24	 *S*-propyl dipropylcarbamothioate	1929-77-7	0,94
25	 2-chloro-*N*-(2,6-diethylphenyl)-*N*-(methoxymethyl)acetamide	15972-60-8	0,75

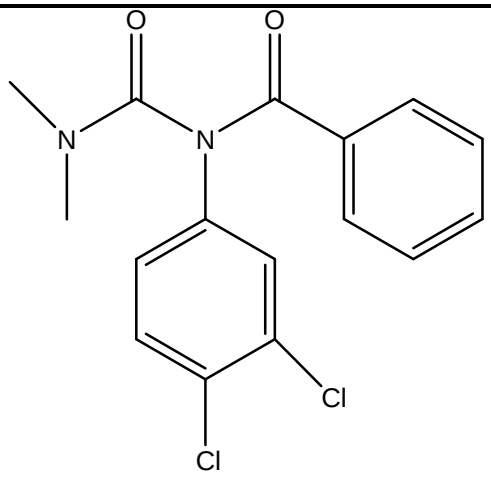
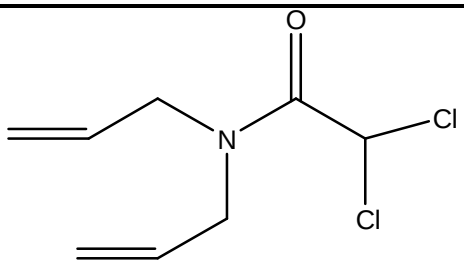
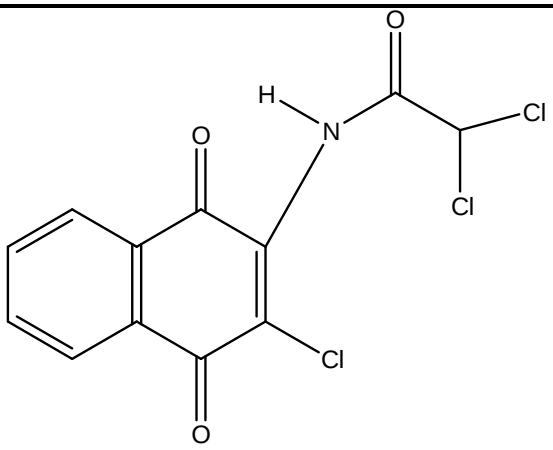
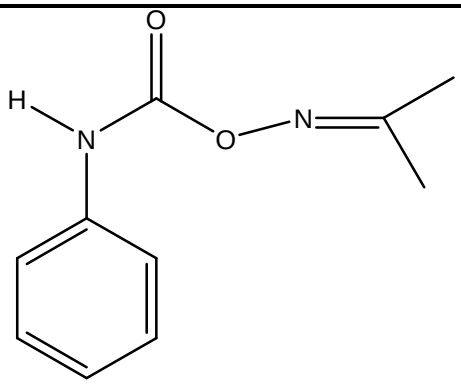
Annexe 1

26	 *S*-ethyl dipropylcarbamothioate	759-94-4	0,94
27	 Ethyl 2-(*N*-(3,4-dichlorophenyl)benzamido)propanoate	22212-55-1	0,63
28	 *S*-4-chlorobenzyl diethylcarbamothioate	28249-77-6	0,7
29	 *N* -(but-3-yn-2-yl)-2-chloro-*N* -phenylacetamide	21267-72-1	0,73

