THÈSE DE DOCTORAT DE L'UNIVERSITÉ PIERRE ET MARIE CURIE

Présentée le 25 septembre 2013 par

Vinca Prana

Pour obtenir le grade de

DOCTEUR de l'UNIVERSITÉ PIERRE ET MARIE CURIE

École Doctorale Chimie Physique et Analytique de Paris VI

Spécialité Chimie théorique et informatique

Approches structure-propriété pour la prédiction des propriétés physico-chimiques des substances chimiques

devant le jury composé de :

Pr. Henry Chermette	Rapporteur
Pr. Alexandre Varnek	Rapporteur
Pr. Esmaïl Alikhani	Examinateur
Dr. David André	Examinateur
Dr. Guillaume Fayet	Examinateur
Dr. Patricia Rotureau	Encadrante INERIS
Pr. Carlo Adamo	Directeur de thèse

"There are three kinds of lies: lies, damned lies, and statistics."

Mark Twain - Chapters from My Autobiography

REMERCIEMENTS

Je voudrais tout d'abord remercier mon directeur de thèse, M. Carlo Adamo de m'avoir donné l'occasion de travailler dans son équipe de Modélisation des Systèmes Complexes de Chimie ParisTech où il encourage l'esprit de convivialité ainsi que pour la confiance qu'il m'a accordée.

J'adresse aussi mes remerciements à ma co-encadrantre de thèse INERIS, Mme Patricia Rotureau, ainsi qu'à M. Guillaume Fayet pour avoir supervisé mon travail.

Je tiens également à remercier Messieurs Henry Chermette et Alexandre Varnek d'avoir accepté d'être les rapporteurs de ce manuscrit ainsi qu'aux examinateurs Messieurs Esmaïl Alikhani et David André d'avoir accepté de faire partie de mon jury de thèse.

Je remercie Ilaria pour ses conseils et ses recommandations, notamment pour l'avenir. Merci à l'ensemble de l'équipe du MSC avec qui j'ai passé de bons moments : Fred (le spécialiste crystal aussi chef des TD), Diane pour les discussions de la pause thé/infusion, Romain, Bertrand et tous ceux qui ont fait des passages plus ou moins longs dans l'équipe. Mais je remercie aussi ceux qui, devenus docteurs, sont parti vers de nouveaux horizons : Cyril, Tangui, Aurélien pour les discussions de fin de journée, Giuseppe, Éric pour son initiation au python et sa bonne humeur quotidienne, Amel pour sa gentillesse (et ses délicieux gâteaux !) et Nils pour ses conseils littéraires et son aide.

Je n'ai pas oublié celles avec qui j'ai eu la chance de profiter du bon air de Verneuil et de partager les trajets Paris-Creil. Un grand merci à Stefania pour son aide à mes débuts et ses bons conseils en termes de séries et Stefanina pour les « sessions zumba® » et les soirées pizza.

Mes remerciements vont aussi à mes parents, Elisa et Yohan pour m'avoir soutenu, chacun à sa manière, ainsi que pour leur patience aux cours de ces années dont le nombre dépasse (largement) trois.

SOMMAIRE

LISTE DES ACRONYMES

ADR : « Accord européen relatif au transport international des marchandises dangereuses par route » est entré en vigueur le 29 janvier 1968

BMLR : « Best Multi-Linear Regression » est une méthode de sélection des descripteurs pour l'obtention d'un modèle MLR.

CAllISto : « Conformational Analysis In Silico » est un programme en python permettant de réduire le nombre de conformations pour une molécule par une méthode de clustering.

CLP : « Classification, Labelling and Packaging of substances and mixtures » est le nouveau système de classification, d'étiquetage et d'emballage de substances et mélanges entré en vigueur en Europe le 20 janvier 2009.

DFT : « Density Functional Theory ». La théorique de la fonctionnelle de la densité est une méthode de chimie quantiques basé sur le postulat suivant : le système peut être caractérisé par la densité électronique.

ECHA : « European chemical Agency » est l'agence européenne créée pour coordonner la mise en place du nouveau règlement REACH au sein de l'Union Européenne.

FDS : « Fiche de données de sécurité » est une fiche technique contenant des données relatives aux propriétés d'une substance chimique. Elle est composé de 18 points réglementaires et obligatoires tels que : l'identification du produit et des dangers, la composition, conditions de stockage et d'utilisation…

HOMO/LUMO : Highest occupied molecular orbital est l'orbitale la plus haute occupée et lowest unoccupied molecular orbital est l'orbitale la plus basse vacante.

MAE : pour Mean Absolute Error ou erreur absolue moyenne (équation chapitre 3)

MLR : (Multi-Linear Regression) ou régression multi-linéaire est une méthode statistique de régression. Elle permet de relier linéairement une variable dépendante Y avec une série de variables indépendantes X_i.

OCDE : « Organisation de Coopération et de Développement Economiques » est une organisation internationale, née le 30 septembre 1961, dont la mission est de promouvoir les politiques qui amélioreront le bien-être économique et social partout dans le monde.

PREDIMOL : « PREDIction des propriétés physico-chimiques des produits par modélisation MOLéculaire » est le projet ANR qui finance cette thèse (voir Chapitre 1).

PCA : (Principal Component Analysis) ou ACP pour Analyse en composantes principales est une méthode d'analyse de données qui permet de réduire le nombre de variable. Elle consiste à transformer des variables liées entre elles en nouvelles variables indépendantes nommées « composantes principales ».

QMRF : « QSAR Model Reporting Format » est un modèle harmonisé qui récapitule et rapporte les informations clés sur les modèles QSAR, y compris les résultats des études de validation. L'information est structurée selon les principes de validation de l'OCDE.

QSPR : Quantitative Structure-Property Relationship est un modèle mathématique qui relie la structure d'une molécule à ses propriétés.

REACH : « Registration, Evaluation, Authorisation and Restriction of Chemicals » est le règlement européen concernant l'enregistrement, l'évaluation et l'autorisation des substances chimique, importées ou produites à plus d'une tonne par an, entré en vigueur le 1er juin 2007.

RMSD : Root Mean Square Deviation (ou écart quadratique moyen) est la mesure de la distance moyenne entre les atomes de deux molécules superposées. Dans l'étude des conformations il s'agit d'une mesure de similarité courante.

RMSE : Root Mean Square Error ou erreur quadratique moyenne (équation chapitre 3)

SGH : « Globally Harmonized System for classification and labelling of chemical products », adopté par les Nations Unies en juillet 2003, vise à harmoniser sur le plan international les critères de

classification et à communiquer, au moyen de l'étiquetage, les dangers liés aux produits chimiques afin d'en garantir la sécurité d'emploi. Il s'agit du texte de base du règlement CLP.

SMILES : « Simplified molecular-input line-entry system » a été inventé par Weininger, il s'agit d'un langage à ligne de texte pour représenter les molécules. A partir d'une chaîne de caractère, la molécule associée peut être convertie en 2D ou 3D. Cette notation permet de construire des bases de données simples.

TDAA : la Température de Décomposition Auto-Accélérée est la température la plus basse à laquelle un peroxyde organique dans son emballage peut subir une décomposition auto-accélérée. Cette température permet de définir la température de stockage et la zone de températures d'utilisation.

TNO : « Nederlandse Organisatie voor Toegepast Natuurwetenschappelijk Onderzoek » (ou « Organisation néerlandaise pour la recherche scientifique appliquée » en français) est un institut de recherche appliquée situé au Pays-Bas pour les entreprises, les organismes gouvernementaux et les organisations publiques.

TMD : arrêté du 29 mai 2009 relatif au « Transport de Marchandises Dangereuses par voies terrestres ».

CHAPITRE 1 : CONTEXTE ET OBJECTIFS

Cette thèse, portée par le projet ANR PREDIMOL (PREDIction des propriétés physico-chimiques des produits par modélisation MOLéculaire), se situe dans un contexte lié au nouveau règlement européen REACH et aux conséquences qu'il entraîne en termes de développement de méthodes alternatives à l'expérimentation. L'objectif est de développer des modèles QSPR prédictifs visant à caractérisier des propriétés physico-chimiques des substances chimiques à enregistrer. En plus de modèles classiques, des modèles simplifiés utilisant des descripteurs calculables sans logiciel seront développés pour permettre une utilisation de ces modèles pour les industriels et leur acceptation règlementaire plus facile.

Dans cette première partie, la réglementation Européenne sur les substances chimiques ainsi que le projet PREDIMOL[1] seront introduits. L'origine et le but de cette réglementation seront exposés. Les peroxydes organiques, qui sont les composés chimiques étudiés dans ces travaux de thèse seront présentés. Enfin, le plan du manuscrit sera détaillé.

I. RÉGLEMENTATIONS EUROPÉENNES DES SUBSTANCES CHIMIQUES

Les produits chimiques, qu'ils soient naturels ou synthétiques, sont présents partout dans notre environnement. Ils ont permis une amélioration[2] de notre mode de vie avec par exemple les médicaments dont certains permettent l'augmentation de l'espérance de vie, la réalisation et la coloration de nouveaux matériaux tel que ceux utilisés pour la fabrication de vêtements. Cependant les produits chimiques sont souvent considérés comme mauvais, notamment au niveau alimentaire malgré les améliorations que la chimie alimentaire a générées telles que l'augmentation de la durée de conservation, la pasteurisation ou la synthèse de produits rares (par exemple suite au blocus continental imposé par Napoléon contre le Royaume-Uni en 1806, le sucre de canne a pu être remplacé par le sucre de betterave). Le mot chimie fait peur, il est associé aux risques (accidents industriels, toxicité…). La gestion des risques chimiques est donc devenue une préoccupation majeure.

1. Le règlement REACH

Le règlement Européen REACH[3] (Registration, Evaluation, Authorisation and Restriction of Chemicals), entré en vigueur le 1er juin 2007, est le résultat d'une coopération des pays européens dans le but d'uniformiser la classification des substances chimiques et d'améliorer leur contrôle. La meilleure connaissance des substances chimiques permettra de mieux maîtriser les risques liés à leur usage et, en cas de besoin, de restreindre ou interdire leur emploi dans le cadre du principe de précaution. L'histoire de REACH commence dans la ville de Seveso au Nord de l'Italie où la première catastrophe industrielle chimique eu lieu sous la forme d'une pollution à la dioxine[4] le 10 juillet 1976. Suite à cette catastrophe la directive 96/82/CE dite directive Seveso est mise en place. Cette directive impose l'identification des sites industriels présentant des risques d'accidents majeurs. En 1981, la directive européenne 67/548/CEE concernant la classification, l'emballage et l'étiquetage des substances dangereuses[5] soumet tous les nouveaux produits chimiques à une évaluation exhaustive des risques à la charge des producteurs. Cependant, celle-ci a comme principal effet l'utilisation des « anciennes » substances par les industriels afin d'engendrer des coûts moindres. Ainsi, le 27 février 2001 a été adopté le livre blanc « Stratégie pour la future politique dans le domaine des substances chimiques »[6], qui supprime la différence entre les produits « anciens » (c'est-à-dire mis sur le marché avant 1981) et les « nouveaux ». Ce livre blanc, dernière étape avant REACH, permet aussi d'avoir un règlement unique en Europe en remplaçant une quarantaine de textes législatifs relatifs aux substances chimiques.

Le règlement REACH présente quatre objectifs majeurs:

- La protection de la santé humaine et la protection environnementale face aux risques potentiels des substances chimiques ;
- La compétitivité et l'innovation de l'industrie chimique européenne ;
- La libre circulation des substances sur le marché intérieur de l'Union européenne ;
- La promotion des méthodes alternatives pour l'évaluation des dangers des substances.

2. Les procédures de REACH

Le règlement REACH s'applique à toutes les substances chimiques produites ou importées dans l'Union européenne à plus d'une tonne par an. Selon le calendrier prévisionnel (en Figure 1), d'ici le 31 mai 2018 ces substances existantes devront toutes être enregistrées auprès de l'Agence Européenne des Produits Chimiques (ECHA[7]) afin d'avoir l'autorisation d'être mises sur le marché européen. La mise en œuvre de cette réglementation et donc de l'enregistrement des substances relève de la responsabilité des industriels. Le règlement REACH met en œuvre quatre procédures principales[8] : l'enregistrement, l'évaluation, l'autorisation et la restriction qui sont décrites ici.

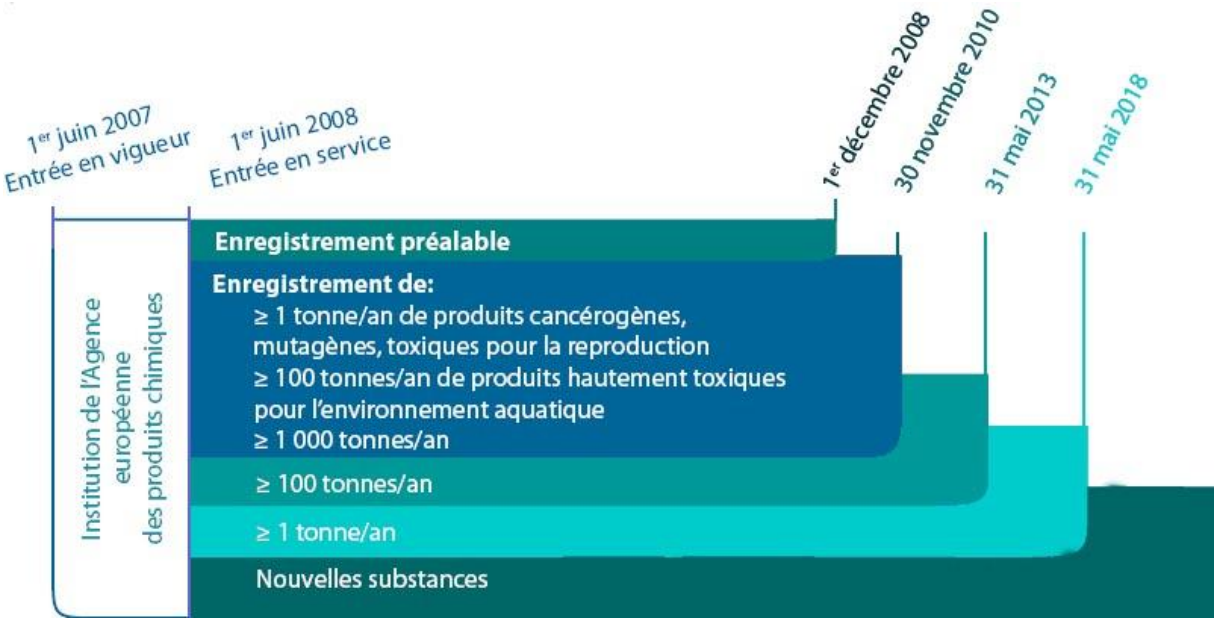


Figure 1 : Calendrier pour l'enregistrement des substances dans le cadre de la réglementation REACH

a) Enregistrement et pré-enregistrement

L'enregistrement constitue l'élément fondamental de REACH. À terme, aucune substance soumise à l'obligation d'enregistrement ne pourra être fabriquée, ni importée, sans avoir été précédemment enregistrée, conformément au principe « Pas de données, pas de marché » (article 5 de la réglementation).

Une première phase de pré-enregistrement est intervenue entre le 1er juin et le 1er décembre 2008. Il s'agissait pour les fabricants et importateurs de préenregistrer les substances circulant déjà sur le marché. Cela a permis aux déclarants, non seulement de pouvoir enregistrer ces substances

ultérieurement mais aussi d'échanger des données afin d'éviter la réalisation d'essais redondants. Au total, ce sont plus de 143 000 substances qui ont été pré-enregistrées par 65 000 industriels.

La véritable phase d'enregistrement des substances existantes produites et importées à plus d'une tonne par an est actuellement en cours. La première étape de l'enregistrement s'est déroulée entre le 1er décembre 2008 et le 30 novembre 2010. Le 20 décembre 2012, 7 884 substances uniques pour 30 601 dossiers étaient enregistrées[7] et 3 000 substances ont été répertoriées comme susceptibles d'être enregistrées en 2013[7,9]. Le 31 mai 2013 nous sommes arrivés à la fin de l'étape d'enregistrement des substances existantes produites et importées à plus de 100 tonnes par an. Au 1er juin 2018 toutes les substances existantes devront être enregistrées. Les substances nouvelles, quant à elles, sont et resteront soumises à enregistrement au-delà de ce délai. Les fabricants et importateurs constituent pour chaque substance un dossier d'enregistrement qui est ensuite soumis à l'ECHA. Les informations à notifier dans ce dossier comprennent différents éléments tels que :

- l'identité et les coordonnées du producteur ou de l'importateur ;
- le ou les numéros d'enregistrement ;
- l'identité de la ou des substances ;
- la classification de la ou des substances ;
- une brève description de la ou des utilisations de la ou des substances ;
- la fourchette de quantité de la ou des substances.

Le dossier d'enregistrement précise les usages prévus pour la substance concernée, encadrant ainsi les usages de la substance par les utilisateurs en aval.

b) Évaluation

L'évaluation des dossiers par l'ECHA comporte deux volets : le contrôle de la conformité des dossiers et celui des essais proposés par les déclarants. L'ECHA indique alors au déclarant s'il doit réaliser l'essai proposé, le modifier ou ne pas réaliser d'essai. L'évaluation des substances se fait en coopération avec les autorités compétentes des états membres.

c) Autorisation et restriction

Si l'évaluation conclut à une substance préoccupante ou extrêmement préoccupante, elle est ajoutée à la liste de ces substances dite « liste candidate » que l'ECHA publie et met à jour régulièrement. Pour ces substances une autorisation est demandée. Le but de l'autorisation est de garantir que les risques relatifs de ces substances soient valablement maîtrisés et que ces substances soient progressivement remplacées par d'autres substances ou technologies appropriées, lorsque celles-ci sont économiquement et techniquement viables.

La procédure de restriction est « un filet de sécurité permettant de gérer les risques qui ne sont pas couverts de manière adéquate par d'autres dispositions du système REACH »[10]. Elle permet de limiter

la fabrication, l'utilisation ou la mise sur le marché de substances entraînant un risque non acceptable. Les restrictions sont suggérées par les États membres ou par l'ECHA.

3. Règlements associés : CLP et TMD

En parallèle du règlement REACH, une règlementation concernant non seulement le classement des substances seules mais aussi les mélanges a été mise en place en décembre 2008. Le règlement CLP[11] (pour *Classification, Labelling and Packaging of substances and mixtures*) est le nouveau système de classification, d'étiquetage et d'emballage de substances et mélanges entré en vigueur en Europe le 20 janvier 2009. Il s'agit de l'application en Europe des recommandations internationales du règlement SGH[12] (pour *Système Général Harmonisé*), ayant été élaborées à partir de systèmes de classification et d'étiquetage existants afin de créer un modèle unique à l'échelle mondiale.

Ce règlement définit de nouvelles règles applicables en matière de classification, d'étiquetage et d'emballage des produits. Aussi 28 classes de danger (16 de danger physique, 10 de danger pour la santé, 2 de danger pour l'environnement) associées à 9 pictogrammes (Tableau 1) ont été définies. À l'intérieur des classes, les catégories indiquent les niveaux de danger.

Tableau 1 : Pictogrammes et classes de danger associées dans le règlement CLP[11]

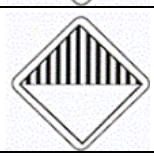
PICTOGRAMMES	CLASSES DE DANGER
	• Explosibles • Substances et mélanges autoréactifs • Peroxydes organiques
	• Gaz inflammables • Aérosols inflammables • Liquides inflammables • Matières solides inflammables • Substances et mélanges autoréactifs • Liquides pyrophoriques • Matières solides pyrophoriques • Substances et mélanges auto-échauffants • Substances et mélanges qui, au contact de l'eau, dégagent des gaz inflammables • Peroxydes organiques
	• Gaz comburants • Liquides comburants • Matières solides comburantes
	• Gaz sous pression

PICTOGRAMMES	CLASSES DE DANGER
	• Substances et mélanges corrosifs pour les métaux • Corrosion/irritation cutanée • Lésions oculaires graves/irritation oculaire
	• Toxicité aiguë
	• Toxicité aiguë • Corrosion/irritation cutanée • Lésions oculaires graves/irritation oculaire • Toxicité spécifique pour certains organes cibles - exposition unique
	• Sensibilisation respiratoire • Mutagénicité sur les cellules germinales • Cancérogénicité • Toxicité pour la reproduction • Toxicité spécifique pour certains organes cibles - exposition unique • Toxicité spécifique pour certains organes cibles - exposition répétée • Danger par aspiration
	• Danger pour le milieu aquatique

Ce nouveau système est obligatoire depuis le 1er décembre 2010 pour les substances. Pour les mélanges, il le sera à partir du 1er juin 2015. En 2015, le système européen préexistant sera abrogé. Les dispositions de ce règlement ne s'appliquent pas au transport des produits chimiques, pour lequel un autre règlement existe.

Les recommandation des Nations Unies sur le transport des marchandises dangereuses (TDG : *Transport of Dangerous Goods*[13]) a été créé dans le but d'harmoniser des différentes réglementations nationales afin de permettre le commerce et le transport des marchandises dangereuses au niveaux international. La réglementation française du transport de marchandises dangereuses est regroupée en un seul texte : l'arrêté TMD[14], qui concerne le transport routier, ferroviaire et fluvial. La réglementation ADR[15] (Transport de Marchandises Dangereuses par Route) comporte 9 classes qui sont elles-mêmes découpées en sous-classes. Chaque matière est ainsi rattachée à une ou plusieurs classes selon les caractéristiques de danger qu'elle présente. Le Manuel d'épreuves et critères des Nations Unies[16], utilisé en relation avec le TDG[13], décrit les méthodes d'épreuves et procédures jugées les plus utiles pour fournir l'information nécessaire au classement correct des matières.

Tableau 2: Classement des matières dangereuses

PICTOGRAMMES	CLASSE	DESIGNATION
	1	Matières et objets explosibles
	2	Gaz
	3	Liquides inflammables
	4.1	Matières solides inflammables, matières autoréactives et matières explosibles désensibilisées solides
	4.2	Matières sujettes à inflammation spontanée
	4.3	Matières qui, au contact de l'eau dégagent des gaz inflammables
	5.1	Matières comburantes
	5.2	Peroxydes organiques
	6.1	Matières toxiques
	6.2	Matières infectieuses
	7	Matières radioactives
	8	Matières corrosives
	9	Matières et objets dangereux divers

4. Méthodes alternatives

En raison du grand nombre de substances et propriétés concernées par la réglementation REACH, la caractérisation expérimentale de l'ensemble des propriétés est contraignante pour des raisons de temps, de coûts, d'éthique (essais sur animaux) et de faisabilité au niveau recherche et développement. Ainsi, le développement de méthodes prédictives, alternatives à l'expérimentation, est recommandé. En particulier l'utilisation des modèles QSPR est indiquée dans l'annexe XI de REACH.