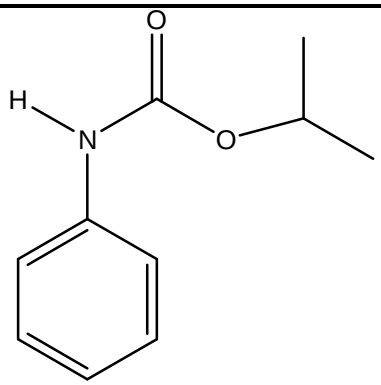
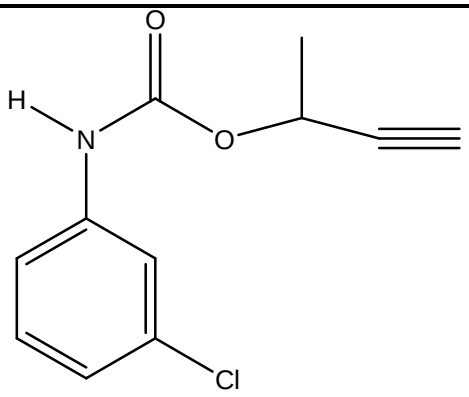
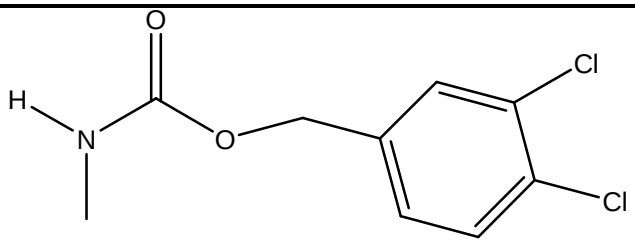
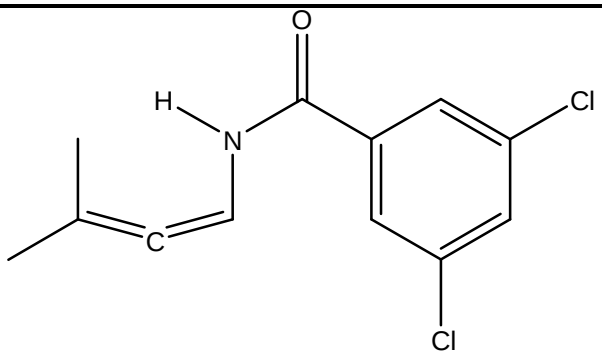
Annexe 1

30	 *S*-2-chlorobenzyl diethylcarbamothioate	34622-58-7	0,63
31	 *S*-2-chloroallyl diethylcarbamothioate	/	0,61
32	 *N*,*N*-diallyl-2-chloroacetamide	93-71-0	0,62
33	 2-Chloro-*N*-isopropyl-*N*-phenylacetamide	1918-16-7	0,53
34	 (2-Benzamidooxy)acetic acid	5251-93-4	1,46

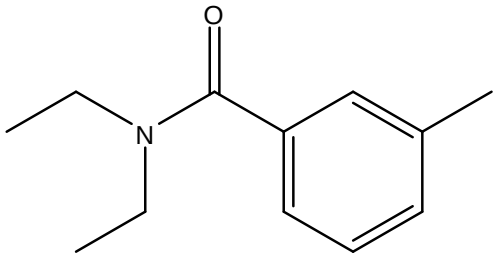
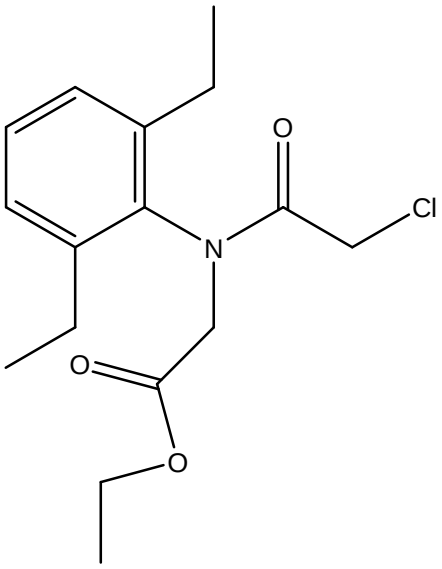
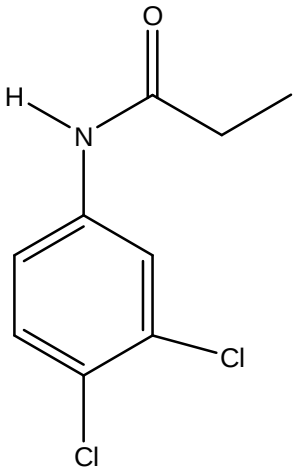
Annexe 1

35	 *N*-(3,4diChlorophenyl)-*N*-(dimethylcarbamoyl)benzamide	3134-12-1	1,17
36	 *N*,*N*-diallyl-2,2diChloroacetamide	37764-25-3	0,98
37	 2,2-diChloro-*N*-(3-Chloro-1,4-diOxo-1,4-dihydronaphthalen-2-yl)acetamide	27541-88-4	1,67
38	 Propan-2-one *O*-phenylcarbamoyl oxime	/	0,90