> *« 1.3. Relation qualitative ou quantitative structure-activité (RSA)*
> *Les résultats obtenus à l'aide des modèles valides de la relation qualitative ou quantitative structure-activité (R(Q)SA) peuvent indiquer la présence ou l'absence d'une certaine propriété dangereuse. Les résultats de la R(Q)SA peuvent être utilisés au lieu de l'essai lorsque les conditions suivantes sont réunies :*
> - *les résultats sont issus d'un modèle R(Q)SA dont la validité scientifique a été établie,*
> - *la substance relève du domaine d'applicabilité du modèle R(Q)SA,*
> - *les résultats conviennent pour la classification et l'étiquetage, et/ou pour l'évaluation des risques, et*
> - *une description suffisante et fiable de la méthode appliquée est fournie. »*

Des principes ont même été mis en place par l'OCDE[17] en 2009 afin de valider scientifiquement et réglementairement les modèles QSAR/QSPR. Cela dans le but d'augmenter la confiance en ces modèles mathématiques prédictifs et de favoriser ainsi leur utilisation par les industriels.

Les 5 principes OCDE[18] sont les suivants :

1) Une propriété ciblée définie (avec un protocole expérimental identifié) ;
2) Un algorithme sans équivoque ;
3) Un domaine d'applicabilité définie ;
4) Des mesures appropriées de la qualité d'ajustement, de robustesse et du pouvoir prédictif ;
5) Si possible, une interprétation des mécanismes sous-jacents.

L'application de ce règlement concerne à la fois la production, la mise sur le marché et l'utilisation des substances elles-mêmes mais aussi des préparations dans lesquelles on les retrouve. Quelques exceptions ne sont pas touchées par ce règlement et les directives déjà en vigueur restent en application: substances radioactives, intermédiaires non isolés, polymères, déchets ou encore transport des marchandises dangereuses.

C'est dans ce contexte que le projet PREDIMOL[1] (PREDIction des propriétés physico-chimiques des produits par modélisation MOLéculaire), financé par l'Agence Nationale de la Recherche, a été conçu.

II. PREDIMOL

Ce projet financé par l'Agence Nationale de la Recherche (dans le cadre de l'appel à projets 2010 « Chimie Durable – Industries - Innovation ») et labellisé par le pôle de compétitivité Axelera a démarré le 15 novembre 2010 pour une durée de trois ans, piloté par l'INERIS associé à plusieurs partenaires publics et privés : IFP Energies Nouvelles, Chimie ParisTech, le Laboratoire de Chimie Physique de Paris XI, Materials Design et Arkema.

L'objectif principal du projet PREDIMOL est de montrer que la modélisation moléculaire peut être une alternative crédible à l'expérimentation pour obtenir des données physico-chimiques (annexes VII et IX illustrées en Tableau 3) manquantes pour les besoins de REACH[3]. Plus précisément, il vise à développer des méthodes et modèles permettant d'estimer de manière précise, quantitative et rapide, les propriétés physico-chimiques nécessaires à l'enregistrement des substances dans REACH à partir de Relations Quantitatives Structure-Propriétés (QSPR) et des méthodes de simulation moléculaire (Dynamique Moléculaire et Monte Carlo). Ce projet vise également le développement d'outils automatisés et de calculs à haut débit pour l'acquisition de données en grande quantité.

1)	État de la substance à 20°C et 101,3 kPa
2)	Point de fusion/congélation
3)	Point d'ébullition
4)	Densité relative
5)	Pression de vapeur
6)	Tension superficielle
7)	Hydrosolubilité
8)	Coefficient de partage n-octanol/eau
9)	Point d'éclair
10)	Inflammabilité
11)	Propriétés explosives
12)	Température d'auto-inflammation
13)	Propriétés comburantes
14)	Granulométrie
15)	Stabilité dans les solvants organiques et identité des produits de dégradation
16)	Constante de dissociation
17)	Viscosité

Tableau 3 : Propriétés physico-chimiques standards exigées dans les annexes VII (propriétés 1 à 14) et IX (propriétés 15 à 17) de REACH

L'organisation du projet PREDIMOL a été structurée en plusieurs parties, comme illustré Figure 2. La première partie, découpée en deux tâches, est la définition des familles de molécules et des propriétés physico-chimiques d'étude, suivie d'un recensement des bases de données expérimentales. Par la suite, une liste de molécules à étudier a été définie afin de consolider les bases de données identifiées avec l'acquisition de données expérimentales. Celles-ci ont été

mesurées par les partenaires dans des conditions expérimentales homogènes, nécessaires pour le développement de modèles. La deuxième partie est le développement de nouvelles méthodes d'études et le développement de modèles prédictifs. Différentes méthodes ont été abordées : méthode QSPR couplée avec la DFT, Monte Carlo, dynamique moléculaire... La validation scientifique et réglementaire constitue une partie à part entière du projet. Les démarches pour l'acceptation des modèles peuvent être assez longues. Tout cela dans le but de pouvoir calculer les propriétés physico-chimiques des substances pour REACH. Les calculs à hauts débits ainsi que la comparaison des différentes méthodes de modélisation utilisées constituent la dernière étape avant l'automatisation.

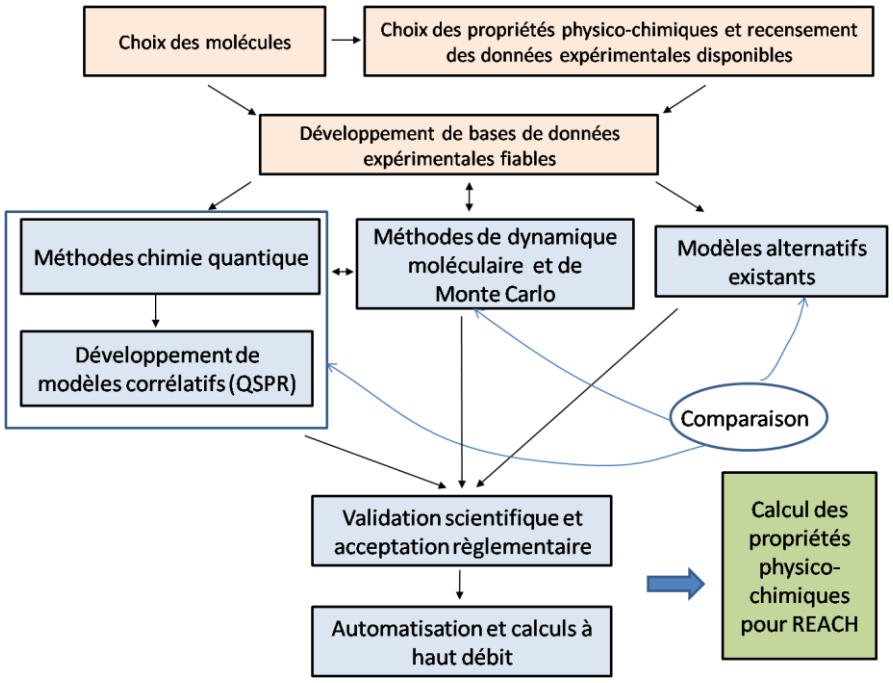


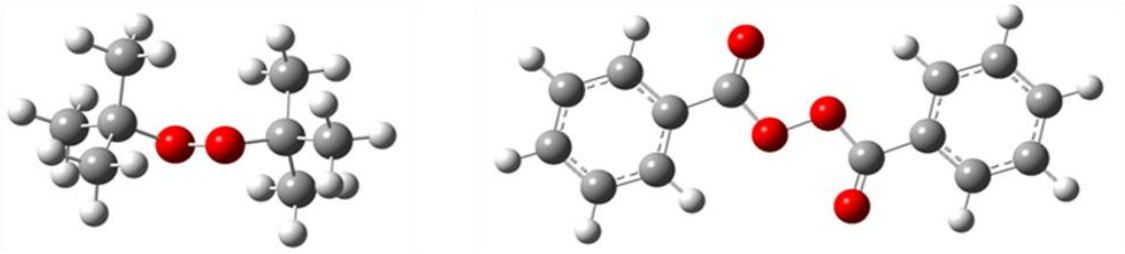
Figure 2 : Organigramme du projet PREDIMOL

L'apport de cette thèse dans ce projet réside dans le développement de modèles QSPR pour prédire les propriétés dangereuses (telle que l'explosibilité, la chaleur et la température de décomposition…) des substances chimiques en tenant compte des mécanismes réactionnels impliqués. Les modèles seront développés dans le but de les diffuser et permettre leur utilisation pour l'enregistrement des substances chimiques. Leur validation scientifique et réglementaire constitue par conséquent un objectif important. La famille de substances chimiques choisie pour cette étude est celle des peroxydes organiques.

III. LES PEROXYDES ORGANIQUES

1. Définitions et propriétés

Les peroxydes organiques[19–22] sont des composés qui contiennent du carbone et possèdent au moins deux atomes d'oxygène reliés ensemble (groupe peroxy) dans la molécule (Figure 3). Ils sont caractérisés par cette liaison très réactive qui rend ces composés instables.



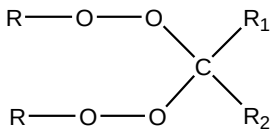
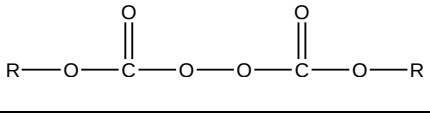
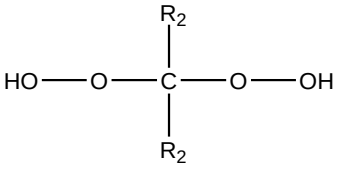
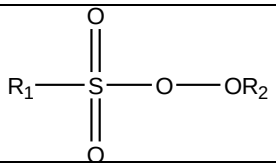
Peroxyde de di-ter-butyle — Peroxyde de benzoyle

Figure 3 : Représentation de molécules de peroxydes organiques

Selon la fiche pratique et sécurité INRS[23], les peroxydes sont généralement sensibles aux sollicitations mécaniques (tels que les chocs et les frottements) et peuvent réagir violemment avec des substances variées[24] (acides forts, bases fortes, amines, alcools, certains métaux et sels métalliques, produits facilement oxydables…). Pour cette raison, ils sont généralement commercialisés en faible concentration, variant en fonction du peroxyde, soit en mélange avec un solvant ou un produit liquide à point d'ébullition élevé (flegmatisant) soit dilué avec une certaine quantité d'eau pour atténuer la sensibilité au choc. Les peroxydes peuvent se classer en différents types de familles comme proposé par INRS[25] dans le Tableau 4.

Tableau 4 : Différents types de peroxydes organiques selon INRS[20]

Types de peroxydes	Formule générale	Exemple connus
Peroxydes di dialkyles	$R_1-O-O-R_2$	Peroxyde de di-*t*-butyle, Peroxyde de dicumyle, Peroxyde de *t*-butyle de de cumyle…
Peroxydes de diacyles	$R-C(=O)-O-O-C(=O)-R$	Peroxyde de dibenzoyle, Peroxyde de dilauroyle, Peroxyde de di-(2,4-dichloro-benzoyle)…
Hydroperoxydes	$R-O-OH$	Hydroperoxyde de *t*-butyle, Hydroperoxyde d'α-cumyle, Hydroperoxyde de 1-phényl-éthyle…
Peroxiacides	$R-C(=O)-OOH$	Acide peroxyacétique, Acide *p*-nitro-peroxybenzoïque…
Peroxyesters	$R_1-C(=O)-O-O-R_2$	Peroxyacétale de *t*-butyle, Peroxypivalate de *t*-butyle, Peroxybenzoate de *t*-butyle…

Types de peroxydes	Formule générale	Exemple connus
Peroxycétals		1,1-di-(*t*-butylperoxy)-3,3,5-triméthylcyclohexane, 1,1-di--(*t*-butylperoxy) cyclohexane, 2,2-di-(cumylperoxy)propane…
Peroxydicarbonates		Peroxycarbonate de diisopropyle Peroxycarbonate de di-*sec*-butyle, Peroxycarbonate de di-(2-butoxyéthyle)…
Peroxydes de cétones		Peroxyde de méthyléthylcétone; Peroxyde d'acétylacétone; Peroxyde de cyclohexanone…
Peroxydes de sulfonyles		Peroxyde d'acétylcyclohexane-sulfonyle…
Peroxydes de silyles	$(R_1OO)_nSi(R_2)_{4-n}$	Vinyltri-(*t*-butylperoxy)silane, Cumylperoxytriméthulsilane…

Selon les règlements CLP[11] (page 179) et ADR[15] (page 187), les peroxydes organiques sont définis comme des « substances organiques liquides ou solides qui contiennent la structure bivalente -O-O- et qui peuvent être considérées comme des dérivés du peroxyde d'hydrogène dans lesquels un ou les deux atomes d'hydrogène ont été remplacés par des radicaux organiques ». Par peroxydes organiques, on entend aussi les mélanges (préparations) contenant au moins un peroxyde organique. Ils sont thermiquement instables et peuvent subir une décomposition exothermique auto-accélérée à température normale ou élevée. Leur décomposition peut s'amorcer sous l'effet de la chaleur, du frottement, du choc, ou du contact avec des contaminants (acides, composés de métaux lourds, amines, etc.). Elle peut entraîner un dégagement de vapeurs ou de gaz inflammables ou nocifs. La vitesse de décomposition croît avec la température et dépend de la nature du peroxyde. Pour certains peroxydes organiques, une régulation de température est obligatoire pendant le transport. Certains peuvent se décomposer en produisant une explosion, surtout sous confinement. Cette caractéristique peut être modifiée par dilution ou l'emploi d'emballages appropriés.

Toutes ces définitions s'accordent pour dire que les peroxydes organiques sont caractérisés par une liaison chimique O-O qui rend ces composés instables. Ils sont capables de se décomposer d'une manière exothermique auto-accélérée. Les peroxydes organiques peuvent exploser ou s'enflammer car ils sont sensibles à la chaleur et aux stimuli tels que les chocs, frottements mais aussi aux contaminations (par d'autres substances chimiques incompatibles). Ils peuvent également être toxiques et corrosifs. La décomposition des peroxydes[26–30] peut s'amorcer facilement à température

ambiante. Elle se fait par la rupture homolytique de la liaison peroxyde O-O qui est de faible énergie (20 à 50 kcal/mol contre 83 kcal/mol pour une liaison carbone-carbone[28,31]). La liaison peroxyde présente deux propriétés principales : un groupe fonctionnel très réactif en raison de son degré d'oxydation et une décomposition qui produit des radicaux.

Les peroxydes sont utilisés dans l'industrie comme amorceurs radicalaires dans les industries du plastique, de la peinture (pour réaliser des polymérisations), du caoutchouc (pour réaliser des vulcanisations) ou pour réticuler (durcir) des résines mais aussi l'oxydation, la décoloration ou encore comme biocides… Les peroxydes peuvent se former par un phénomène d'auto-oxydation (dit peroxydation), en particulier lorsqu'ils sont conservés longtemps. Ce sont des intermédiaires d'oxydation par l'air de plusieurs produits. La formation de peroxydes dans certains solvants tels que le di-éthyle ether peut mener à des résidus hautement explosifs[32]. Pour prévenir la formation de peroxydes indésirables des antioxydants sont ajoutés.

Les peroxydes présentent trois types de danger : l'explosion, le feu et la toxicité. En effet, leur décomposition peut entraîner un dégagement de vapeurs ou de gaz inflammables ou nocifs, en plus de dégager une forte chaleur. Dans le cas d'une décomposition vive (généralement quand le peroxyde est sous confinement) une explosion peut parfois avoir lieu. Les risques peuvent être diminués par l'ajout de solvants pour diluer les peroxydes et/ou des flegmatisants ainsi que par l'emploi d'emballages appropriés. En conclusion, ces composés peuvent être dangereux.

2. Caractérisation

Un peroxyde est caractérisé notamment par sa température de demi-vie T½ (donnée pour des demi-vies t½ de 10 h et 1 h), son taux massique en oxygène actif, sa température de décomposition auto-accélérée TDAA, sa température maximale de stockage et sa zone de températures d'utilisation ; certains paramètres permettent un classement selon la stabilité.

La teneur en oxygène actif (en %) d'une préparation de peroxyde organique est donnée par la formule :

$$16\sum n_i c_i / m_i$$

Avec n_i le nombre de groupes peroxy par molécule du peroxyde organique i ;

c_i la concentration (% en masse) du peroxyde organique i ;

et m_i la masse moléculaire du peroxyde organique i.

La TDAA[33], mesurée pour un conditionnement donné, dépend du type d'emballage utilisé et de son volume (rapport masse sur surface d'échange). C'est la température la plus basse à laquelle un

peroxyde organique dans son emballage peut subir une décomposition auto-accélérée. Cette température permet de définir la température de stockage et la zone de températures d'utilisation.

3. Classification des peroxydes organiques

L'ADR[15] consacre un classement particulier à tous les peroxydes organiques en leur dédiant une classe à part : la classe 5.2. Tout peroxyde organique est censé être dans cette classe sauf si la préparation de peroxyde organique :

a) ne contient pas plus de 1% d'oxygène actif pour 1% au maximum de peroxyde d'hydrogène ;
b) ne contient pas plus de 0,5% d'oxygène actif pour plus de 1% mais 7% au maximum de peroxydes d'hydrogène.

A l'intérieur de cette classe, les peroxydes sont répartis en 7 types (voir Tableau 5) en fonction des résultats d'essais des épreuves des séries A à H du Manuel d'épreuves et de critères des Nations Unies qui détermine leur degré de danger.

Tableau 5: Principe de classification des peroxydes organiques parmi les 7 types selon la réglementation CLP[11]

Type du PO	Description des dangers
A	Peroxyde organique qui, tel qu'emballé, peut détoner ou déflagrer rapidement
B	Peroxyde organique ayant des propriétés explosives qui, tel qu'emballé, ne peut pas détoner ni déflagrer rapidement, mais peut exploser sous l'effet de la chaleur dans cet emballage
C	Peroxyde organique ayant des propriétés explosives qui, tel qu'emballé, ne peut pas détoner, déflagrer rapidement ni exploser sous l'effet de la chaleur
D	Peroxyde organique qui, lors d'épreuves de laboratoire: i) Détone partiellement, mais ne déflagre pas rapidement et ne réagit pas violemment au chauffage sous confinement ; ii) Ne détone pas, déflagre lentement mais ne réagit pas violemment au chauffage sous confinement ; iii) Ne détone pas et ne déflagre pas et réagit modérément au chauffage sous confinement
E	Peroxyde organique qui, lors d'épreuves de laboratoire, ne détone pas, ne déflagre pas et n'a qu'une réaction faible ou nulle au chauffage sous confinement
F	Peroxyde organique qui, lors d'épreuves de laboratoire, ne détone pas à l'état cavité, ne déflagre pas, n'a qu'une réaction faible ou nulle au chauffage sous confinement et n'a qu'une puissance explosive faible ou nulle Peroxyde organique qui n'est pas thermiquement stable (c'est-à-dire que la température de décomposition auto-accélérée ou point de décomposition exothermique (TDAA) soit de 60 °C ou plus pour un colis de 50 kg) ou si le diluant utilisé comme flegmatisant a un point d'ébullition inférieur à 150 °C

Type du PO	Description des dangers
G	Peroxyde organique qui, lors d'épreuves de laboratoire, ne détone pas à l'état cavité et ne déflagre pas, ne réagit pas au chauffage sous confinement et a une puissance explosive nulle, à condition qu'il soit thermiquement stable (c'est-à-dire que la température de décomposition auto-accélérée ou point de décomposition exothermique (TDAA) soit de 60 °C ou plus pour un colis de 50 kg) et, pour les mélanges liquides, que le diluant utilisé comme flegmatisant ait un point d'ébullition d'au moins 150 °C

Le règlement CLP[11] considère qu'un peroxyde organique possède des propriétés explosives si, lors d'essais de laboratoire, la préparation se révèle susceptible de détoner, de déflagrer brusquement ou de réagir violemment à un chauffage sous confinement. Le Tableau 6 indique l'étiquetage associé à chaque type de peroxyde.

Tableau 6 : Etiquetage des peroxydes organiques[34]

Classification	Etiquetage
Peroxyde organique Type A H240 : peut exploser sous l'effet de la chaleur	 Danger H240
Peroxyde organique Type B H241 : peut s'enflammer ou exploser sous l'effet de la chaleur	 Danger H241
Peroxyde organique Types C et D H242 : peut s'enflammer sous l'effet de la chaleur	 Danger H242
Peroxyde organique Types E et F H242 : peut s'enflammer sous l'effet de la chaleur	 Attention H242
Peroxyde organique Type G	-

Selon la nomenclature ICPE[35] (Installations Classées pour la Protection de l'Environnement), les peroxydes organiques et les préparations en contenant sont répartis en quatre groupes de risques décroissants de Gr 1 à Gr 4 :

Gr 1 : Produits présentant un risque de décomposition violente ou de combustion très rapide ;

Gr 2 : Produits présentant un risque de combustion rapide ;

Gr 3 : Produits présentant un risque de combustion moyenne similaire à celle du bois ou des solvants organiques ;

Gr 4 : Produits présentant un risque de combustion nulle ou lente.

En pratique, le classement des peroxydes organiques s'appuie sur celui de l'ONU pour le transport des matières dangereuses (groupes B à G) et sur la vitesse de combustion mesurée (« quantité de substances brulée par minute pour un lot de 10000 kg dévoré par le feu »[36]).

Tableau 7 : Classement générique des peroxydes organiques entre les différents groupes de risque

Type de danger selon l'arrêté ADR en vigueur	Groupe de risque			
A	1	1	1	1
B	1	1	1	1
C	2	2	2	1
D	3	3	2	
E	4	3	2	
F	4	3	3	
G	4	3		
Vitesse de combustion	kg/min (Test grande échelle) 1	10	60	300
	kg/m².min (Test laboratoire)		0,9	9

4. Accidentologie liées aux peroxydes organiques

Les peroxydes organiques sont des produits dangereux qui peuvent produire des nuages toxiques, des explosions ou déclencher des feux. Deux accidents lors du transport de peroxydes organiques ont été identifiés[37] : le premier aux USA et le second au Royaume-Uni. Entre 1978 et 1996, 10 explosions à Taïwan[38] ont été causés par les peroxydes, parmi eux, 8 sont dus à la décomposition thermique, 1 à un feu extérieur et le dernier à une incompatibilité. Quatre de ces 10 accidents ont été provoqués par le peroxyde de la méthyl-éthyl-cétone (MEKPO) qui est un durcisseur pour résines polyester insaturées.