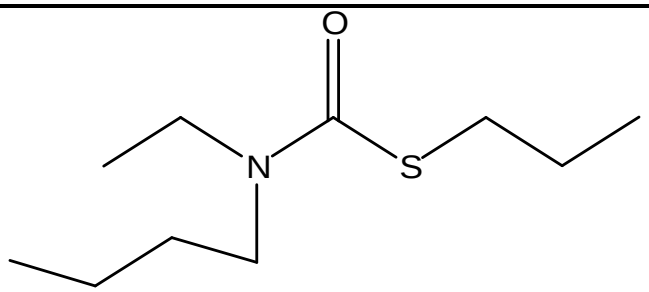
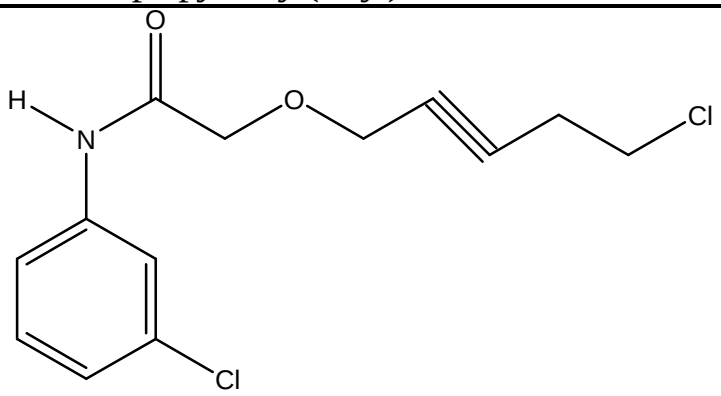
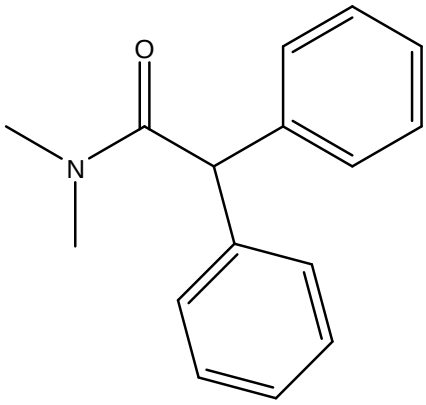
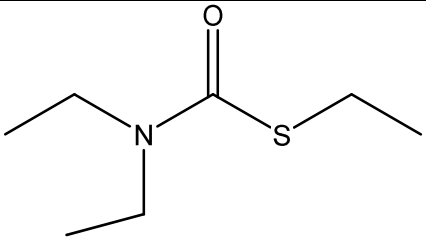
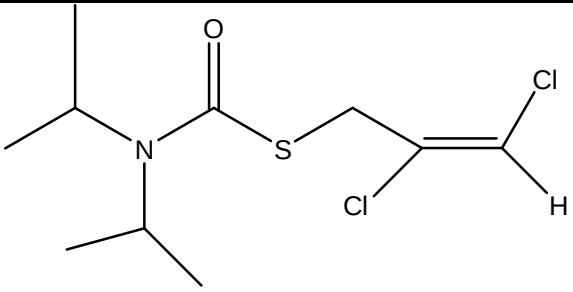
Annexe 1

39	 Isopropyl phenylcarbamate	122-42-9	1,45
40	 But-3-yn-2-yl 3-Chlorophenylcarbamate	1967-16-4	1,03
41	 3,4-diChlorobenzyl methylcarbamate	1966-58-1	0,92
42	 3,5-diChloro-*N*-(3-methylbuta-1,2-dienyl)benzamide	/	1,51

Annexe 1

43	 *N*,*N*-diethyl-3-methylbenzamide	134-62-3	1,02
44	 Ethyl 2-(2Chloro-*N*-(2,6-diethylphenyl)acetamido)acetate	/	0,87
45	 *N*-(3,4-diChlorophenyl)propionamide	709-98-8	0,81

Annexe 1

46	 *S*-propyl butyl(ethyl)carbamothioate	1114-71-2	0,77
47	 2-(5-Chloropent-2-ynyloxy)-*N*-(3-Chlorophenyl)acetamide	/	0,70
48	 *N*,*N*-dimethyl-2,2-diphenylacetamide	957-51-7	0,62
49	 *S*-ethyl diethylcarbamothioate	2941-55-1	0,39
50	 (*E*)-*S*-2,3-diChloroallyl diisopropylcarbamothioate	2303-16-4	0,16