Tableau 8: Explosion causée par un peroxyde à Taiwan entre 1978 et 1996)[38]

Date	Substances	Nombre de blessés	Nombre de mort	Cause
14/07/1978	MEKPO	49	33	Décomposition
21/04/1981	Hydroperoxyde de cumène	3	1	Décomposition
18/02/1984	MEKPO	55	5	Décomposition
02/02/1986	Hydroperoxyde de cumène	0	0	Décomposition
05/09/1987	H_2O_2	20	0	Incompatibilité
25/07/1988	t-butyl hydroperoxide	19	0	Décomposition
14/03/1989	Organic Peroxide	0	0	Décomposition
04/08/1989	Organic Peroxide	0	0	Décomposition
01/09/1989	MEKPO	5	7	Décomposition
10/07/1996	MEKPO	47	10	Feu

Un grand nombre d'accidents liés au MEKPO ont été recensés au Japon, Chine, Taïwan[38–40]. L'accident le plus grave a eu lieu à Tokyo au Japon le 14 juillet 1964 où 3600 kg de MEKPO ont explosé, tuant 19 personnes et blessant 114 autres personnes. A Taiwan, l'accident le plus sérieux a eu lieu en 1996 à Yung-Hsin (réaction d'oxydation non contrôlée). Les articles de Chang[40] et Tseng[39]

illustrent l'incompatibilité des peroxydes avec certaines autres substances, appelées contaminant, avec le MEKPO qui est moins stable en présence de H_2SO_4, NaOH ou HCl par exemple.

D'autres accidents, liés au cumène hydroperoxyde (CHP), ont été identifiés par Kletz[41] et Chen[42]. Le CHP est utilisé comme initiateur pour la polymérisation de l'acrylonitrile butadiène styrène (ABS) ou encore pour la production de phénol et d'acétone. Fishwick[43] rapporte un accident avec le di-chloro benzoyle peroxyde(DCLBP) survenu lors de la fabrication du 2,4-di-chloro benzoyle chloride. La base de données ARIA[44] recense des incidents ou accidents qui ont ou auraient pu porter atteinte à la santé ou la sécurité publiques, l'agriculture, la nature et l'environnement en France notamment mais aussi dans le reste du monde.

5. Sécurité

La dangerosité de ces peroxydes organiques n'est donc plus à démontrer et, pour réduire au minimum les accidents, des mesures de sécurité ont été mises en place aux différents niveaux du parcours des substances : transport, stockage et utilisation.

a) Transport

Il a été vu dans la partie « Accidentologie liées aux peroxydes organiques » que les accidents peuvent se produire lors du transport des peroxydes organiques. Pour assurer la sécurité pendant le transport des peroxydes organiques, ils sont désensibilisés, souvent en ajoutant des matières organiques liquides ou solides, des matières inorganiques solides ou de l'eau. La dilution doit être telle qu'en cas de fuite, le peroxyde organique ne puisse pas se concentrer dans une mesure dangereuse. Il existe plusieurs types de diluants défini de la manière suivante dans l'ADR :

- Les diluants de type A sont des liquides organiques qui sont compatibles avec le peroxyde organique et qui ont un point d'ébullition d'au moins 150 °C. Les diluants de type A peuvent être utilisés pour désensibiliser tous les peroxydes organiques;
- Les diluants de type B sont des liquides organiques qui sont compatibles avec le peroxyde organique et qui ont un point d'ébullition inférieur à 150 °C mais au moins égal à 60 °C et un point d'éclair d'au moins 5 °C. Les diluants du type B peuvent être utilisés pour désensibiliser tout peroxyde organique à condition que le point d'ébullition du liquide soit d'au moins 60 °C plus élevé que la TDAA dans un colis de 50 kg.

De plus, pour certains peroxydes organiques, une régulation de température est obligatoire pendant le transport. En particulier, selon l'ADR, les peroxydes suivant sont soumis à la régulation de température pendant le transport :

- Type B et C ayant une TDAA ≤ 50 °C ;

- Type D manifestant un effet moyen lors de chauffage sous confinement et ayant une TDAA ≤ 50°C, ou manifestant un faible ou aucun effet lors de chauffage sous confinement et ayant une TDAA ≤ 45 °C ;
- Type E et F ayant une TDAA ≤ 45 °C.

En revanche, les peroxydes de type A pouvant détoner ou déflagrer à grande vitesse dans leurs emballages sont interdits au transport. Quant aux peroxydes de type G, ils ne sont pas assujettis aux prescriptions de la classe 5.2.

La température de transport doit donc être inférieure à la température maximale de contrôle (T1) mais aussi, dans certains cas, choisie de manière à éviter la turbidité, la séparation de phase, le dépôt de cristaux ou la solidification. La température de maximale de contrôle est déterminée à partir de la TDAA pour des peroxydes emballés et stockés en emballages (voir Tableau 9).

Afin de réduire les risques, la taille maximale de l'emballage d'un peroxyde organique est clairement définie ainsi que le type de l'emballage (notamment dans l'ADR). Elle est déterminée à l'issue d'une série de tests définissant le caractère détonant, déflagrant d'un produit, son comportement lorsqu'il est chauffé sous confinement, ainsi que sa puissance explosive. Les tests à réaliser sont définis à l'aide du diagramme de décision pour le classement des matières réactives et des peroxydes organiques disponible en annexe I et dans le Manuel d'épreuves et critères[16].

b) Stockage

Pour prévenir les accidents et réduire leurs effets lors du stockage, de bonnes pratiques existent. Le moyen de prévention le plus important est le contrôle des températures. Des mesures de température de l'espace de stockage et des peroxydes doivent être effectuées afin de vérifier que la température maximale de régulation du stockage (T1) n'est pas dépassée. Si la température atteint ou dépasse le seuil de température d'urgence (T2, voir Tableau 9), il faut mettre en place des mesures compensatoires immédiates : inondation par l'eau pour refroidir mais aussi l'extinction du feu en cas d'incendie, mise en place d'un bassin de captable pour éviter la diffusion du feu et la pollution des sols… D'autres recommandations ont été faites, comme éviter les contaminations, l'absence de source de chaleur ou d'ignition des vapeurs (tels que les installations électriques), période de stockage limitée, ventilation et distance de sécurité.

Tableau 9: Définition des températures de contrôle et d'urgence en fonction de la TDAA

TDAA	T1	T2
≤ 20°C	TDAA – 20°C	TDAA – 10°C
20°C < TDAA ≤ 35°C	TDAA – 15°C	TDAA – 10°C
> 35°C	TDAA – 10°C	TDAA – 5°C

A l'intérieur d'une cellule de stockage, une bonne circulation de l'air doit être assurée entre les palettes de peroxydes : un espace minimum de 15 cm entre les palettes et la paroi du stockage doit être garanti par un dispositif approprié. De même, un espace suffisant d'au moins 10 cm doit être laissé entre les palettes. La ventilation doit réduire la concentration en pression à une valeur maximale de 20% de la limite inférieure d'inflammabilité des produits. Pour cela, la mise en place de conduits permettant la redirection des vapeurs produites lors de la décomposition est une bonne solution.

c) Utilisation

La première recommandation lors de l'utilisation est de porter une protection personnelle : gants, lunettes de sécurité, blouse… Des fiches de sécurité ont été rédigées pour chaque substance afin de prévenir l'utilisateur sur les dangers des peroxydes qu'il utilise. L'annexe 4 du règlement SGH est un guide sur l'élaboration des fiches de données de sécurité (FDS). La FDS doit indiquer quels dangers présente une substance ou un mélange et comment les stocker, les manipuler ou les éliminer dans des conditions satisfaisantes de sécurité. Cette fiche doit être compréhensible facilement pour toute personne y ayant accès. Les informations devront être rédigées de manière cohérente et exhaustive : pas d'abréviation, indiqué clairement si l'information demandée n'est pas pertinente…

Les informations devront figurer sur la FDS sous 16 rubriques, dans l'ordre établi ci-dessous :

1. Identification
2. Identification du ou des dangers
3. Composition/information sur les composants
4. Premiers soins
5. Mesures à prendre en cas d'incendie
6. Mesures à prendre en cas de déversement accidentel
7. Manutention et stockage
8. Contrôles de l'exposition/protection individuelle
9. Propriétés physiques et chimiques
10. Stabilité et réactivité
11. Données toxicologiques
12. Données écologiques
13. Données sur l'élimination
14. Informations relatives au transport
15. Informations sur la réglementation
16. Autres informations.

Dès lors que le peroxyde est sorti de son emballage de transport, les risques liés au confinement et à l'auto-échauffement doivent être réévalués. Le produit n'a plus la valeur de TDAA caractéristique de cet emballage. D'autres propriétés peuvent être très différentes en fonction des conditions d'utilisation. Ainsi le confinement accroît la violence des décompositions. Aussi dans le cas des réacteurs métalliques, la mise en place de soupapes, de disques de rupture est à prévoir pour éviter la rupture mécanique en cas de surpression due à une décomposition du peroxyde.

Les modèles QSPR sont une approche alternative pour la prédiction de propriétés. Il existe plusieurs modèles QSPR pour la prédiction de propriétés thermiques pour différents type de composés chimiques : les nitroaromatiques[45–48], nitramines [49–51], liquides ioniques [52], polymères [52,53]... Cependant, à notre connaissance, seulement un modèle QSPR [54] existe pour la prédiction des propriétés dangereuses des peroxydes organiques. Des modèles ont donc été développés au cours de ce travail de thèse et sont présentés dans ce manuscrit.

IV. PLAN DU MANUSCRIT DE THÈSE

Le manuscrit a été organisé de la manière suivante :

Après ce chapitre d'introduction sur le contexte et les objectifs, un deuxième chapitre présentera les bases de la chimie théorique ainsi qu'un programme permettant de réduire le nombre de conformations développé au cours de cette thèse. Dans un troisième chapitre, les principes et les méthodes de développement de modèle QSPR seront introduits. Les étapes du développent des modèles QSPR, depuis de la préparation de la base de données jusqu'à la validation du modèle en passant par la mise en place du modèle, seront expliquées. Dans les quatrième et cinquième chapitres, des modèles QSPR obtenus au cours de cette thèse seront présentés : tout d'abord les modèles ayant été obtenus avec les données de la Datatop 2005, puis le calcul des énergies de dissociation de la liaison peroxy ainsi que les différentes corrélations. Puis seront proposés les modèles développés pour différentes propriétés physico-chimiques à partir de la base de données de 38 peroxydes organiques, construite dans le cadre du projet PREDIMOL.

Cette thèse a pour objectif le développement de modèles QSPR validés suivant les principes OCDE, liés à la prédiction de propriétés physico-chimiques demandées pour permettre l'enregistrement des peroxydes organiques dans REACH. Les propriétés dangereuses sont particulièrement visées comme la chaleur de décomposition, la température de début de décomposition…

V. RÉFÉRENCES

(1) PREDIMOL www.ineris.fr/predimol/ (accessed Sep 4, 2012).

(2) Université de Rennes 1 - Service commun de documentation Genèse de la chimie moderne http://www-scd.univ-rennes1.fr/themes/presentation-SCD/phototheque/expo_chimie/ (accessed Mar 1, 2013).

(3) *Règlement (CE) N° 1907/2006 Du Parlement Européen et Du Conseil Du 18 Décembre 2006 Concernant L'enregistrement, L'évaluation et L'autorisation Des Substances Chimiques, Ainsi Que Les Restrictions Applicables à Ces Substances (REACH), Instituant Une Agence Européenne Des Produits Chimiques, Modifiant La Directive 1999/45/CE et Abrogeant Le Règlement (CEE) N° 793/93 Du Conseil et Le Règlement (CE) N° 1488/94 de La Commission Ainsi Que La Directive 76/769/CEE Du Conseil et Les Directives 91/155/CEE, 93/67/CEE, 93/105/CE et 2000/21/CE de La Commission*; 2006.

(4) Antenne 2 Journal télévisé du 27 juillet 1976 (archive INA) http://www.ina.fr/video/CAB04013276/ja2-20h-emission-du-27-juillet-1976.fr.html.

(5) Directive Du Conseil N° 67/548/CEE Du 27 Juin 1967 Concernant Le Rapprochement Des Dispositions Législatives, Réglementaires et Administratives Relatives à La Classification, L'emballage et L'étiquetage Des Substances Dangereuses.

(6) Stratégie Pour La Future Politique Dans Le Domaine Des Substances Chimiques, Livre Blanc, COM(2001)88 Final, Commission Européenne.

(7) European Chemicals Agency (ECHA) http://echa.europa.eu/.

(8) Helpdesk REACH & CLP http://www.reach.lu/ (accessed Jan 23, 2013).

(9) UFCC La règlementation Reach au jour le jour http://www.reach.ufcc.fr/ (accessed Jan 23, 2013).

(10) Parlement Européen *Document de Base Législatif 2003/0257(COD)*; 2003.

(11) Règlement CLP. Règlement (CE) N° 1272/2008 Du Parlement Européen et Du Conseil Du 16 Décembre 2008 Relatif à La Classification, à L'étiquetage et L'emballage Des Substances et Des Mélanges, Modifiant et Abrogeant Les Directives 67/548/CEE et 1999/45/CE et Modifiant Le Règlement CE N° 1907/2006.

(12) ST/SG/AC.10/30/Rev.4 Système Général Harmonisé de Classification et D'étiquetage Des Produits Chimiques (SGH), Nations Unies, New York et Genève, 2011.

(13) UNECE Transport of Dangerous Goods http://www.unece.org/trans/danger/danger.html (accessed Jan 31, 2013).

(14) Arrêté Du 29 Mai 2009 Relatif Au Transport de Marchandises Dangereuses Par Voies Terrestres (dit "Arrêté TMD") NOR: DEVP1241087A Version Consolidée Au 01 Janvier 2013 **2013**.

(15) Accord Européen Relatif Au Transport International Des Marchandises Dangereuses Par Route, ECE/TRANS/225, ADR En Vigueur Le 1er Janvier 2013, Nations Unies, 2012.

(16) *Recommandations Relatives Au Transport Des Marchandises Dangereuses - Manuel D'épreuves et de Critères ST/SG/AC.10/11/Rev.5*; Nations Unies, 2010.

(17) Organisation de Coopération et de Développement Economiques (OCDE).

(18) Organisation de Coopération et de Développement Economiques (OCDE) *Principles for the Validation, for Regulatory Purposes, of (Quantitative) Structure-Activity Relationship Models*; Paris, 2009.

(19) Swern, D.; Swern, D. *Organic Peroxides. Volume 1*; Wiley-Interscience: New York, 1970.

(20) Lemarquand, J.; Triolet, J. Les Peroxydes et Leur Utilisation. *Cahiers de notes documentaires – Hygiène et sécurité du travail* **2002**, *186*, 5–14.

(21) Centre canadien d'hygiène et de sécurité au travail Les peroxydes organiques et leurs dangers http://www.cchst.ca/oshanswers/chemicals/organic/organic_peroxide.html (accessed Mar 7, 2013).

(22) Médard, L.; Chovin, P. *Les explosifs occasionnels*; I.P.E., industries, productions, environnement; Technique et documentation: Paris, France, 1979.

(23) INRS Fiche Pratique et Sécurité ED 41 - Peroxydes. Risques à L'utilisation et Mesures de Sécurité **2005**.

(24) Arkema Organic peroxide safety: Video of contamination http://www.arkema-inc.com/media/orgper/contam_broadband.wmv (accessed Mar 18, 2013).

(25) Institut National de Recherche et de Sécurité (INRS) www.inrs.fr.

(26) Benassi, R.; Taddei, F. Homolytic Bond-dissociation in Peroxides, Peroxyacids, Peroxyesters and Related Radicals: Ab-initio MO Calculations. *Tetrahedron* **1994**, *50*, 4795–4810.

(27) Benassi, R.; Folli, U.; Sbardellati, S.; Taddei, F. Conformational Properties and Homolytic Bond Cleavage of Organic Peroxides. I: An Empirical Approach Based Upon Molecular Mechanics and Ab Initio Calculations. *J. Comput. Chem.* **1993**, *14*, 379–391.

(28) Duh, Y.-S.; Hui wu, X.; Kao, C.-S. Hazard Ratings for Organic Peroxides. *Proc. Safety Prog.* **2008**, *27*, 89–99.

(29) Swain, C. G.; Stockmayer, W. H.; Clarke, J. T. Effect of Structure on the Rate of Spontaneous Thermal Decomposition of Substituted Benzoyl Peroxides. *Journal of the American Chemical Society* **1950**, *72*, 5426–5434.

(30) Litinskii, A. O.; Shreibert, A. I.; Balyavichus, L.-M. Z.; Bolotin, A. B. Electronic Structure, Stability, and Reactivity of Alkyl Peroxides. *Theor Exp Chem* **1974**, *7*, 673–675.

(31) Zumdahl, S. S. *Chimie générale*; De Boeck Université; Les Ed. CEC: Paris; [S.l], 1998.

(32) Di Tommaso, S.; Rotureau, P.; Crescenzi, O.; Adamo, C. Oxidation Mechanism of Diethyl Ether: a Complex Process for a Simple Molecule. *Phys. Chem. Chem. Phys.* **2011**, *13*, 14636–14645.

(33) Arkema Organic peroxide safety: Video of improper storage http://www.arkema-inc.com/media/orgper/handle_broadband.wmv (accessed Mar 18, 2013).

(34) INERIS Service national d'assistance réglementaire sur le réglement CLP http://clp-info.ineris.fr/ (accessed Jan 31, 2013).

(35) Arrêté Du 20/03/07 Relatif à La Définition et à La Classification Des Peroxydes Organiques Entre Les Différents Groupes de Risque Définis à La Rubrique 1210 de La Nomenclature Des Installations Classées.

(36) CPR 3E (Committee for the prevention of disasters caused by dangerous substances) *Storage of Organic Peroxides*; 2nde ed.; Sdu Ultgevers, 1997.

(37) Long, R. Two Explosion in the Transport of Organic Peroxides. *Loss Prevention Bulletin* **2000**, *153*, 17.

(38) Ho, T.-C.; Duh, Y.-S.; Chen, J. R. Case Studies of Incidents in Runaway Reactions and Emergency Relief. *Process Safety Progress* **1998**, *17*, 259–262.

(39) Tseng, J.-M.; Chang, Y.-Y.; Su, T.-S.; Shu, C.-M. Study of Thermal Decomposition of Methyl Ethyl Ketone Peroxide Using DSC and Simulation. *Journal of Hazardous Materials* **2007**, *142*, 765–770.

(40) Chang, R. H.; Tseng, J. M.; Jehng, J. M.; Shu, C. M.; Hou, H. Y. Thermokinetic Model Simulations for Methyl Ethyl Ketone Peroxide Contaminated with H2SO4 OR NaOH by DSC and VSP2. *J Therm Anal Calorim* **2006**, *83*, 57–62.

(41) Kletz, T. A. Fires and Explosions of Hydrocarbon Oxidation Plants. *Plant/Operations Progress* **1988**, *7*, 226–230.

(42) Chen, K.-Y.; Wu, S.-H.; Wang, Y.-W.; Shu, C.-M. Runaway Reaction and Thermal Hazards Simulation of Cumene Hydroperoxide by DSC. *Journal of Loss Prevention in the Process Industries* **2008**, *21*, 101–109.

(43) Fishwick, T. An Uncontrolled Chemical Decomposition. *Loss Prevention Bulletin* **2004**, *177*, 8.

(44) Bureau d'Analyse des risques et Pollutions Industriels (BARPI) ARIA : Analyse, Recherche et Information sur les Accidents http://www.aria.developpement-durable.gouv.fr/ (accessed Mar 8, 2013).

(45) Fayet, G.; Rotureau, P.; Joubert, L.; Adamo, C. On the Prediction of Thermal Stability of Nitroaromatic Compounds Using Quantum Chemical Calculations. *Journal of Hazardous Materials* **2009**, *171*, 845–850.

(46) Fayet, G.; Rotureau, P.; Joubert, L.; Adamo, C. Development of a QSPR Model for Predicting Thermal Stabilities of Nitroaromatic Compounds Taking into Account Their Decomposition Mechanisms. *Journal of Molecular Modeling* **2010**, *17*, 2443–2453.

(47) Sang, P.; Zou, J.-W.; Xu, L.; Liu, Y.-H. QSPR of Thermal Stability of Nitroaromatic Explosives Using Theoretical Descriptors Derived from Electrostatic Potentials on the Molecular Surface. *chinese journal of structural chemistry* **2011**, *30*, 533–537.

(48) Sang, P.; Zou, J.; Xu, L.; Zhou, P. Linear and Nonlinear QSPR Models for Predicting Thermal Stabilities of Nitroaromatic Compounds. *Chemical research in chinese university* **2011**, *27*, 891–895.

(49) Atalar, T.; Zeman, S. A New View of Relationships of the N-N Bond Dissociation Energies of Cyclic Nitramines. Part I. Relationships with Heats of Fusion. *Journal of Energetic Materials* **2009**, *27*, 186–199.

(50) Keshavarz, M. H. Predicting Heats of Fusion of Nitramines. *Indian journal of engineering & materials sciences* **2007**, *14*, 386–390.

(51) Zeman, S. Some Predictions in the Field of the Physical Thermal Stability of Nitramines. *Thermochimica Acta* **1997**, *302*, 11–16.

(52) Gharagheizi, F.; Sattari, M.; Ilani-Kashkouli, P.; Mohammadi, A. H.; Ramjugernath, D.; Richon, D. Quantitative Structure-property Relationship for Thermal Decomposition Temperature of Ionic Liquids. *Chemical Engineering Science* **2012**, *84*, 557–563.

(53) Yu, X.; Xie, Z.; Yi, B.; Wang, X.; Liu, F. Prediction of the Thermal Decomposition Property of Polymers Using Quantum Chemical Descriptors. *European Polymer Journal* **2007**, *43*, 818–823.

(54) Lu, Y.; Ng, D.; Mannan, M. S. Prediction of the Reactivity Hazards for Organic Peroxides Using the QSPR Approach. *Industrial & Engineering Chemistry Research* **2011**, *50*, 1515–1522.

CHAPITRE 2 –DE L'ATOME À LA MOLÉCULE : RAPPELS DE THÉORIE

Dans cette partie, les différentes méthodes théoriques utilisées dans cette thèse pour l'étude des molécules, du niveau électronique au niveau moléculaire, seront présentées. Tout d'abord, les méthodes de la chimie quantique[1,2] seront rappelées, en particulier la théorie de la fonctionnelle de la densité (DFT) qui considère les molécules au niveau électronique (de l'électron aux orbitales moléculaires et au calcul de l'énergie). Puis, les bases de la mécanique moléculaire[3,4] seront décrites. Les molécules étant des objets flexibles qui existent sous différentes géométries, la dernière partie de ce chapitre sera consacrée à l'analyse conformationnelle avec la présentation d'un programme permettant de réduire le nombre de géométries à considérer par molécule.

I. De l'équation de Schrödinger à Hartree-Fock

Les méthodes quantiques peuvent être utilisées pour déterminer la structure électronique des molécules, l'énergie, la géométrie, mais aussi pour la recherche d'états de transition dans le cadre d'une étude sur la réactivité, ou encore pour l'étude des effets de solvant.

1. L'équation de Schrödinger

Les méthodes de chimie quantique reposent toutes sur le même postulat de départ : tout système peut être décrit par une fonction d'onde. Celle-ci est une fonction des coordonnées des noyaux mais aussi des électrons. Cette fonction est solution de l'équation de Schrödinger.

L'équation de Schrödinger non relativiste et indépendante du temps se présente sous la forme suivante :

$$\hat{H}\Psi = E\Psi \qquad \textbf{(2. 1)}$$

dans laquelle Ψ est la fonction d'onde décrivant le système (noyau + électrons), $\hat{H}$ est l'opérateur Hamiltonien du système, et *E* est son énergie, valeur propre de l'équation.

L'opérateur Hamiltonien est la somme des différentes contributions à l'énergie du système :

$$E_{totale} = E_{cinétique}(e) + E_{cinétique}(N) + E_{attraction}(N-e) + E_{répulsion}(e-e) + E_{répulsion}(N-N) \qquad \textbf{(2. 2)}$$

Où N sont les noyaux et e^- les électrons.

Pour un système composé de M noyaux et N électrons, l'opérateur Hamiltonien en unités atomiques est égal à :

$$\hat{H} = -\frac{1}{2}\sum_{i=1}^{N}\nabla_i^2 - \frac{1}{2}\sum_{A=1}^{M}\frac{1}{M_A}\nabla_A^2 - \sum_{i=1}^{N}\sum_{A=1}^{M}\frac{Z_A}{r_{iA}} + \sum_{i=1}^{N}\sum_{j>i}^{N}\frac{1}{r_{ij}} + \sum_{A=1}^{M}\sum_{B>A}^{M}\frac{Z_A Z_B}{r_{AB}} \qquad \textbf{(2. 3)}$$

Où *A* et *B* sont les indices courants sur les noyaux, *i* et *j* ceux sur les électrons, les variables *r* représentent les distances inter-particules, *M* et *Z* respectivement les masses et les charges et ∇_q^2 l'opérateur Laplacien.

Cette équation est actuellement impossible à résoudre sans approximation pour des systèmes « réels ».

2. Born-Oppenheimer

L'opérateur Hamiltonien peut être simplifié en utilisant l'approximation de Born-Oppenheimer [5] : les noyaux sont considérés comme fixes par rapport aux électrons. En effet, étant donné que les noyaux ont des masses au moins un millier de fois plus grandes que celle des électrons, on peut considérer que leur mouvement est négligeable. L'énergie cinétique des noyaux devient nulle et l'énergie de répulsion entre noyaux constante. L'énergie relative aux noyaux devient donc un paramètre et

l'énergie du système devient la somme de l'énergie électronique et du terme constant de la répulsion nucléaire.

(2. 4)
$$E_{totale} = E_{el} + \sum_{A=1}^{M}\sum_{B>A}^{M}\frac{Z_A Z_B}{r_{AB}}$$

L'Hamiltonien du système peut donc être réduit à l'Hamiltonien électronique, tout en se rappelant qu'il faudra ajouter l'énergie relative aux noyaux :

(2. 5)
$$\hat{H}_{el} = -\frac{1}{2}\sum_{i=1}^{N}\nabla_i^2 - \sum_{i=1}^{N}\sum_{A=1}^{M}\frac{Z_A}{r_{iA}} + \sum_{i=1}^{N}\sum_{j>i}^{N}\frac{1}{r_{ij}}$$

Soit de façon plus simple :

(2. 6)
$$E_{el} = T_e + V_{Ne} + V_{ee}$$

Avec T_e l'énergie cinétique des électrons, V_{Ne} celle d'attraction noyau-électron et V_{ee} celle de répulsion électron-électron.

L'Hamiltonien électronique peut être exprimé comme la somme d'un terme mono-électronique et d'un terme bi-électronique :

(2. 7)
$$\hat{H}_{el} = \sum_{i=1}^{N}\left(-\frac{1}{2}\nabla_i^2 - \sum_{A=1}^{M}\frac{Z_A}{r_{iA}}\right) + \sum_{i=1}^{N}\sum_{j>i}^{N}\frac{1}{r_{ij}} = \sum_{i=1}^{N}\hat{h}(\vec{r}_i) + \sum_{i=1}^{N}\sum_{j<i}^{N}\frac{1}{r_{ij}}$$

La nouvelle équation permet le calcul de façon exacte de la fonction d'onde électronique Ψ_{el} d'un atome à un électron uniquement. Dans le cas d'un système poly-électronique, on ne peut pas obtenir de solution analytique exacte à l'équation.

3. Approximation orbitalaire

Selon l'approximation orbitalaire, la fonction d'onde électronique peut être décomposée comme le produit de plusieurs fonctions mono-électroniques φ_i (appelées spinorbitales) dans lequel les électrons sont indiscernables.

(2. 8)
$$\Psi_{el} = \varphi_1(\vec{r}_1)\varphi_2(\vec{r}_2)...\varphi_N(\vec{r}_N)$$

Pour respecter le principe de Pauli, la fonction doit être antisymétrique, et pour cela le produit est écrit sous la forme de déterminant de Slater (voir l'équation (2. 9)). Dans ce déterminant, chaque ligne représente les différentes façons de placer un électron dans les spinorbitales : l'échange de deux électrons correspond à l'échange de deux lignes, ce qui conduit au changement de signe de la fonction d'onde.

(2. 9)
$$\Psi^{SD} = \frac{1}{\sqrt{N!}}\begin{vmatrix} \varphi_1(\vec{r}_1) & \varphi_2(\vec{r}_1) & \dots & \varphi_N(\vec{r}_1) \\ \varphi_1(\vec{r}_2) & \varphi_2(\vec{r}_2) & \dots & \varphi_N(\vec{r}_2) \\ \dots & \dots & \dots & \dots \\ \varphi_1(\vec{r}_N) & \varphi_2(\vec{r}_N) & \dots & \varphi_N(\vec{r}_N) \end{vmatrix} = \frac{1}{\sqrt{N!}}\left\| \varphi_1(\vec{r}_1)\varphi_2(\vec{r}_2)\dots\varphi_N(\vec{r}_N) \right\|$$

où $\vec{r}_i$ est le vecteur position de l'électron i.

L'énergie associée au déterminant de Slater est la suivante :

(2. 10)
$$E_{el} = \langle \Psi | H_{el} | \Psi \rangle = \langle \Psi | \sum_{i=1}^{N} \hat{h}(\vec{r}_i) | \Psi \rangle + \langle \Psi | \sum_{i=1}^{N} \sum_{j<i}^{N} \frac{1}{r_{ij}} | \Psi \rangle$$

(2. 11)
$$E_{el} = \sum_{i=1}^{N} h(\vec{r}_i) + \sum_{i=1}^{N} \sum_{j<i}^{N} (J_{ij} - K_{ij})$$

Avec J_{ij} l'intégrale bi-électronique coulombienne et K_{ij} l'intégrale bi-électronique d'échange.

Une méthode de résolution variationnelle est utilisée, avec la contrainte que les spinorbitales sont orthogonales, pour obtenir l'énergie minimale.

4. Equations de Hartree-Fock

Cette dernière approximation amène aux équations de Hartree-Fock, équations aux valeurs propres de l'opérateur de Fock :

(2. 12)
$$\hat{F}(\vec{r}_1) = \hat{h}(\vec{r}_1) + \sum_j \left[\hat{J}_j(\vec{r}_1) - \hat{K}_j(\vec{r}_1) \right]$$

Avec $\hat{J}$ l'opérateur de Coulomb et $\hat{K}$ celui d'échange.

L'opérateur de Fock s'exprime en fonction des orbitales moléculaires que l'on recherche. Il est donc nécessaire d'utiliser une procédure itérative (dite auto-cohérente) pour résoudre l'équation. A partir d'un ensemble d'orbitales d'essai, un nouvel ensemble d'orbitales est obtenu avec le déterminant de Slater et l'énergie associés. Ce processus est répété jusqu'à convergence de l'énergie.

La résolution des équations de Hartree-Fock est possible en utilisant une troisième approximation, LCAO (*Linear Combination of Atomic Orbitals*) : chaque orbitale moléculaire s'exprime comme une combinaison linéaire de fonctions de bases appelées χ_p. De manière générale, ces fonctions de bases sont les orbitales atomiques et les coefficients c_{pi} sont les paramètres variationnels.

(2. 13)
$$\varphi_i = \sum_{p=i}^{M} c_{pi}\chi_p$$

5. Fonctions de bases

Le choix des fonctions de bases doit prendre en compte la nature des fonctions et avoir un sens chimique. Pour cela, les fonctions de Slater (STO – Slater Type Orbitals) sont utilisées. La forme pour une fonction 1s normalisée pour l'atome d'hydrogène est la suivante :

(2. 14) $$\chi_{1s}^{STO}(\zeta,r)=\left(\zeta^3/\pi\right)^{1/2}e^{-\zeta r}$$

Où ζ est l'exposant de la fonction de Slater.

Mais la forme des fonctions doit aussi permettre un calcul efficace des intégrales. Or, les fonctions de Slater ne permettant pas un calcul aisé des intégrales, des fonctions de type gaussiennes sont aussi utilisées : *Gaussian Type Orbitals* (*GTO*). La forme est la suivante pour une fonction 1s normalisée pour l'atome d'hydrogène :

(2. 15) $$\chi_{1s}^{GTO}(\alpha,r)=\left(2\alpha/\pi\right)^{3/4}e^{-\alpha r^2}$$

Où α est l'exposant de la fonction gaussienne.

L'inconvénient des ces dernières est qu'elles représentent mal les orbitales atomiques au voisinage du noyau, contrairement aux fonctions de Slater. Pour pallier ce problème, des combinaisons de fonctions gaussiennes (dites primitives) sont utilisées pour donner de nouvelles fonctions gaussiennes appelées contractées (voir Figure 4). Les fonctions de bases χ_p des orbitales moléculaires sont donc des fonctions contractées de type :

(2. 16) $$\chi_p=\sum_s d_{sp}g_s$$

où g est une fonction gaussienne et d_{sp} les coefficients de contraction.

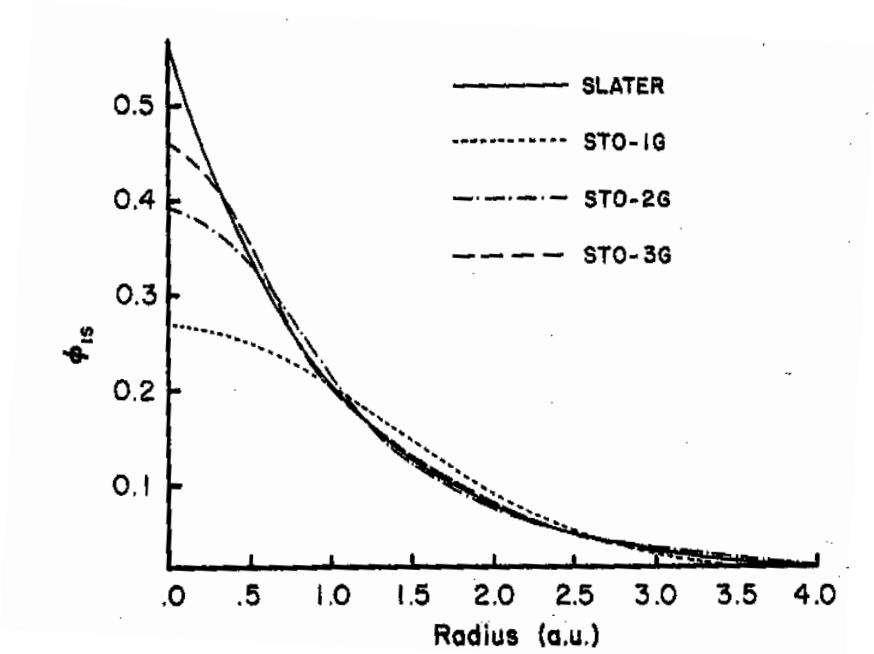


Figure 4: Comparaison de la représentation d'une fonction de Slater 1s à partir de gaussiennes contractées STO-1G, STO-2G et STO-3 dans la référence[2].

Des fonctions supplémentaires peuvent encore être ajoutées, comme les fonctions de polarisation qui donnent de la flexibilité angulaire aux orbitales de valence. Elles sont notamment utilisées pour la description des systèmes polarisés. Les fonctions diffuses sont des fonctions gaussiennes caractérisées par des exposants très faibles qui servent à décrire les régions loin du noyau. Elles sont en particulier nécessaires pour décrire des anions et pour la description correcte des interactions intermoléculaires.

II. Au-delà de Hartree-Fock

La méthode Hartree-Fock (HF) ne prend pas en compte la corrélation électronique (i.e. le mouvement des électrons) : chaque électron voit tous les autres comme un champ approché. L'énergie de corrélation E_{corr} est ainsi définie comme la différence entre l'énergie exacte E_{exacte} et l'énergie calculée en HF pour une base complète E_{HF} .

(2. 17) $$E_{corr} = E_{exacte} - E_{HF} < 0$$

D'autres approximations ont été développées pour résoudre ce problème.

La méthode qui a été choisie dans cette thèse pour l'optimisation des structures et le calcul de l'énergie et des fréquences est la théorie de la fonctionnelle de la densité (DFT). Cette méthode a l'avantage de pouvoir modéliser des molécules de taille assez grande avec un temps de calcul raisonnable.

1. Théorie de la fonctionnelle de la densité

La Théorie de la Fonctionnelle de la Densité est fondée sur le postulat datant de 1927 issu des travaux de Thomas et Fermi [6]. Il s'agit de dire que le système peut être caractérisé par la densité électronique.

L'un des avantages de cette méthode est l'utilisation d'une observable physique mesurable : la densité électronique $\rho(\vec{r})$, au lieu de la fonction d'onde Ψ elle même. La densité électronique est définie comme le nombre d'électrons par unité de volume et est liée à la fonction d'onde par l'équation suivante :

(2. 18) $$\rho(\vec{r}) = \Psi^*(\vec{r})\Psi(\vec{r}) = \left|\Psi^2(\vec{r})\right|$$

L'intégration de la densité, fonction des coordonnées spatiales xyz sur tout l'espace, est égale au nombre d'électrons du système.

a) Théorèmes de Hohenberg-Kohn

En 1964, Hohenberg et Kohn reformulent l'approche de Fermi en théorie exacte d'un système à plusieurs corps, à l'aide de plusieurs théorèmes. Le premier théorème [7], appelé « théorème

d'existence », démontre que pour tout système de particules en interaction dans un potentiel externe $V_{ext}(\vec{r})$, le système est déterminé de manière unique par la densité à l'état fondamental.

$$E[\rho] = F_{HK}[\rho] + \int \rho(\vec{r}) V_{ext}(\vec{r}) d\vec{r} \quad \textbf{(2. 19)}$$

$$F_{HK}[\rho] = T[\rho] + V_{ee}[\rho] \quad \textbf{(2. 20)}$$

Où $F_{HK}[\rho]$ est la fonctionnelle universelle. Celle-ci est constituée des termes indépendants de $V_{ext}(\vec{r})$ et regroupe l'énergie cinétique des électrons ($T[\rho]$) et la répulsion inter-électronique ($V_{ee}[\rho]$).

En d'autres termes, un système à l'état fondamental et toutes ses propriétés observables peuvent entièrement être déterminés par sa densité électronique totale ρ en tout point.

L'énergie, en particulier, est donc une fonctionnelle de la densité : $E_0 = F[\rho]$

Le second théorème est analogue au principe variationnel : l'énergie, fonctionnelle d'une densité électronique approchée, est supérieure ou égale à l'énergie exacte du système dans son état fondamental.

$$F_{HK}[\rho] + V_{eN}[\rho] = E[\rho] \geq E[\rho_0] = E_0 \quad \textbf{(2. 21)}$$

Déterminer la densité ρ pour laquelle l'énergie est minimale à partir de la fonction d'onde correspondante Ψ peut se faire via l'équation (2. 18), mais une infinité de fonctions d'onde peuvent être associées à une même densité. Obtenir la fonction d'onde de l'état fondamental à partir de la densité électronique n'est pas chose évidente. La résolution de l'équation de Schrödinger est donc, encore une fois, impossible à cause de la présence du terme d'interaction électronique dans l'Hamiltonien électronique.

b) Approche Kohn-Sham

En 1965, une nouvelle approche va permettre de pallier ce problème : l'approche Kohn-Sham [8]. Elle repose sur l'utilisation d'un système fictif d'électrons sans interaction, dans lequel chaque électron individuel voit simplement un potentiel constant avec lequel il interagit. Ce potentiel a la même densité électronique que le système réel.

Pour calculer $T[\rho]$, on utilise un système fictif d'électrons sans interaction ayant la même densité que le système réel, c'est-à-dire que chaque électron individuel voit simplement un potentiel avec lequel il interagit. L'équation finale devient alors :

$$E[\rho] = T_{ni}[\rho] + V_{Ne}[\rho] + J[\rho] + E_{xc}[\rho] \quad \textbf{(2. 22)}$$

Où $T_{ni}[\rho]$ est l'énergie cinétique des électrons sans interaction, $V_{Ne}[\rho]$ l'énergie d'attraction électrons-noyaux, $J[\rho]$ l'énergie d'interaction coulombienne, $E_{xc}[\rho]$ l'énergie d'échange et corrélation.

L'artifice du système fictif d'électrons non-interagissant, mais ayant la même densité que le système réel, permet de réintroduire un déterminant de Slater composé de spinorbitales φ_i appelées orbitales de Kohn-Sham.

c) Fonctionnelles d'échange et corrélation

Finalement, la seule inconnue est l'expression exacte du terme d'énergie d'échange et corrélation $E_{xc}[\rho]$. La connaissance de ce terme permettrait de résoudre l'équation en utilisant la méthode variationnelle.

L'énergie d'échange et corrélation est calculée à l'aide de fonctionnelles d'échange et corrélation définies comme :

(2. 23) $$E_{xc}[\rho] = \int F_{xc}[\rho] d\vec{r}$$

Pour plus de clarté, il est utile d'adopter des notations usuelles. Par exemple, la fonction de dépendance de E_{xc} de la densité électronique est exprimée comme interaction entre ρ et une densité d'énergie ε_{xc} qui dépend de ρ,

(2. 24) $$E_{xc}[\rho] = \int \rho . \varepsilon_{xc}[\rho] d\vec{r}$$

où l'énergie d'échange et corrélation s'exprime comme la somme de deux contributions différentes, une d'échange, l'autre de corrélation.

(2. 25) $$E_{xc}[\rho] = E_x[\rho] + E_{corr}[\rho]$$

Plusieurs fonctionnelles ont ainsi été développées pour traiter ces deux contributions.

- <u>Approximation de la densité locale (LDA)</u>

Dans cette approximation, la densité électronique est supposée localement uniforme [9]. L'énergie E_{xc}^{LDA} est calculée suivant l'équation (2. 24) et est considérée comme dépendante uniquement de la densité électronique. Ces méthodes fournissent souvent d'assez bonnes propriétés moléculaires (géométrie, fréquences) mais conduisent généralement à de très mauvaises données énergétiques telles que les énergies de liaison.

- <u>Approximation du gradient généralisé (GGA)</u>

L'approximation du gradient généralisé considère des fonctions d'échange et corrélation dépendant non seulement de la densité électronique en chaque point, mais aussi de son gradient.

(2. 26) $$E_{xc}^{GGA}[\rho] = \int \rho \cdot \varepsilon_{xc}[\rho, \nabla\rho] d\vec{r}$$

Un grand nombre de fonctionnelles de type GGA ont ainsi été proposées, par exemple les fonctionnelles d'échange proposées par Becke (B88 [10]ou B97 [11]) ou la fonctionnelle de corrélation proposée par Lee, Yang et Parr (LYP) [12].

- Fonctionnelles hybrides

Pour le troisième type de fonctionnelle, il s'agit de combiner les méthodes LDA ou GGA avec la méthode HF qui traite correctement l'énergie d'échange. En pratique, une fraction d'échange Hartree-Fock est intégrée. Les fonctionnelles hybrides se composent donc de deux contributions distinctes, DFT et HF.

La fonctionnelle hybride la plus utilisée dans la littérature est B3LYP [13,14], elle est comptée parmi les fonctionnelles hybrides les plus utilisées pour l'étude de systèmes moléculaires. Cette fonctionnelle à 3 paramètres repose sur les termes d'échange de Becke [10] et de corrélation de Lee, Yang et Parr [12].

$$E_{xc}^{B3LYP} = (1-a)E_x^{LSDA} + aE_x^{HF} + b\Delta E_x^{B} + (1-c)E_c^{LSDA} + cE_c^{LYP} \quad \textbf{(2. 27)}$$

où les paramètres a, b et c ont été ajustés respectivement à 0,20, 0,72 et 0,81 [13] par calcul des énergies d'atomisation, des potentiels d'ionisation et des affinités électroniques d'une série de petites molécules [15].

Dans ces travaux de thèse, la fonctionnelle PBE0 [16] a été employée pour l'optimisation des structures. Cette fonctionnelle peut être considérée comme non-paramétrée, le pourcentage d'échange Hartree-Fock (25%) n'ayant pas été ajusté sur des critères empiriques mais fixé sur des considérations physiques.

$$E_{xc}^{PBE0} = E_{xc}^{PBE} + \frac{1}{4}\left(E_x^{HF} - E_x^{PBE}\right) \quad \textbf{(2. 28)}$$

Où E_{xc}^{PBE} et E_x^{PBE} sont respectivement l'énergie d'échange et corrélation et l'énergie d'échange calculée avec la fonctionnelle PBE et E_x^{HF} l'énergie d'échange calculée avec la méthode Hartree-Fock.

Les fonctionnelles hybrides sont très couramment employées pour leur bonne précision dans une gamme importante de propriétés moléculaires. Alors que les fonctionnelles de la densité présentées auparavant avaient tendance à entraîner des surestimations de certaines grandeurs, la théorie HF les sous-estime. C'est, par exemple, le cas pour les longueurs de liaisons [17]. Une combinaison de ces deux approches propose des résultats plus fiables.

2. DFT conceptuelle

La DFT « classique » (appelée DFT computationnelle par Parr et Yang [18]) calcule des structures électroniques tandis que la DFT conceptuelle [19–21], développée à partir des années 1980, permet de calculer des propriétés liées à la réactivité. Elle a pour but de donner un cadre théorique et des définitions précises à des concepts largement utilisés par les chimistes expérimentateurs. De nombreux descripteurs de la réactivité ont été proposés à partir de dérivées successives de l'énergie. La Figure 5 présente comment ils peuvent être calculés.

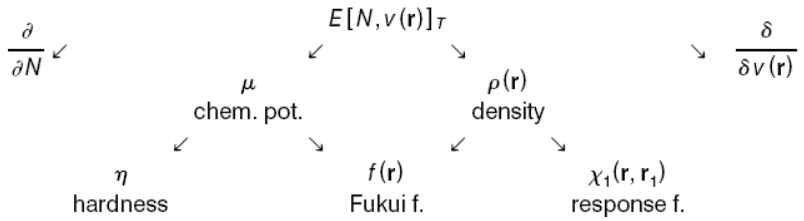


Figure 5: Dérivées de l'énergie aux 1er et 2nd ordre [19]

Ces dérivées de l'énergie peuvent être classées en trois catégories (illustrées Figure 5) :

- Grandeurs globales : dérivées par rapport à N, le nombre d'électrons, (tel que le potentiel chimique μ et l'électronégativité χ, la dureté η, la mollesse S et l'indice d'électrophilicité ω)
- Locales : dérivées qui dépendent d'une coordonnée spatiale uniquement
- Non locales : dérivées qui dépendent de deux coordonnées spatiales

Le potentiel chimique μ mesure la tendance du nuage électronique à s'échapper de la molécule. L'électronégativité est l'opposé du potentiel chimique et mesure donc la tendance à ne pas laisser s'échapper les électrons du système. Cette définition de l'électronégativité est celle selon la DFT (Iscowski et Margave) dont la différence finie amène à retrouver celle de Mulliken.

$$\mu = -\chi = \left(\frac{\partial E}{\partial N}\right)_{v(r)} = -\frac{1}{2}(I + A) \quad \textbf{(2. 29)}$$

Où I est l'énergie de première ionisation et A l'électroaffinité.

En 1963, Pearson [22] énonce le concept de la dureté et de la mollesse des acides et bases de Lewis. Ce principe (HSAB : *Hard and Soft Acids and Bases*) considère que les bases et les acides durs préfèrent réagir ensemble alors que les bases molles préfèrent réagir avec les acides mous. La dureté η [23] est une valeur positive qui peut être vue comme la résistance d'un système moléculaire au transfert d'électron. Il s'agit d'une mesure de la stabilité d'une molécule. La mollesse qui est la propriété inverse correspond à la capacité d'un système à conserver une charge acquise [24].

$$\eta = \left(\frac{\partial^2 E}{\partial N^2}\right)_{v(r)} = I - A \quad \textbf{(2. 30)}$$

Selon la théorie des orbitales frontières (ou théorème de Fukui) [25] et le théorème de Koopmans [26] dans l'approximation des orbitales gelées, la dureté peut aussi être calculée de la manière suivante :

$$\eta = E_{LUMO} - E_{HOMO} \quad \textbf{(2. 31)}$$

L'indice d'électrophilicité, quant à lui, mesure le caractère électrophile d'une molécule, c'est-à-dire sa capacité à attirer les électrons. Il est défini comme la stabilisation énergétique due au transfert de charge.

$$\omega = \frac{\mu^2}{2\eta} \quad \textbf{(2. 32)}$$

Au niveau des grandeurs locales, les fonctions de Fukui [27] f(r) sont définies comme la réponse de la densité électronique lorsque le nombre d'électrons change. Elles indiquent les sites du système les plus réactifs. La mollesse locale s(r), définie à potentiel constant, permet de déterminer la réactivité au sens local des molécules[28].

$$f(r) = \left(\frac{\partial \rho(r)}{\partial N} \right)_V \quad \textbf{(2. 33)}$$

$$s(r) = \left(\frac{\partial \rho(r)}{\partial \mu} \right) \quad \textbf{(2. 34)}$$

Un programme a été développé au laboratoire par Li Rao pour calculer les fonctions de Fukui et la mollesse locale en se basant sur l'approximation des orbitales gelées. Les équations (2. 35) et (2. 36) sont donc utilisées.

$$\begin{cases} f^-(r) = |\Phi(HOMO)|^2 \\ f^+(r) = |\Phi(LUMO)|^2 \end{cases} \quad \textbf{(2. 35)}$$

$$\begin{cases} s^-(r) = \dfrac{1}{E_{LUMO} - E_{HOMO}} . f^-(r) \\ s^+(r) = \dfrac{1}{E_{LUMO} - E_{HOMO}} . f^+(r) \end{cases} \quad \textbf{(2. 36)}$$

Ces dérivées pourront être utilisées dans le développement de modèle QSPR afin de comprendre la réactivité reliée aux propriétés étudiées.

III. MÉCANIQUE MOLÉCULAIRE

Dans cette partie, les bases de la mécanique moléculaire[3,4] qui sera utilisée au cours de l'étude des conformations seront brièvement décrites. Cette méthode, contrairement aux méthodes quantiques, permet le traitement de plus grosses molécules, comme les protéines, en un temps de calcul raisonnable. La mécanique moléculaire revient à appliquer les méthodes de mécanique classique à

une molécule. Les atomes (électrons + noyau) sont considérés comme des sphères rigides (de rayon et charge définis) et les liaisons comme des ressorts. Le calcul de l'énergie d'une molécule se fait en fonction des coordonnées des atomes. L'énergie totale de la molécule est calculée comme une somme de termes additifs sans interaction :

(2. 37) $$E_{tot} = E_{liaison} + E_{angle} + E_{torsion} + E_{non-liées}$$

Où $E_{liaison}$ est l'énergie de déformation des liaisons, E_{angle} celle des angles, $E_{torsion}$ celle des angles dièdres et $E_{non-liées}$ l'énergie des interactions non liées.

La plupart des champs de forces calculent l'énergie de déformations des liaisons et angles en considérant le système comme un oscillateur harmonique.

(2. 38) $$E_{liaison} = \sum_{Nliaisons} \frac{1}{2} k_l (r - r_0)^2$$

où r est la longueur de la liaison (en Å), r_0 la valeur de la liaison de référence et k_l la constante de force (en kcal/mol).

(2. 39) $$E_{angle} = \sum_{Nangles} \frac{1}{2} k_\theta (\theta - \theta_0)^2$$

où θ est l'angle (en degré), θ_0 la valeur de référence et k_θ la constante de force associée (en kcal/mol).

Contrairement aux deux énergies précédentes, l'énergie de torsion est décrite par une équation contenant un cosinus qui reproduit la périodicité de la rotation autour d'une liaison.

(2. 40) $$E_{diedre} = \sum_{Ndiedre} k_d \left[1 + \cos(n\phi - \phi_0)\right]$$

Où est ϕ l'angle de torsion, ϕ_0 celui de référence, k_d est la constante de force associées à chaque angle dièdre et n la périodicité.

Les énergies non liées sont les interactions électrostatiques et de Van der Waals (VDW) modélisées respectivement par le potentiel de Coulomb et les équations de Lennard-Jones [29].

(2. 41) $$E_{non-liée} = \sum_{N_{nb}} \left\{ \frac{q_i q_j}{r_{ij}} + 4\varepsilon_{ij} \left[\left(\frac{\sigma_{ij}}{r_{ij}} \right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}} \right)^{6} \right] \right\}$$

Où N_{nb} est le nombre de paires d'atomes pour lesquels il y a des interactions non-liées, r_{ij} est la distance entre les atomes i et j tandis que q_i et q_j sont leur charge partielle, ε_{ij} est la profondeur du puits de potentiel et σ_{ij} la distance à laquelle l'interaction entre atomes est considérée comme nulle.

(2. 42) $$\varepsilon_{ij} = \sqrt{\varepsilon_i \varepsilon_j}$$

(2. 43)
$$\sigma_{ij} = \frac{1}{2}\left(\sigma_i + \sigma_j\right)$$

Les termes d'interactions non-liantes pour les atomes éloignés de plus d'une ou deux liaisons sont possibles mais généralement non considérés.

1. Paramétrisation des champs de forces

Tout comme il existe différentes méthodes de calcul *ab initio*, la mécanique moléculaire présente plusieurs méthodes qui sont caractérisées par un champ de force. Un champ de force est généralement conçu pour prédire certaines propriétés et sera paramétré dans ce sens. Le principe des champs de force est la transférabilité : les mêmes paramètres sont utilisés pour une série de molécules (par exemple les alcanes), dans le but de faire de la prédiction. Les champs de force sont empiriques, il n'y a pas de forme « correcte ». Ils sont paramétrés à partir d'une base de données expérimentale la plus grande possible (RMN, RX) voire obtenue par des calculs *ab initio*. Les paramètres sont choisis selon un compromis entre la précision et le temps de calcul.

A chaque atome est associé un type atomique, plus précis que le numéro atomique, qui prend en compte les informations sur l'hybridation et l'environnement local. L'atome de carbone peut ainsi avoir plusieurs types : C_{sp}, C_{sp}^2, C_{sp}^3 et encore plus de constantes de force de liaison : C_{sp}^3-C_{sp}^3, C_{sp}^3-C_{sp}^2, C_{sp}^2= C_{sp}^2, C_{sp}^2=O, C_{sp}^3 –N_{sp}^3 et C-N_{amide}.

Les principaux champs de force sont : MM2/MM3/MM4 d'Allinger[30] pour les petites molécules organiques, CHARMM[31] pour les biomolécules et les macromolécules, AMBER[32] pour les protéines et peptides et UFF[33] pour les molécules organométalliques.

2. Limites de la méthode

La mécanique moléculaire présente cependant des limites : par définition, les électrons ne sont pas pris en compte, ce qui rend cette méthode non adaptée aux problèmes dans lesquels les effets électroniques sont prédominants. De plus, les champs de force sont optimisés pour une famille de molécules et ne peuvent pas être généralisés à toutes les molécules.

Dans cette thèse, des composés organiques aliphatiques, et donc flexibles, mais de petite taille sont étudiés. La mécanique moléculaire, avec le champ de force MM3 [30], spécialisé dans le traitement de petites molécules organiques, sera donc utilisée dans la partie sur l'analyse conformationnelle. En effet, le logiciel Scigress[34], générateur de toutes les conformations possibles pour une molécule donnée, optimise la géométrie des structures obtenues à ce niveau de calcul pour supprimer les structures identiques.

IV. ANALYSE CONFORMATIONNELLE

Les propriétés des molécules sont dépendantes de la conformation. De nombreux domaines de la chimie nécessitent la structure la plus stable c'est-à-dire celle d'énergie minimale. L'utilisation de programmes générateurs de conformations permet de parcourir l'espace chimique à la recherche de cette structure. Cependant, un très grand nombre de conformations sont alors générées (par exemple, on obtient 235 conformations pour le 1,1,1,6,6,6-hexanitro-3-hexyne en utilisant le logiciel Scigress[34]. L'analyse et le calcul de l'énergie de toutes ces conformations sont très longs, voire impossibles, en utilisant des méthodes telles que la DFT. Dans le but de faciliter cette recherche, un programme qui réduit le nombre de conformations à étudier, tout en conservant la répartition des conformations dans l'espace chimique, a été développé. Ce programme, appelé Callisto (pour Conformational Analysis In Silico et qui doit être utilisé après un échantillonnage conformationnel), a pour objectif une rapidité qui ne peut être atteinte en utilisant des méthodes telles que la dynamique moléculaire ou une analyse de Monte Carlo. L'installation, l'utilisation et les résultats du programme sont expliqués en annexe III.

1. Principe et fonctionnement

a) Principe du programme

La méthode utilisée pour réduire le nombre de conformations est une analyse par clustering de toutes les conformations possibles suivie d'une analyse de population de Boltzmann. La méthode de clustering hiérarchique basée sur la matrice des RMSD (root-mean-square deviation) est utilisée. Cette matrice est une matrice carrée de taille N*N, où N est le nombre de conformations, qui contient la valeur de RMSD entre les coordonnées de chaque paire de conformation. Les deux types de clustering hiérarchique ont été programmés : l'agglomératif car c'est celui qui est utilisé dans la littérature[35] et le divisif car il est moins enclin à former des singletons (clusters contenant une seule conformation). Finalement, les conformations d'énergies minimales restant après l'analyse de Boltzmann sont sélectionnées. Ce programme a l'avantage d'être rapide et présente la possibilité de valider la pertinence du nombre de conformations finales sélectionnées.

b) Méthodes de clustering

La classification[36,37] (aussi appelé *clustering*) est une méthode descriptive et non prédictive. La classification est une méthode d'apprentissage non supervisée qui permet le regroupement de données sur un critère de similarité (dans notre cas, la valeur de RMSD) : les données ayant des caractéristiques similaires sont regroupées dans un même sous-ensemble, les données aux caractéristiques différentes sont séparées. Le nombre de classes n'est pas défini à l'avance et la classe à laquelle appartiennent les données n'est pas connue.

Il existe plusieurs types de clustering[36] : méthodes de partitionnement pour lesquelles le nombre de clusters finaux doit généralement être déterminé à l'avance (k-mean, réseaux kohonen, densité…), méthodes hiérarchiques, ou encore les méthodes « floues » (*fuzzy*) lorsque les données peuvent être dans plusieurs clusters. Les méthodes de partitionnement ont l'inconvénient de devoir fixer à l'avance le nombre k de clusters finaux. Les méthodes floues sont des méthodes compliquées et rarement utilisées. Ce sont donc les méthodes hiérarchiques qui sont utilisées dans ce programme car elles permettent un classement progressif, sous forme d'arbre appelé dendrogramme, illustré en Figure 6.

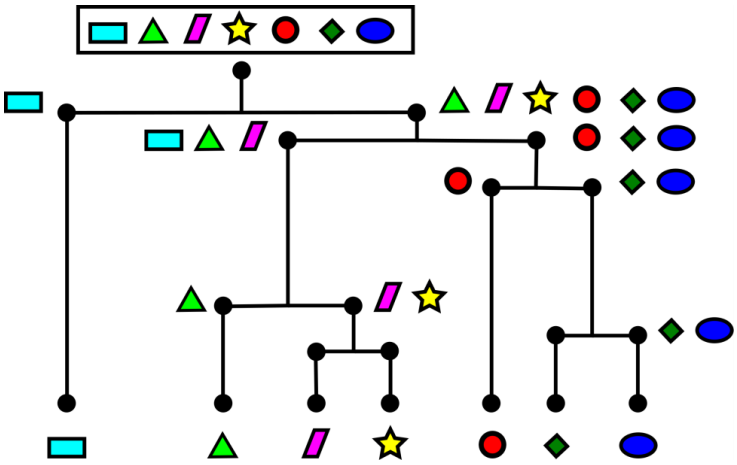


Figure 6: Exemple de dendrogramme - Clustering hiérarchique

Une valeur seuil est choisie par l'utilisateur pour couper le dendrogramme : le nombre de clusters finaux ou une valeur seuil de similarité. Ce dernier critère ne peut être utilisé que dans le cas de d'une méthode hiérarchique. Les deux types de clustering hiérarchique sont implémentés dans callisto.

La fusion des clusters dite *linkage* peut être faite de différentes façons. Par exemple le « single link » correspond à la fusion des deux clusters dont la distance minimale des objets de classes différentes est la plus petite, au contraire le « complete link » fusion ceux dont la distance maximale entre deux objets de classes différentes est minimale (exemple illustré par la Figure 7). De même pour « l'average link » où la distance considère est la moyenne des distances.

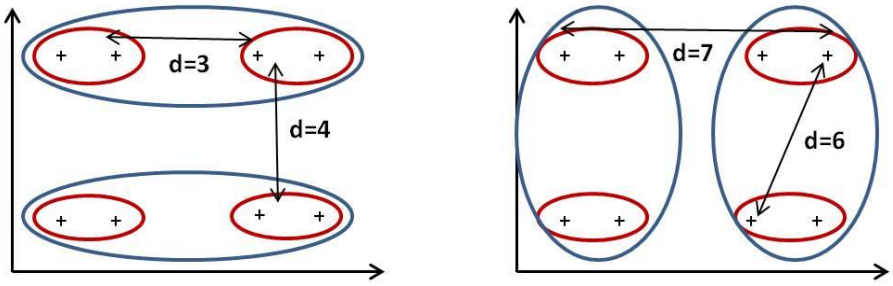


Figure 7: Illustration des différents linkage (single link à gauche, complete link à droite)

c) Validation du clustering

Il existe plusieurs méthodes de validation[38–40] du nombre de clusters choisis qui permettent la sélection du nombre de clusters le plus adapté. Elles peuvent être graphiques comme la PCA mais aussi plus quantitatives avec le calcul de différents indices tels que : l'indice de Dunn[41], l'indice de Davies Bouldin[42], l'indice SD validation[43]… La méthode de validation utilisée dans le programme est la méthode « silhouette », développée par Rousseeuw en 1987[40], qui ne dépend pas de celle utilisée pour réaliser la classification. Cette méthode est à la fois graphique et quantitative avec le calcul d'un coefficient S, moyenne des coefficients s(i). Le coefficient s(i) d'un objet i de la base de données, dont la valeur se situe entre -1 et 1, se calcule avec la formule générale suivante :

$$s(i) = \frac{b(i) - a(i)}{\max\left(a(i), b(i)\right)} \qquad \textbf{(2. 44)}$$

Avec a(i) la dissimilarité moyenne de l'objet i par rapport à tous les objets du cluster A
et b(i) la dissimilarité moyenne de l'objet i par rapport à tous les objets du cluster B.

Cette formule peut aussi s'écrire :

$$s(i) = \begin{cases} 1 - a(i)/b(i) & \text{si a(i) < b(i),} \\ 0 & \text{si a(i) = b(i),} \\ b(i)/a(i) - 1 & \text{si a(i) > b(i).} \end{cases} \qquad \textbf{(2. 45)}$$

La dissimilarité moyenne d'un cluster mesure à quel point ce cluster est compact. Le coefficient s(i) mesure l'homogénéité et la séparation des clusters. Quand s(i) est proche de 1, on peut dire que l'objet i est « bien classé ». Quand s(i) est zéro alors a(i) et b(i) sont à peu près égaux, la classification de i dans le cluster A ou B n'est pas claire dans ce cas. La pire des situations est lorsque la valeur de s(i) est négative. Dans ce cas, a(i) est beaucoup plus grand que b(i) ce qui signifie que i est en moyenne plus proche du cluster B que du cluster A et donc que l'objet a été « mal classé ».

La moyenne de s(i) de tous les objets correspond au S de la base de données qui mesure comment les données ont été convenablement groupées. Plus ce coefficient (qui doit être supérieur à 0) est élevé, meilleur est le découpage. Pour comparer le choix du nombre de clusters (k) choisi, il suffit donc de regarder pour quel k la valeur de S est maximum. La silhouette de la classification obtenue peut être représentée de manière graphique comme illustré Figure 8.

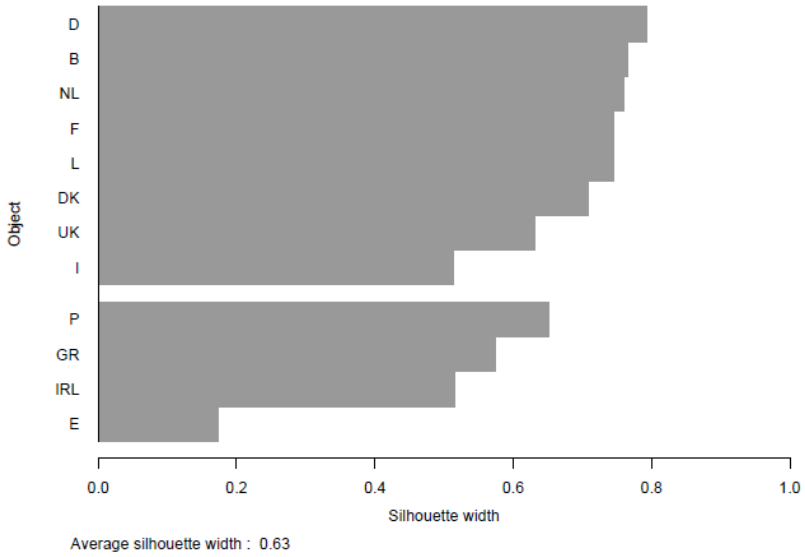


Figure 8: Exemple de graphe du profil silhouette obtenue avec le logiciel R[44]

2. Validation du programme

L'installation du programme, rédigé en python, nécessite la librairie python numpy[45]. Le clustering a été effectué grâce au package fortran « twins » et à la librairie R[44] « sildist ».

a) RMSD et clustering

La valeur seuil de RMSD a évidement une influence sur le nombre de clusters sélectionnés. Plus la valeur est basse, plus le nombre de clusters augmente. Cela s'explique facilement par le fait que la valeur de RMSD mesure la dissimilarité entre paires de molécules et donc la distance entre les clusters. La valeur seuil de RMSD est la valeur minimale entre chaque cluster non fusionné, chaque cluster doit donc avoir une valeur de dissimilarité supérieure à la valeur seuil. Plus cette valeur est faible, plus il est difficile de grouper les clusters. L'analyse en composante principale (PCA expliquée dans le chapitre 3) permet de représenter les données dans l'espace et d'observer l'influence de la valeur seuil choisie. Un graphique utilisant les deux premières composantes principales représente la position des données les unes par rapport aux autres. Les données proches sur le graphique sont similaires et peuvent être assez facilement regroupées. Une façon de valider la méthode de clustering choisie dans cette étude est de représenter les clusters par une PCA en colorant les molécules d'un même cluster avec la même couleur et de vérifier si elles sont effectivement proches. En représentant par PCA les clusters obtenus pour le 2,2,4,6,6-pentanitroheptane, la Figure 9 confirme que les molécules présentes dans un même cluster sont proches dans l'espace chimique.

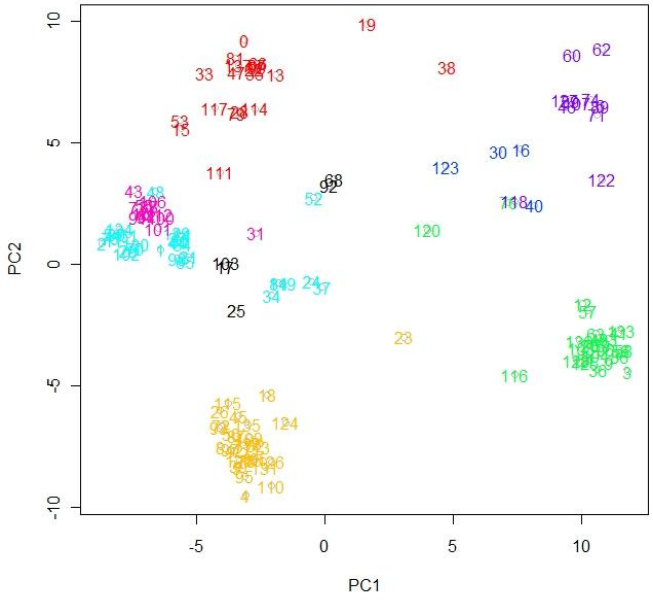
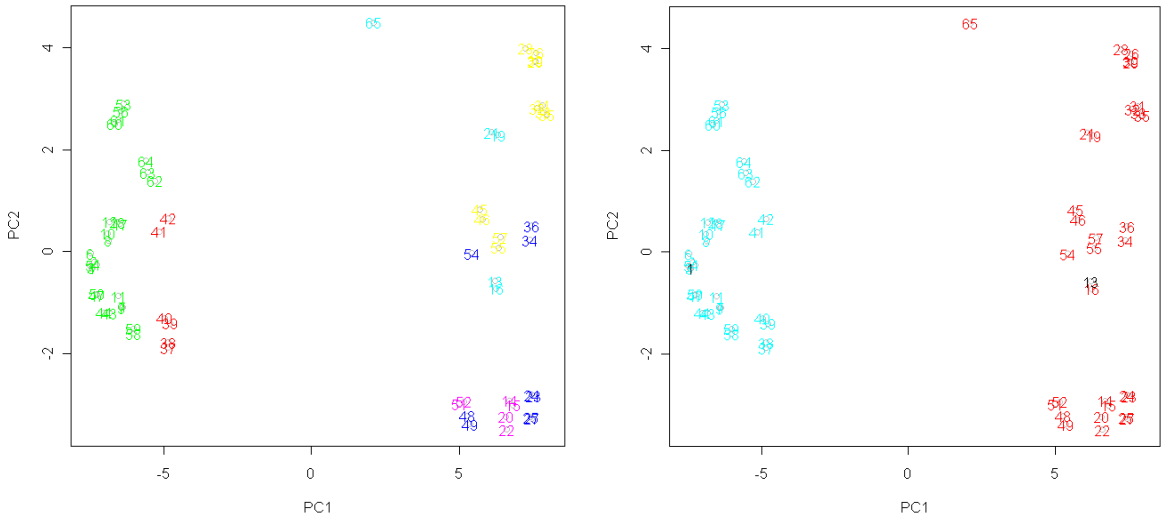


Figure 9: PCA – Clustering agglomératif average link du 2,2,4,6,6-pentanitroheptane avec RMSD =1,5

Les Figures 10 a et b démontrent l'importance du choix de la valeur seuil utilisée. Ainsi un seuil trop faible peut découper en un nombre de clusters qui n'est pas réellement représentatif de la répartition des conformations dans l'espace chimique. C'est pour cette raison que la validation du nombre de clusters par la méthode « silhouette » a été ajoutée dans Callisto.



Figures 10: PCA – Clustering agglomératif average link du di-2,4-dichlorobenzoyl peroxide a) RMSD=1,5 et b) RMSD=2

b) Choix de la méthode

Une base de données de 49 composés nitroaliphatiques a été utilisée[46] pour tester le programme. Une analyse conformationnelle a été effectuée pour chacune des molécules puis les conformations ont été optimisées par le champ de force MM3[30] (choix par default) avec le logiciel Scigress[34]. Chaque ensemble de conformations générées a ensuite été traité par Callisto avec trois types de clustering : divisif, agglomératif avec single link et avec average link.

Pour l'un de ces composés, le 1,1,1,6,6,6-hexanitro-3-hexyne qui a 235 conformations, le profil silhouette moyen est calculé en fonction du nombre de clusters finaux (Figure 11.a). Pour les trois types de clustering, le profil silhouette moyen présente un maximum local (0,39) pour 10 clusters et un maximum global (0,53) pour environ 148 clusters. D'une part, ces 235 conformations peuvent donc être regroupées en 148 clusters, d'autre part en fonction des applications ultérieures elles peuvent être regroupées en seulement 10 clusters mais avec une perte de précision. L'algorithme divisif et l'agglomératif average link donnent un profil silhouette similaire. Le clustering agglomératif single link donne les moins bonnes valeurs de silhouette mais correspond aux courbes obtenues avec les deux autres algorithmes aux alentours des maxima.
Le même procédé a été effectué pour les 48 autres molécules et leur ensemble de conformations. La valeur maximale du profil silhouette moyen a été représentée pour chaque molécule sur la Figure 11.b. Pour chaque algorithme, la moyenne du profil silhouette moyen calculée sur l'ensemble des 49 molécules est représenté par une droite horizontale. La moyenne maximale est obtenue avec le clustering agglomératif average link avec une valeur de 0,35 contre une moyenne de 0,33 et 0,32 pour les clustering divisif et agglomératif single link respectivement. Cette observation confirme que le clustering agglomératif single link est le moins performant.

La Figure 11.c représente le nombre de clusters finaux associé à ces valeurs maximales pour chaque molécule. La tendance observée précédemment est conservée avec une moyenne de 25, 28 et 31 clusters pour les clustering divisif, agglomératif « average link » et agglomératif « single link » respectivement. La Figure 11.d représente les valeurs seuil de RMSD correspondant à ce nombre de clusters pour chaque molécule. On peut remarquer que malgré une méthode de sélection des conformations identique, on obtient une valeur seuil très différente selon le type de clustering choisi. Le clustering agglomératif « average link » semble être le meilleur choix ici. Les molécules sélectionnées par Callisto peuvent ensuite être optimisées par une méthode telle que la DFT pour sélectionner la géométrie de plus basse énergie sans avoir à optimiser toutes les conformations.
Una application de ce programme sur le calcul des propriétés sera faite dans le chapitre 5.

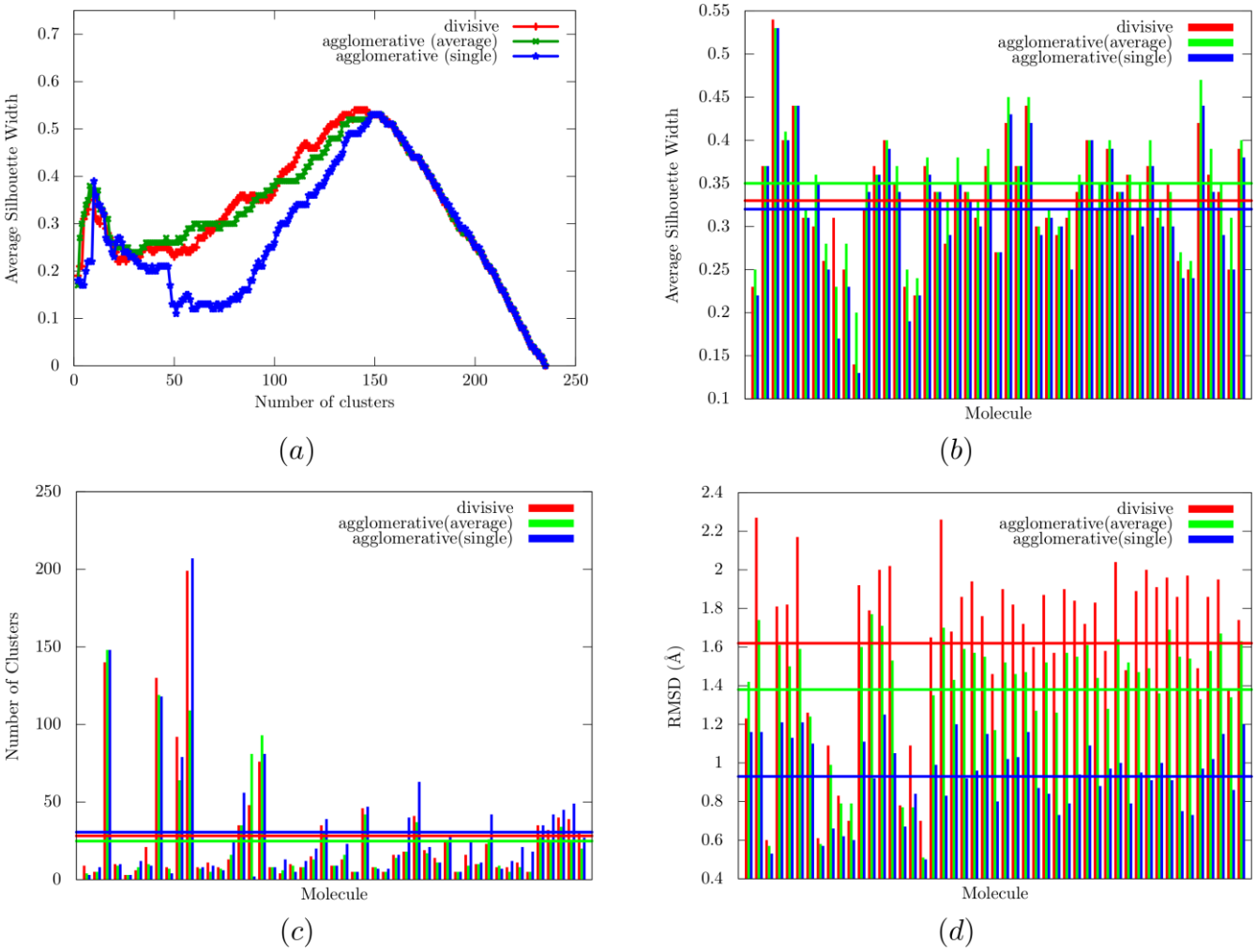


Figure 11: Représentation des résultats pour les 49 composés nitroaliphatiques après l'utilisation de Callisto. (a) Pour le « 1,1,1,6,6,6-hexanitro-3-hexyne » (composé numéro 3), la silhouette moyenne est représentée en fonction du nombre de clusters sélectionnés. (b) Représentation de la silhouette maximale pour chaque molécule, (c) du nombre de clusters associé et (d) de la valeur seuil de RMSD associée.

V. CONCLUSION

Cette partie a présenté les bases de la chimie théorique, des méthodes quantiques de l'équation de Schrödinger à la DFT, ainsi que la mécanique moléculaire. Ces méthodes nous servirons pour l'optimisation de la géométrie des molécules avant d'utiliser leur structure pour le développement de modèles QSPR.

Puis, la problématique de l'analyse conformationnelle, qui est un très vaste domaine, a été abordée. Au cours de nos travaux nous nous sommes intéressés à l'effet de la conformation sur les modèles et sur la recherche de la structure la plus stable.

VI. RÉFÉRENCES

(1) Cramer, C. J. *Essentials of computational chemistry*; Wiley: Chichester, 2004.

(2) Szabo, A.; Ostlund, N. S. *Modern quantum chemistry : introduction to advanced electronic structure theory*; Dover Publications: Mineola, N.Y., 1996.

(3) Leach, A. R. *Molecular modelling: principles and applications*; Addison Wesley: Harlow, 1996.

(4) Frenkel, D.; Smit, B. *Understanding molecular simulation : from algorithms to applications*; Academic Press: San Diego, 2002.

(5) Born, M.; Oppenheimer, R. Zur Quantentheorie der Molekeln. *Annalen der Physik* **1927**, *389*, 457–484.

(6) Thomas, L. H. The calculation of atomic fields. *Mathematical Proceedings of the Cambridge Philosophical Society* **2008**, *23*, 542–548.

(7) Hohenberg, P.; Kohn, W. Inhomogeneous Electron Gas. *Physical Review* **1964**, *136*, B864–B871.

(8) Kohn, W.; Sham, L. J. Self-Consistent Equations Including Exchange and Correlation Effects. *Physical Review* **1965**, *140*, A1133–A1138.

(9) Dirac, P. A. M. Note on Exchange Phenomena in the Thomas Atom. *Mathematical Proceedings of the Cambridge Philosophical Society* **2008**, *26*, 376–385.

(10) Becke, A. D. Density-functional exchange-energy approximation with correct asymptotic behavior. *Physical Review A* **1988**, *38*, 3098–3100.

(11) Becke, A. D. Density-functional thermochemistry. V. Systematic optimization of exchange-correlation functionals. *The Journal of Chemical Physics* **1997**, *107-114*, 8554.

(12) Lee, C.; Yang, W.; Parr, R. G. Development of the Colle-Salvetti correlation-energy formula into a functional of the electron density. *Physical Review B* **1988**, *37*, 785–789.

(13) Becke, A. D. Density-functional thermochemistry. III. The role of exact exchange. *The Journal of Chemical Physics* **1993**, *98*, 5648–5653.

(14) Stephens, P. J.; Devlin, F. J.; Chabalowski, C. F.; Frisch, M. J. Ab Initio Calculation of Vibrational Absorption and Circular Dichroism Spectra Using Density Functional Force Fields. *The Journal of Physical Chemistry* **1994**, *98*, 11623–11627.

(15) Curtiss, L. A.; Raghavachari, K.; Trucks, G. W.; Pople, J. A. Gaussian-2 theory for molecular energies of first- and second-row compounds. *The Journal of Chemical Physics* **1991**, *94*, 7221–7231.

(16) Adamo, C.; Barone, V. Toward reliable density functional methods without adjustable parameters: The PBE0 model. *Journal of Chemical Physics* **1999**, *110*, 6158–6170.

(17) Johnson, B. G.; Gill, P. M. W.; Pople, J. A. The performance of a family of density functional methods. *The Journal of Chemical Physics* **1993**, *98*, 5612–5626.

(18) Parr, R. G.; Yang, W. Density-Functional Theory of the Electronic Structure of Molecules. *Annual Review of Physical Chemistry* **1995**, *46*, 701–728.

(19) Chermette, H. Chemical reactivity indexes in density functional theory. *Journal of Computational Chemistry* **1999**, *20*, 129–154.

(20) Geerlings, P.; De Proft, F.; Langenaeker, W. Conceptual density functional theory. *Chemical Reviews* **2003**, *103*, 1793–1873.

(21) Morell, C. Un nouveau descripteur de la réactivité chimique : étude théorique et applications à la selectivité de quelques réactions chimiques, Université Joseph-Fourier, 2006.

(22) Pearson, R. G. Hard and Soft Acids and Bases. *J. Am. Chem. Soc.* **1963**, *85*, 3533–3539.

(23) Parr, R. G.; Pearson, R. G. Absolute hardness: companion parameter to absolute electronegativity. *J. Am. Chem. Soc.* **1983**, *105*, 7512–7516.

(24) Politzer, P. A relationship between the charge capacity and the hardness of neutral atoms and groups. *The Journal of Chemical Physics* **1987**, *86*, 1072–1073.

(25) Fukui, K.; Yonezawa, T.; Shingu, H. A Molecular Orbital Theory of Reactivity in Aromatic Hydrocarbons. *The Journal of Chemical Physics* **1952**, *20*, 722–725.

(26) Koopmans, T. Über die Zuordnung von Wellenfunktionen und Eigenwerten zu den Einzelnen Elektronen Eines Atoms. *Physica* **1934**, *1*, 104–113.

(27) Parr, R.; Yang, W. Density functional approach to the frontier-electron theory of chemical reactivity. *Journal of the American Chemical Society* **1984**, *106*, 4049–4050.

(28) Chandra, A. K.; Nguyen, M. T. Use of Local Softness for the Interpretation of Reaction Mechanisms. *International Journal of Molecular Sciences* **2002**, *3*, 310–323.

(29) Jones, J. E. On the Determination of Molecular Fields. II. From the Equation of State of a Gas. *Proc. R. Soc. Lond. A* **1924**, *106*, 463–477.

(30) Allinger, N. L.; Yuh, Y. H.; Lii, J. H. Molecular mechanics. The MM3 force field for hydrocarbons. 1. *Journal of the American Chemical Society* **1989**, *111*, 8551–8566.

(31) Brooks, B. R.; Bruccoleri, R. E.; Olafson, B. D.; States, D. J.; Swaminathan, S.; Karplus, M. CHARMM: A program for macromolecular energy, minimization, and dynamics calculations. *Journal of Computational Chemistry* **1983**, *4*, 187–217.

(32) Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Gould, I. R.; Merz, K. M.; Ferguson, D. M.; Spellmeyer, D. C.; Fox, T.; Caldwell, J. W.; Kollman, P. A. A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules. *J. Am. Chem. Soc.* **1995**, *117*, 5179–5197.

(33) Rappe, A. K.; Casewit, C. J.; Colwell, K. S.; Goddard, W. A.; Skiff, W. M. UFF, a full periodic table force field for molecular mechanics and molecular dynamics simulations. *Journal of the American Chemical Society* **1992**, *114*, 10024–10035.

(34) *Scigress*; FUJITSU, 2008.

(35) Shenkin, P. S.; McDonald, D. Q. Cluster analysis of molecular conformations. *J. Comput. Chem.* **1994**, *15*, 899–916.

(36) Kaufman, L. *Finding groups in data: an introduction to cluster analysis*; Wiley: Hoboken N.J., 2005.

(37) Tufféry, S. *Data mining et statistique décisionnelle: l'intelligence des données*; Éd. Technip: Paris, 2012.

(38) Halkidi, M.; Batistakis, Y.; Vazirgiannis, M. On Clustering Validation Techniques. *Journal of Intelligent Information Systems* **2001**, *17*, 107–145.

(39) Kovács, F.; Babos, A.; Legány, C. Cluster Validity Measurement Techniques. *Proc. Sixth Int'l Symp. Hungarian Researchers on Computational Intelligence (CINTI)* **2005**.

(40) Rousseeuw, P. J. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics* **1987**, *20*, 53–65.

(41) Dunn, J. C. A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters. *Journal of Cybernetics* **1973**, *3*, 32–57.

(42) Davies, D. L.; Bouldin, D. W. A Cluster Separation Measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **1979**, *PAMI-1*, 224–227.

(43) Pal, N. R.; Biswas, J. Cluster validation using graph theoretic concepts. *Pattern Recognition* **1997**, *30*, 847–857.

(44) R. Development Core (Nom) *R: A Language and Environment for Statistical Computing*; Vienna, Austria, 2012.

(45) *Scientific Computing Tools For Python — Numpy*; 2011.

(46) Storm, C. B.; Stine, J. R.; Kramer, J. F. *Sensitivity relationships in energetic materials. Chemistry and physics of energetic materials*; Kluwer Academic Publishers, 1990.

CHAPITRE 3 - PRINCIPE ET MÉTHODES DES MODÈLES QSPR

Cette partie présente le principe des modèles « Quantitative Structure-Property/Activity Relationship » (QSPR/QSAR) ainsi que les méthodes d'apprentissage utilisées depuis leur développement jusqu'à leur validation scientifique. De plus, un exemple de développement de modèle QSPR sera présenté en fin de chapitre pour la prédiction de la sensibilité à l'impact des composés nitroaliphatiques.

Les modèles QSAR sont récents[1,2] puisque c'est en 1964 que Hansch et Fujita[3] présentent le paradigme QSAR, suite à l'observation de divers cas. En 1868, 100 ans avant, Brown et Fraser[4] suggèrent pour la première fois que l'activité physiologique d'une substance est fonction de sa composition chimique et de leur constitution : « activity = f(structure) ». Leur étude porte sur des dérivés de la strychnine dont certains possèdent une activité similaire à celle de la curare dans un muscle paralysé. En 1869, Richardson[5] observe une évolution de l'effet narcotique de divers alcools primaires proportionnelle à leur poids moléculaire. En 1893, Richet[6] montre que la toxicité de composés organiques (esters, alcools, cétones) est reliée à l'inverse de leur solubilité dans l'eau. Indépendamment, Meyer[7] (1899) et Overton[8] (1901) découvrent une variation de l'effet anesthésiant avec le logP (coefficient de partage eau/octanol).

En 1937, Hammett[9] étudie la réactivité chimique de benzènes substitués, et aboutit à l'équation suivante pour calculer leur constante d'équilibre K :

$$\log K = logK^0 + \sigma\rho$$

avec σ la constante liée au substituant et ρ la constante de réaction (type de réaction, température).
Dans les années 1950, Taft[10,11] développe une procédure pour séparer les différents effets : polaires, stériques et la résonance. Il développe une équation de Hammett modifiée pouvant s'appliquer aux composés non aromatiques, l'équation de Taft, prenant aussi en compte les effets stériques (δE_S).

$$\log \frac{k_Y}{k_0} = \rho^* \sigma^* + \delta E_S$$

Avec ρ^* et σ^* les constantes de Hammett modifiées correspondant à l'effet de polarisation et à l'effet d'induction du substituant.
D'après les travaux de Hammett et Taft, Hansch[3] et Fujita développent l'équation biologique de Hammett, dite équation de Hansch, qui permet le calcul de l'activité d'une molécule.

$$\log\left(\frac{1}{C}\right) = a(\log P)^2 + b\log P + \rho\sigma + \delta E_S + \text{constante}$$

avec a et b les coefficients, C la concentration, logP l'hydrophobicité et δE_S les effets stériques.

À partir des années 1970, les principes du QSAR ont été développés avec de nouveaux descripteurs, un formalisme mathématique mais aussi des méthodes de validation de plus en plus poussées. Les modèles QSAR/QSPR sont maintenant reconnus comme une méthode de prédiction sérieuse et utilisés dans de nombreux domaines tels que la biologie[12], la toxicologie[13,14] et la pharmacie[14,15]. Les modèles pour les propriétés physico-chimiques se développent de plus en plus, notamment dans le cadre de l'enregistrement des substances chimiques pour REACH[16]. Cependant, ils ne sont pas encore généralisés et un manque existe pour les propriétés explosives ou encore la stabilité thermique[16,17].

I. PRINCIPE

Les modèles QSPR/QSAR relient une propriété (ou activité biologique) à la structure d'une molécule en proposant une relation mathématique plus ou moins complexe. Les méthodes QSPR[18] reposent sur le principe suivant : des molécules ayant les mêmes propriétés sont proches dans l'espace chimique. L'espace chimique (Figure 12) est un espace à n dimensions qui correspondent à des variables décrivant la structure moléculaire. Ce sont ces variables, appelées descripteurs, qui vont être reliées à la propriété étudiée par l'intermédiaire de différentes méthodes statistiques. La propriété (Y) est exprimée comme une fonction de la structure.

(3. 46) $$Y = f\left(structure\right) = f\left(descripteurs\right)$$

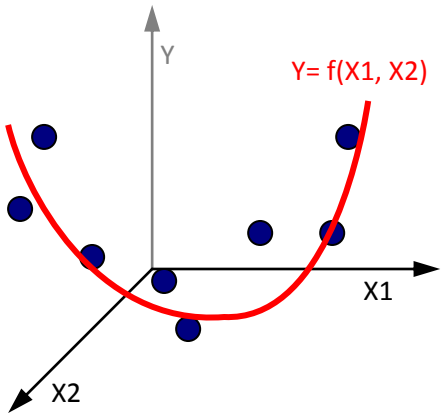


Figure 12: Espace chimique à 2 variables (X1, X2)

À première vue, le développement d'un modèle pourrait être schématisé comme dans la Figure 13. En partant de la structure, des descripteurs sont calculés puis utilisés pour développer un modèle permettant de prédire la propriété ciblée.

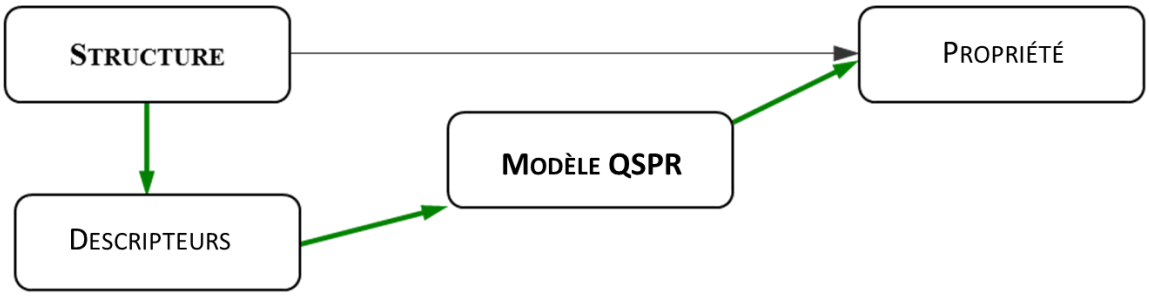


Figure 13: Schéma simplifié du développement d'un modèle QSAR/QSPR

II. BASE DE DONNÉES

Une partie préliminaire primordiale pour le développement ainsi que pour la validation des modèles QSPR est la sélection de la base de données. En effet, sans valeurs expérimentales fiables et homogènes, il est difficile, voire impossible, d'obtenir un modèle prédictif. La qualité des modèles QSPR dépend fortement de la qualité des données expérimentales car les erreurs expérimentales (qui ne sont pas toujours connues) se propagent lors du développement des modèles. L'homogénéité des données étant importante[19], les données doivent être mesurées, autant que possible, par le même protocole dans les mêmes conditions expérimentales. La taille de la base de données est aussi importante : une base de données trop petite rend difficile l'obtention de modèles prédictifs et parfois impossible leur validation (pas de molécule disponible). Par exemple, un modèle développé sur une base de données de 10 molécules n'a pas de sens et ne peut évidemment pas être validé par un jeu de molécules n'intervenant pas dans le développement.

Il faut aussi veiller aux erreurs dans les structures des molécules, qui sont la base des modèles QSPR/QSAR, dont la bonne représentation est cruciale. Il est indispensable de ne pas avoir d'erreur à ce niveau. Il a été observé[20,21] que les bases de données publiques ou privées ont des taux d'erreurs (entre 0,1 et 3,4%) qui semblent peu significatifs mais qui influent de manière non négligeable sur les prédictions. Il faut donc être très attentif aux structures et corriger les éventuelles erreurs[22] comme : les incohérences entre numéro CAS et nom de la molécule, les mélanges stéréo-isomères, la présence de sels ou celle de valeurs de propriétés/activités différentes pour un même composé. Une seule erreur à ce niveau peut imposer de redévelopper complètement les modèles et avoir une grande influence sur leur prédictivité si l'erreur n'est pas repérée. Il faut aussi vérifier l'absence de redondance[19] : plusieurs numéros CAS pour une seule molécule (par exemple la *cyanoguanidine*), même molécule mais sous un nom différent[23] (comme la *2-butanone peroxyde* aussi appelée *Méthyle éthyle cétone peroxyde* ou *Dioxydi-2,2-butanediyle dihydroperoxyde*) et de molécules mal définies (nom incorrect ou ambigu comme *tert-butyl peroxyneodecanoate* ou encore règles de valence non respectées). Les données aberrantes (outliers), molécules ayant une structure trop différente des autres molécules de la base de données ou une valeur de propriété complètement différente, doivent être supprimées afin de ne pas perturber le développement des modèles. La base de données doit donc être épurée avant de pouvoir être utilisée. Il est à noter que ce type d'erreurs aura d'autant plus d'importance que la base de données est petite.

Les bases de données relient souvent le nom de la molécule uniquement à une valeur qualitative ou quantitative de la propriété étudiée (la structure ou le numéro CAS ne sont pas toujours donnés). Les noms ne respectant pas toujours la nomenclature IUPAC (*International Union of Pure and Applied*

Chemistry)[24], il est parfois compliqué d'identifier les structures. Le numéro CAS peut être une bonne piste mais cela nécessite cependant d'y avoir accès.

Cette recherche de valeurs expérimentales homogènes et de structures correctes fait partie intégrante du développement d'un modèle et peut être longue et fastidieuse. Il faut trouver, *a minima*, la structure 1D des molécules de la base de données expérimentale pour pouvoir envisager de développer un modèle.

III. REPRÉSENTATION DES STRUCTURES

Il existe différents niveaux de représentation de la structure des molécules : la nomenclature avec le nom uniquement, la formule brute, la structure 2D et la structure 3D.

Le niveau le plus simple de la représentation des structures est le nom de la molécule ainsi que la formule brute. Les informations disponibles et les descripteurs pouvant être calculés à partir de ces données sont assez limités. Mais il est difficile de stocker les structures 2D et 3D de plusieurs molécules pour construire une base de données informatique simple. C'est pour cela que le langage SMILES[25] (*Simplified Molecular Input Line Specification*) a été développé à la fin des années 1980 par the *Environmental Research Laboratory-Duluth QSAR research program*[26]. À partir d'une simple ligne de caractères, les structures 2D et 3D peuvent être retrouvées facilement. Cependant, cette notation présente quelques désavantages : elle n'est pas unique, il existe plusieurs façons de créer un SMILES pour une même molécule, notamment au niveau de l'aromaticité des liaisons. Le problème d'unicité peut être pallié en utilisant des algorithmes basés sur la théorie des graphes moléculaires comme CANGEN[18,27] qui permet l'obtention de SMILES unique. Selon le type de descripteurs à calculer, la structure nécessaire change. En effet, le calcul du nombre d'atomes de la molécule nécessite uniquement la formule brute alors que l'obtention de la surface accessible au solvant nécessite la structure 3D. L'obtention de la « bonne » structure 3D, c'est-à-dire la plus stable, nécessite une analyse conformationnelle.

Dans cette thèse, pour obtenir la structure 3D à partir de la structure 2D, le logiciel Scigress[28] a été utilisé. Il permet d'optimiser rapidement la structure à partir d'un champ de force MM3[29] et surtout une analyse conformationnelle peut être faite. La méthode CONFLEX[30,31] a été utilisée pour obtenir la liste des minima et leur énergie. Il s'agit d'identifier les conformations stables par rotation autour des différentes liaisons de la molécule et d'optimiser la structure des conformations obtenues. La géométrie de la conformation la plus stable obtenue par Scigress pour chaque molécule a été retenue dans cette analyse. Le logiciel Gaussian09[32] a ensuite été utilisé pour optimiser la géométrie des molécules. La structure obtenue est optimisée au niveau DFT (base 6-31G+(d,p), fonctionnelle PBE0[33]). La Figure 14 résume la procédure suivie.

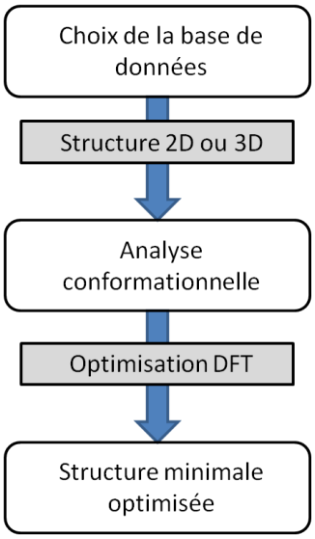


Figure 14: De la base de données à la structure utilisée

IV. DESCRIPTEURS

1. Définition

À partir des structures calculées précédemment, des variables les représentant, appelées descripteurs, peuvent être obtenues. L'une des définitions possible des descripteurs est donnée ci-dessous par Todeschini et Consonni dans le "Handbook of Molecular Descriptor"[34] : "*The molecular descriptor is the final result of a logic and mathematical procedure which transforms chemical information encoded within a symbolic representation of a molecule into a useful number or the result of some standardized experiment*". Le mot "useful" est à considérer ici avec un double sens : le descripteur aide à l'interprétation des propriétés moléculaires et/ou il est capable de participer à l'amélioration du pouvoir prédictif. En effet, il est fondamental de prédire mais aussi de comprendre les faits expérimentaux d'un point de vue chimique à partir d'une représentation des molécules par les descripteurs. Ces descripteurs peuvent être obtenus par l'observation des structures, l'utilisation d'algorithmes avec des logiciels tels que Dragon[35], Codessa[36], ISIDA[37] (In SIlico design and Data Analysis) ou être des propriétés physico-chimiques expérimentales.

2. Classement par type

Il existe plusieurs types de descripteurs moléculaires et plusieurs façons de les classer. Ils peuvent être classés, par exemple, selon la dimension de la structure nécessaire pour son calcul (1D, 2D, 3D), le temps de calcul nécessaire pour leur obtention ou encore en fonction du type d'informations fournies par le descripteur.

Dans ce manuscrit, la nomenclature utilisée par le logiciel Codessa est respectée car il a été utilisé pour le calcul des descripteurs. Voici donc le classement choisi par type de descripteurs :

- Constitutionnels : descripteurs basés sur la formule brute d'une molécule (nombre d'atomes, masse moléculaire). Les descripteurs correspondant à de simples comptages, mais qui

nécessitent la formule semi développée, comme le nombre de groupes fonctionnels, le nombre de liaisons (simples, doubles, triples) sont aussi considérés dans ce groupe.

- Topologiques : basés sur le graphe moléculaire c'est-à-dire la matrice de connectivité des atomes, ces descripteurs sont des indices topologiques tels que les indices de Wiener[38], Randić[39] donnant notamment des informations sur la taille et la forme des molécules.
- Géométriques : basés sur la structure 3D de la molécule, ces descripteurs peuvent être le volume moléculaire, la surface moléculaire (surface accessible au solvant) ou encore la distance interatomique.
- Electrostatiques : aussi basés sur la structure 3D, ces descripteurs nécessitent également le calcul des charges des atomes. Ces descripteurs sont, par exemple, la charge d'atomes, le potentiel électrostatique de la molécule. Parmi ces descripteurs, les descripteurs CPSA (*Charged Partial Surface Area*[40]) font l'objet d'une classe à part.
- Quantiques : basés sur des données obtenues par calculs quantiques, ces descripteurs sont par exemple les énergies électroniques, l'énergie orbitalaire (HOMO/LUMO), des descripteurs de réactivité (énergie de dissociation).
- Thermodynamiques : liés à la thermodynamique de la molécule, il s'agit de descripteurs tels que l'enthalpie et l'entropie de la molécule

Plus de 6000 descripteurs ont été recensés par Todeschini[34]. Parmi eux d'autres types de descripteurs comme les propriétés moléculaires ou encore les descripteurs WHIM[41] (*Weighted Holistic Invariant Molecular*, basés la projection des atomes sur les composantes principales) sont présents.

Dans cette thèse, des descripteurs supplémentaires à ceux proposés par Codessa ont aussi été calculés et utilisés en fonction des composés et de la propriété étudiés. Ces derniers ont été extraits de la littérature, obtenus par l'observation des structures de la base de données ou encore calculés par DFT (la DFT conceptuelle[42,43] notamment abordée dans le chapitre 2).

3. Sélection des descripteurs

Ainsi, plus d'un millier de descripteurs peuvent être obtenus mais tous ne sont pas nécessaires au développement du modèle. Des méthodes de sélection de variables sont disponibles pour réduire ce nombre, notamment afin de ne pas obtenir des équations sur-paramétrées. De manière générale, la réduction des descripteurs commence par la suppression des données redondantes c'est-à-dire très corrélées entre elles. De plus, les descripteurs considérés comme pertinents sont ceux ayant une grande corrélation avec la propriété et ayant une variance significative sans laquelle le descripteur ne permet pas la distinction des différentes données entre elles. Il existe un grand nombre de méthodes

de sélection automatique de variables[44–55] plus ou moins rapides et faciles à programmer : méthodes pas à pas (dite stepwise), PLS[56] (régression des moindres carrés partiels), algorithme génétique.

Dans cette thèse, nous avons utilisé la *Best Multilinear regression* (BMLR)[57] implémentée dans le logiciel Codessa[36] qui est une approche point par point basée sur la présélection des descripteurs pour les régressions multilinéaires. Pour commencer, les descripteurs ayant des valeurs manquantes ou une valeur constante sur l'ensemble des données sont écartés. Le logiciel calcule ensuite, pour chaque descripteur, sa corrélation avec la propriété et ceux qui ont le moins de corrélation avec celle-ci sont alors écartés (R^2 inférieur à 0,1 par défaut). Pour des descripteurs inter-corrélés (R^2 supérieur à 0,6 par défaut), seul celui ayant la plus grande corrélation avec la propriété est gardé. Ainsi, des descripteurs inter-corrélés, c'est-à-dire décrivant la même information, ne peuvent pas être présents dans un même modèle. Ces seuils sont des paramètres par défaut qui peuvent être modifiés si nécessaire. Les corrélations à deux paramètres sont ensuite calculées pour tous les couples de descripteurs non corrélés entre eux. Les couples de descripteurs ayant la plus grande corrélation avec la propriété expérimentale sont gardés. A partir de 3 descripteurs, tant que la valeur de la corrélation augmente significativement (ΔR^2>0,001 par défaut), des descripteurs non corrélés au couple sélectionné lui sont ajoutés un à un. La BMLR sélection les descripteurs en se basant sur les performances des modèles associés. Au final, les meilleures régressions choisies pour chaque rang (nombre de descripteurs) sont proposées. Il s'agit ensuite de choisir parmi elles la régression présentant le meilleur compromis entre les performances (par rapport aux coefficients R^2 mesurant l'ajustement et Q^2 mesurant la robustesse – défini dans le paragraphe VI.1 « Validation croisée ») et le nombre de descripteurs (voir Figure 15).

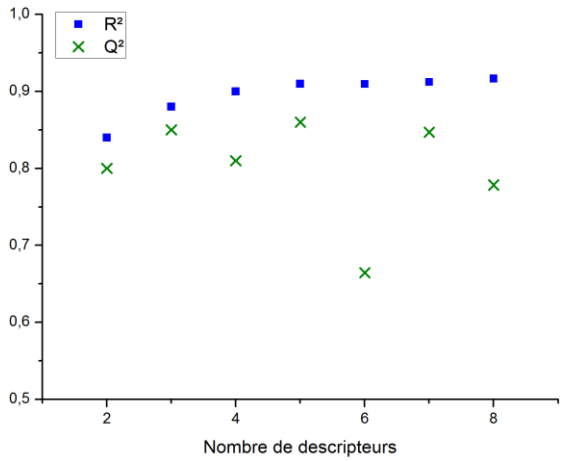


Figure 15: Sélection du nombre de descripteurs du modèle en fonction des performances (R^2 et Q^2)

V. DÉVELOPPEMENT DE MODÈLES

Contrairement à ce que la Figure 13 pourrait laisser penser, le développement d'un modèle QSPR est complexe. La Figure 16 représente un schéma de la mise en place d'un modèle QSPR. La recherche et l'optimisation des structures ne sont qu'une étape de la première partie du schéma. La seconde étape constitue le calcul des descripteurs. La question suivante se pose alors : comment ces « useful number » vont-ils permettre d'obtenir un modèle prédictif ? La deuxième partie du schéma répond à cette question à l'aide de la case « entraînement, domaine d'applicabilité, validation ». En effet, ce sont ces trois étapes qui vont permettre l'obtention d'un modèle prédictif. La première étape de cette deuxième partie est l'entraînement : il s'agit de développer une équation, un arbre de décision, un réseau de neurones permettant de calculer la propriété des données utilisées.

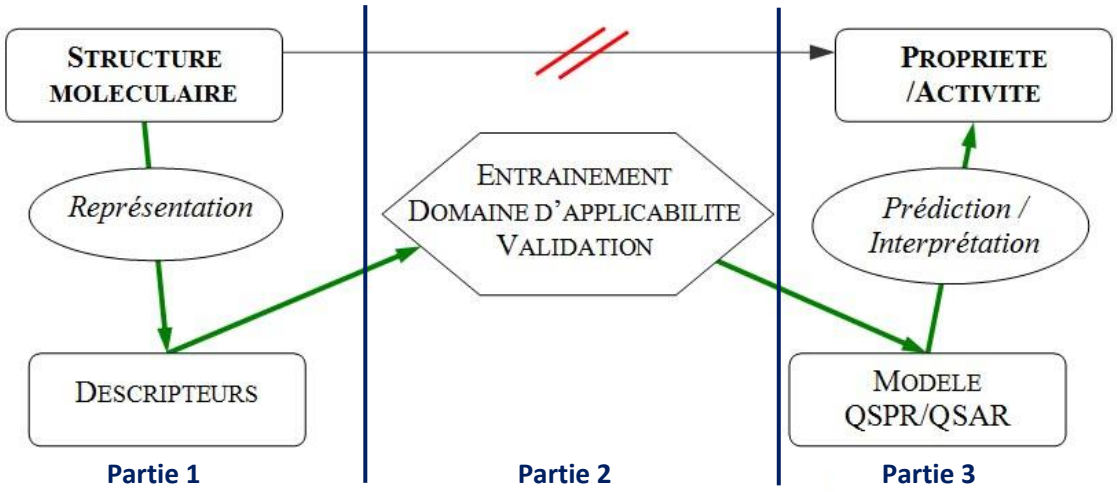


Figure 16 : Schéma détaillé de la mise en place d'un modèle QSPR/QSAR

1. Jeux d'entraînement et de validation

Le développement d'un modèle prédictif et validé nécessite le partage de la base de données expérimentale en deux jeux distincts. Le jeu d'entraînement va permettre la construction d'un modèle tandis que le jeu de validation sert à calculer la prédictivité du modèle. Ce partage en deux jeux et leur utilité est illustré par la Figure 17. Peduzzi[58] et Babyak[59] estiment que pour obtenir des estimations correctes, le nombre minimal de molécules par descripteur est compris entre 10 et 15. En effet, lorsque le nombre de descripteurs du modèle est élevé par rapport au nombre de molécules on se trouve souvent face à un problème dit de sur-paramétrisation (ou *overfitting*[60]) : le nombre de descripteurs choisi ne correspond pas au nombre minimal de descripteurs nécessaires (principe de parcimonie). La sur-paramétrisation lors du développement de modèle entraîne généralement des difficultés en termes de prédictivité puisque le modèle n'est applicable qu'aux molécules utilisées pour l'entraînement. Il n'y a pas de règle concernant la taille optimale du jeu de validation mais un minimum de 10 molécules semble être accepté[61].

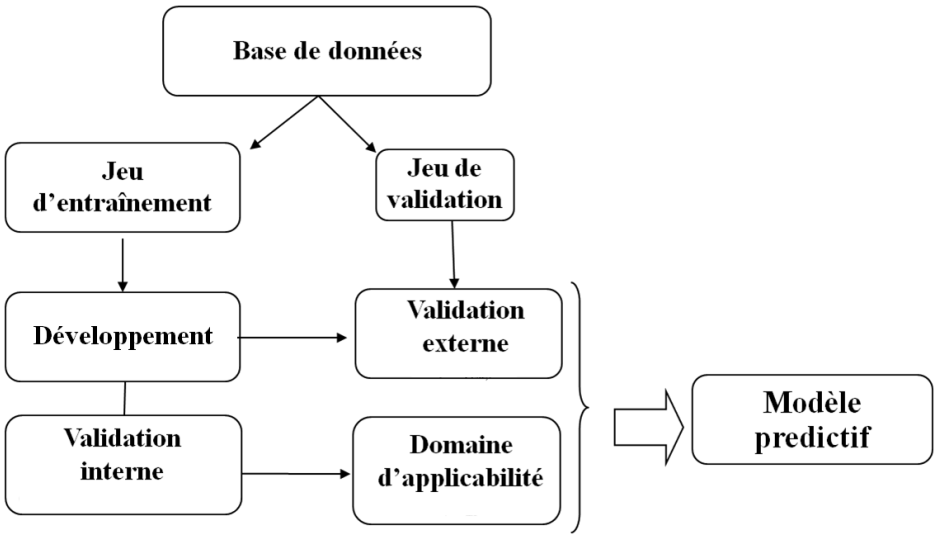


Figure 17 : Partage des données expérimentales pour le développement d'un modèle

a) Partage des données en deux jeux

Les deux jeux, représentatifs de l'ensemble de la base de données, doivent avoir la même distribution au niveau de la propriété et de l'espace des descripteurs (c'est-à-dire que les structures ne doivent pas être trop dissimilaires). Pour cela plusieurs méthodes sont disponibles pour répartir les données telles que : la sélection basée sur la valeur de la propriété[62,63], la division aléatoire[62,64], le clustering[62,65,66], les sphères d'exclusion[64] ou encore l'algorithme génétique[48].

Dans cette thèse, la répartition des données basée sur la valeur de la propriété a été choisie dans un premier temps. En raison du faible nombre de données expérimentales, 1/3 des données sont dans le jeu de validation, afin que celui-ci ait un nombre de molécules suffisamment élevé pour réaliser une validation externe correcte.

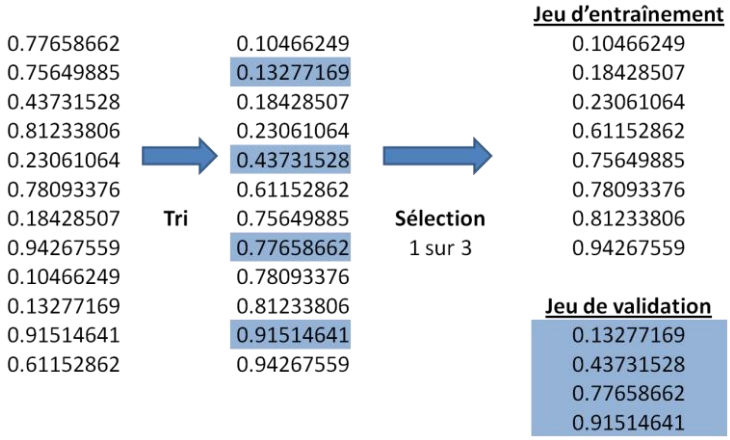


Figure 18: Illustration de la méthode de sélection des données utilisée

En pratique, les molécules ont été classées par ordre croissant de valeur de la propriété puis une molécule sur trois a été sélectionnée pour constituer le jeu de validation comme illustré sur la Figure 18. La première (et si possible la dernière) molécule de la base de données sont attribuées au jeu d'entraînement afin de permettre un domaine d'applicabilité plus grand. Dans le cas d'un partage

1/3-2/3 sur la valeur de la propriété, la 2ème ou 3ème molécule de la base de données est la première du jeu de validation. Le choix des molécules du jeu de validation a été réalisé de manière à garantir une distribution similaire dans les deux jeux[67].

b) Représentation des données par l'analyse en composantes principales

Une façon efficace de vérifier la bonne répartition des molécules dans les deux jeux de données dans l'espace des descripteurs est de les représenter dans celui-ci en utilisant les composantes principales. L'analyse en composantes principales[18] (ACP ou PCA) est une méthode puissante permettant de diminuer le nombre de variables décrivant un jeu de données en minimisant la perte d'information. Il s'agit d'une transformation dont les nouvelles variables obtenues par combinaison linéaire des variables initiales sont appelées *composantes principales ou PC*. Il faut d'abord calculer la matrice de covariance C des i données x_j, j=1, …, i centrées en zéro avec l'équation (3. 47) où X est la matrice de ces données.

(3. 47)
$$C = \frac{1}{\mathrm{i}} \sum_{j=1}^{\mathrm{i}} X_j X_j^T$$

Il s'agit ensuite de résoudre un problème aux valeurs (λ) et vecteurs propres (v) avec l'équation :

(3. 48)
$$\lambda v = Cv$$

La projection des anciennes coordonnées sur les vecteurs propres donne les nouvelles coordonnées ou composantes principales. Les composantes principales sont orthogonales et non corrélées entre elles comme illustré Figure 19 . La première composante principale (PC1, associée à la valeur propre la plus élevée) contient la plus grande quantité de variation et décrit le mieux les données. Tandis que la dernière (ième) composante est celle qui décrit les détails.

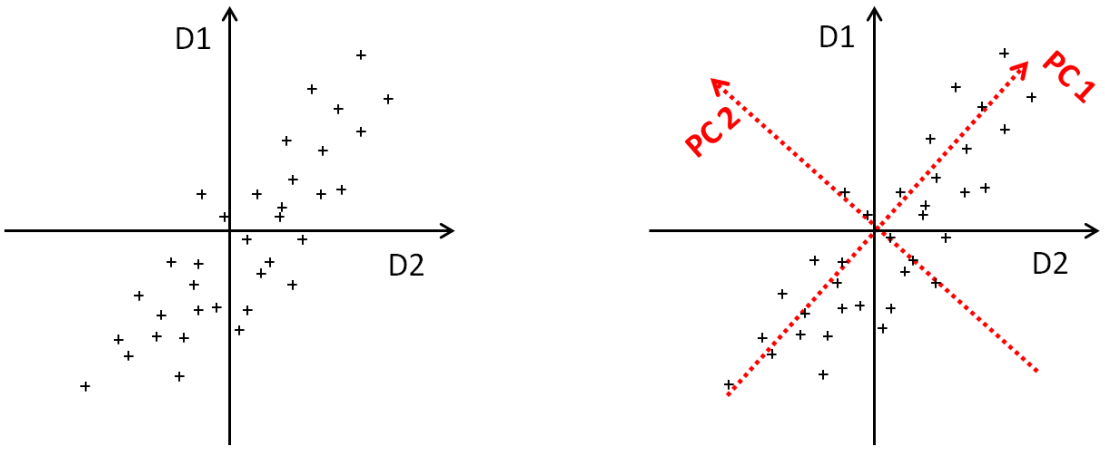


Figure 19: Représentation géométrique des composantes principales : D1 et D2 sont les variables, PC1 et PC2 sont la 1ere et la 2nde composantes principales (droite rouge en pointillé)

Il s'agit ensuite de sélectionner les molécules dans cet espace de manière à couvrir autant que possible la surface avec les molécules du jeu d'entraînement et celles du jeu de validation.

2. Méthodes d'entraînement des données

Pour relier la structure des molécules à la propriété expérimentale, différentes méthodes sont utilisées, allant de la régression multilinéaire aux cartes de Kohonen[18,68] en passant par les arbres de décisions[69]. Dans cette partie, la régression multilinéaire utilisée au cours de cette thèse sera détaillée.

La régression multilinéaire (MLR) part de l'hypothèse selon laquelle il existe une relation linéaire entre une variable Y (la propriété) et plusieurs variables x (les descripteurs). La MLR est une généralisation à plusieurs variables de la régression linéaire qui permet de déterminer pour une série de points, l'équation d'une droite passant le plus près de l'ensemble des points.

Le principe de la régression multilinéaire consiste donc à exprimer une variable Y comme une combinaison linéaire de N variables x_i avec une équation de type (3. 49).

(3. 49)
$$Y = a_0 + \sum_{i}^{N} a_i x_i$$

Pour déterminer la valeur des coefficients a_i, la méthode des moindres carrés est utilisée. Elle a pour but de minimiser le carré des résidus ou encore RSS (*Residual Sum of Squared*) représenté sur la Figure 20 c'est-à-dire la somme des carrées des écarts entre les valeurs prédites et les valeurs réelles sur toute la base de données de p molécules.

(3. 50)
$$RSS = \sum_{i=1}^{p} (y_{\exp,i} - y_{calc,i})^2$$

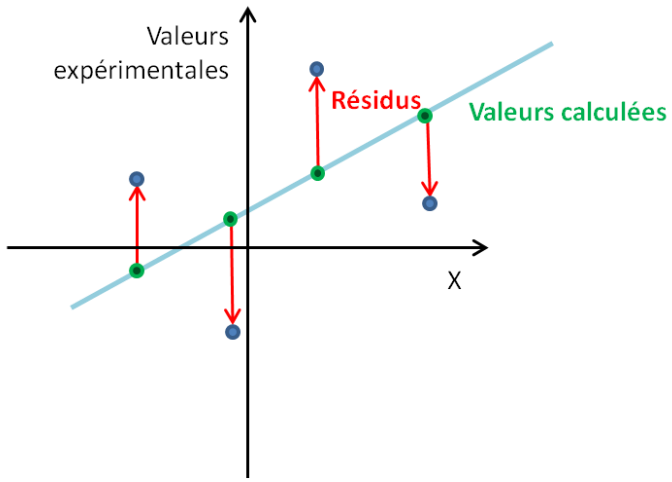


Figure 20: Représentation graphique des résidus

En pratique, il s'agit de résoudre un système à p équations (correspondant au nombre de molécules) pour N variables (nombre de descripteurs) avec N<p en minimisant le RSS. Ce système peut être résolu en utilisant une notation matricielle :

(3. 51)
$$A = (X^T X)^{-1} X^T Y$$

Où A est la matrice des coefficients a_i, X celle des variables x_i et Y le vecteur contenant les valeurs de la propriété.

3. Mesure de l'ajustement

Pour évaluer les performances d'ajustement des modèles, différents coefficients ont été calculés. La racine carrée de l'erreur quadratique (ou *root mean square error* i.e. RMSE) entre les valeurs des données prédites et expérimentales se calcule avec la formule suivante :

(3. 52)
$$RMSE = \sqrt{\frac{RSS}{n-p-1}}$$

Où RSS est le carré des résidus, n le nombre de données, p le nombre de paramètres (i.e. de descripteurs dans l'équation).

La corrélation peut aussi être mesurée par le coefficient de détermination R^2 :

(3. 53)
$$R^2 = 1 - \frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

Où y_i est la valeur expérimentale de la propriété, $\hat{y}_i$ est celle prédite pour propriété, $\bar{y}$ est la moyenne des valeurs expérimentales et n le nombre de molécules.

L'erreur moyenne absolue (*Mean absolute error* i.e. MAE), pouvant également être donnée en pourcentage, est une autre mesure de l'ajustement des valeurs des données calculées avec celles des données expérimentales.

(3. 54)
$$MAE = \frac{\sum_{i=1}^{n}|y_i - \hat{y}_i|}{n}$$

En plus de ces coefficients, le coefficient du test de Student (t-test) est calculé avec un niveau de confiance de 95% sur les descripteurs obtenus dans l'équation afin de valider leur pertinence. Le descripteur est considéré comme significatif si la valeur t est supérieure à celle tabulée à 95% pour un nombre de degrés de liberté (n-p-1). Le descripteur ayant la valeur absolue de t-test la plus élevée est le plus pertinent.

(3. 55)
$$t = \frac{b_i}{s_{b_i}}$$

Où b_i est le coefficient du descripteur dans l'équation et s_{b_i} la déviation standard du descripteur.

L'importance statistique de la régression peut être évaluée au moyen du test de Fisher (valeur F). Plus la valeur de F est grande, plus la probabilité que l'équation soit pertinente augmente. L'équation est considérée comme significative si la valeur F est supérieure à celle tabulée à 95% pour un nombre de degrés de liberté (n-p-1).

$$F = \frac{ESS/p}{RSS/(n-p-1)} \quad \textbf{(3. 56)}$$

Où n est le nombre de molécules, p le nombre de descripteurs et ESS est la somme des carrés des résidus due à la régression calculée par l'équation suivante.

$$ESS = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 \quad \textbf{(3. 57)}$$

VI. VALIDATION

La validation des modèles est une partie intégrante et importante du développement d'un modèle. L'OCDE a mis en place des principes[70] pour une validation scientifique et réglementaire, déjà évoqués dans le chapitre 1 « contexte et objectifs ». La validation sert à démontrer que le modèle est prédictif et que les bonnes performances mesurées jusque là ne sont pas dues au sur-apprentissage ou à la chance. Les modèles sont créés sur une partie de la base de données, appelée jeu d'entraînement, comme illustré en Figure 17. Mais avant de pouvoir envisager d'utiliser un modèle, il faut le valider. Il existe plusieurs types de validation qui se complètent : la validation dite interne qui utilise le jeu d'entraînement et la validation externe réalisée sur le jeu de validation. Les différentes méthodes de validation seront décrites dans cette partie.

1. Validation croisée

La validation interne ou validation croisée mesure la robustesse du modèle c'est-à-dire sa capacité à rester corrélé à la propriété quand on modifie légèrement les données (suppression d'une ou plusieurs données). Il existe plusieurs méthodes de validation croisée : LOO (*Leave One Out*) et LMO (*Leave Many Out*) ou n-fold validation[71]. Il s'agit de modifier le jeu de validation en le séparant à son tour en deux groupes (voir Figure 21). Le modèle est ajusté sur le plus grand groupe (en bleu foncé), à partir des descripteurs auparavant sélectionnés sur l'ensemble du jeu d'entraînement (en bleu clair). La nouvelle équation est appliquée sur le second groupe (en gris) pour obtenir la prédiction associée à cet ensemble de molécule. Cette opération est effectuée N fois afin d'obtenir les prédictions de tout le jeu d'entraînement.

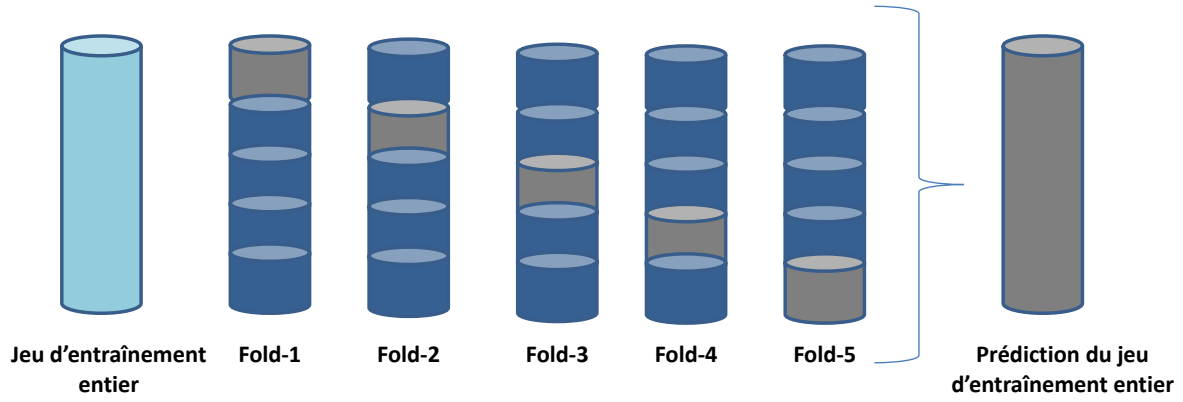


Figure 21: Validation croisée pour 5 folds - Prédiction des données du jeu d'entraînement

Dans le cas du Leave One Out (LOO), une seule molécule du jeu d'entraînement est retirée et les coefficients de la régression sont optimisés sur les N-1 autres données. La propriété Y_{pred} est recalculée à partir de cette nouvelle équation pour la molécule isolée. Cette manipulation est effectuée pour les N molécules du jeu d'entraînement, puis le coefficient de corrélation de validation croisée noté Q^2 est calculé avec l'équation suivante :

(3. 58)

$$Q^2_{LOO} = 1 - \frac{\sum_{i=1}^{n} (\hat{y}_{i/i} - y_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}$$

Où $\hat{y}_{i/i}$ est la valeur prédite en excluant le $i^{ème}$ élément dans le développement du modèle.

Dans le cas du Leave Many Out (LMO), un groupe de molécules du jeu d'entraînement est retiré au lieu d'une seule molécule. Une faible valeur de Q^2 implique que le modèle n'est pas robuste et ne sera pas prédictif, mais la réciproque n'est pas nécessairement vraie[72]. En effet, le modèle est considéré comme robuste quand les différents coefficients de validation Q^2 (voir dans le paragraphe VIII) ont des valeurs très proches et quand la différence entre les Q^2 et le R^2 est faible (comme on peut le voir dans la Figure 15).

Le bootstrap[73–75] est une autre méthode de validation interne mais qui n'a pas été utilisée au cours de cette thèse. Le bootstrapping fait parti des méthodes conseillées par le JRC[76] pour l'enregistrement des modèles (fichier QMRF[77]). Il s'agit d'une technique de rééchantillonnage avec « remise » d'un grand nombre d'itérations (de 50 à 2000) qui permet d'évaluer l'intervalle de confiance et des estimateurs statistiques tels que la variance. Il existe beaucoup de méthodes plus ou moins sophistiquées.

2. Corrélation par chance : Y-scrambling

Le Y-scrambling permet de déterminer si la corrélation obtenue n'est pas due à la chance. Deux cas sont illustrés dans l'article de Hutter[78] : l'article de Sies[79] qui présente une corrélation de R^2=0,99 entre le nombre de cigognes et le nombre de nouveau-nés, mais avec seulement 7 données, et celui de Johnson[80] qui remarque une corrélation étrange entre la quantité de citrons importée et la mortalité routière (R^2=0,97 avec 5 données). En effet, dans le cas où le nombre de variables est grand et la base de données petite, le risque de corrélation par accident est important. Un exemple illustrant l'importance de cette procédure de validation est l'étude de Tropsha[81] à partir les modèles développé par Wilcox[82] avec CoMFA[83] (*Comparative Molecular Field Analysis*).

Afin d'évaluer la part de chance dans les modèles, les valeurs de la propriété sont échangées entre les molécules afin de voir si une corrélation significative est obtenue avec des valeurs de propriété « fausses ». La première étape consiste à mélanger les valeurs de la propriété Y entre elles sans changer la valeur des descripteurs X_i comme illustré en Figure 22. Il suffit ensuite de redévelopper le modèle avec les propriétés « fausses ». Ce mélange se fait entre 500 et 1000 fois généralement[44]. La signification du terme « redévelopper » dans ce contexte n'est pas toujours claire[84–86]. En effet, la sélection des descripteurs doit-elle être incluse dans ce redéveloppement ? Deux méthodes semblent exister : celle qui considère la sélection des descripteurs[19,87] et celle qui ne le fait pas[55]. Dans cette thèse, pour une raison pratique (comme la difficulté à automatiser la sélection des descripteurs) la sélection des descripteurs n'a pas été prise en compte lors du Y-scrambling.

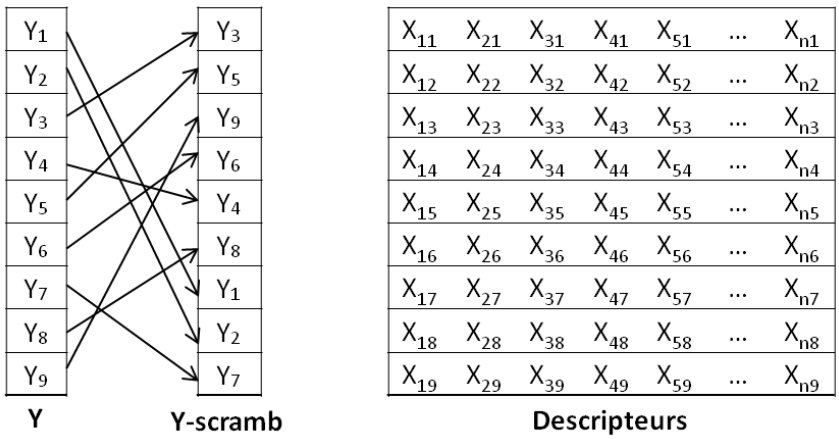


Figure 22: Illustration de la méthode « Y-scrambling »

Les corrélations entre le vecteur contenant les données expérimentales et celui contenant les nouvelles après « mélange » peuvent être calculées afin d'obtenir le graphique en Figure 23. Moins les Y_i sont corrélés entre eux, moins la valeur de R^2 doit être grande. Un bon modèle doit se distinguer sur la figure et aucun des R^2 obtenus avec des données mélangées ne doit être supérieur au R^2 du modèle développé. Les modèles fortuits doivent avoir des performances nettement

inférieures à celles du modèle initial. Selon Rücker[88], pour que la probabilité que le modèle ne soit pas dû au hasard soit de 1%, l'équation (3. 59) doit être vraie.

(3. 59) $$R^2 - R^2_{YS} > 2{,}3\sigma_{YS}$$

Avec R^2_{YS} la moyenne des R² des modèles fortuits et σ_{YS} l'écart type.

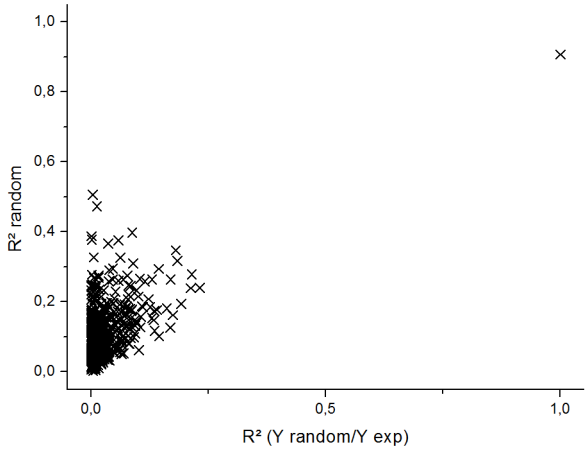


Figure 23: Coefficient de détermination R² du modèle obtenu avec les nouveaux Y (R²random) en fonction de la corrélation entre ces nouveaux Y et les Y expérimentaux ($R^2(Y_{random}/Y_{exp})$)

Cependant, en pratique ce critère semble souvent respecté, même pour les modèles avec de faibles performances (comme eux du chapitre 4).

3. Validation externe ou prédictivité

La mesure de la prédictivité la plus utilisée est le R^2_{ext}, le coefficient de détermination associé à la prédiction de la propriété pour les données du jeu de validation qui n'ont pas été utilisées pour le développement du modèle.

(3. 60) $$R^2_{ext} = \frac{\sum_{i=1}^{N}\left(y_{calc,i} - \langle y_i \rangle\right)^2}{\sum_{i=1}^{N}\left(y_i - \langle y_i \rangle\right)^2} = 1 - \frac{\sum_{i=1}^{N}\left(y_i - y_{calc,i}\right)^2}{\sum_{i=1}^{N}\left(y_i - \langle y_i \rangle\right)^2} \quad \textbf{avec} \quad \langle y_i \rangle = \frac{1}{N}\sum_{i=1}^{N} y_i$$

De même, tous les autres évaluateurs de l'ajustement (MAE et RMSE) des données pour le jeu d'entraînement peuvent être calculés pour le jeu de validation afin d'évaluer la prédictivité. L'article de Chirico et Gramatica[89] répertorie différents coefficients de calcul de la prédictivité présentés ci-après.

Le coefficient Q^2_{F1} proposé par Tropsha[65,72,81] n'est pas une vraie validation externe car il fait intervenir des informations sur le jeu d'entraînement.

(3. 61)
$$Q_{F1}{}^2 = 1 - \frac{\sum_{i=1}^{n_{ext}} (\hat{y}_i - y_i)^2}{\sum_{i=1}^{n_{ext}} (y_i - \overline{y}_{TR})^2}$$

Avec y_i la valeur expérimentale de la propriété, $\hat{y}_i$ la valeur prédite/calculée de la propriété et $\overline{y}_{TR}$ la moyenne des valeurs y_i du jeu d'entraînement.

En 2008, une autre mesure de la prédictivité, proposée par Schüürmann[85], est le Q^2_{F2} qui se différencie de Q^2_{F1} par le fait que la moyenne utilisée au dénominateur est celle du jeu de validation et non celle du jeu d'entraînement : il s'agit donc bien d'une validation externe car aucune donnée du jeu d'entraînement n'est nécessaire. De plus, Q^2_{F1} est plus optimiste[89] car supérieur ou égal à Q^2_{F2} et par conséquent accepte plus facilement les modèles. Le risque d'avoir un modèle non prédictif accepté est moins grand avec Q^2_{F2}.

(3. 62)
$$Q_{F2}{}^2 = 1 - \frac{\sum_{i=1}^{n_{ext}} (\hat{y}_i - y_i)^2}{\sum_{i=1}^{n_{ext}} (y_i - \overline{y}_{EXT})^2}$$

Avec $\overline{y}_{EXT}$ la moyenne des valeurs y_i du jeu de validation

En 2009, Le coefficient Q^2_{F3} a été proposé par Consonni[90] afin de supprimer le biais introduit par la distribution des données. De plus, selon Consonni, l'absence d'information sur le jeu d'entraînement est un désavantage[90]. En effet, il a été observé que la valeur de Q^2_{F3} est identique quel que soit la distribution du jeu de validation. Il semble également être insensible au nombre de molécules. En effet, la valeur de Q^2_{F3} ne change pas avec la taille du jeu de validation, contrairement à Q^2_{F2} dont la valeur augmente avec le nombre de molécules. Cependant, tout comme Q^2_{F1}, ce coefficient n'est pas une vraie validation externe car il fait intervenir des informations sur le jeu d'entraînement.

(3. 63)
$$Q^2_{F3} = 1 - \frac{\left[\sum_{i=1}^{n_{EXT}} (\hat{y}_i - y_i)^2\right] / n_{EXT}}{\left[\sum_{i=1}^{n_{TR}} (y_i - \overline{y}_{TR})^2\right] / n_{TR}}$$

Avec n_{TR} le nombre de molécules du jeu d'entraînement et n_{ext} le nombre de molécules dans le jeu de validation.

Le dernier coefficient CCC[91,92] mesure à la fois la précision (distance par rapport à l'équation) et la justesse (c'est-à-dire à quel point la ligne de la régression dévie de la droite x=y dite « concordance

line »). Il s'agit d'une validation externe car aucune information du jeu d'entraînement n'est nécessaire.

(3. 64)
$$CCC = \frac{2\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2 + \sum_{i=1}^{n}(y_i - \bar{y})^2 + n(\bar{x} - \bar{y})^2}$$

Tous ces coefficients ainsi que les cinq principes de l'OCDE ont pour but l'améliorer la validation du modèle et ainsi d'augmenter la confiance en ce type de méthode. Le but étant de pouvoir utiliser les modèles QSPR avec assurance pour prédire les propriétés physico-chimiques et toxicologiques, pour répondre aux exigences des différents règlements, notamment dans l'industrie.

4. Critères de validation

Parmi les critères de validation les plus utilisés, Tropsha[81] propose d'accéder à la prédictivité du modèle en mesurant les coefficients de détermination lorsque la ligne de régression passe par zéro, R_0^2 (valeurs prédites vs valeurs expérimentales) et $R'_0{}^2$ (valeurs expérimentales vs valeurs prédites) ainsi que les pentes k et k' de ces lignes de régressions :

- $Q^2>0,5$
- $R^2>0,6$
- $\frac{\left(R^2 - R_0^2\right)}{R^2} < 0,1$ ou $\frac{\left(R^2 - R_0'^2\right)}{R^2} < 0,1$
- $0,85<k$ ou $k'<1,15$

La validation est en évolution permanente avec l'utilisation de nouveaux coefficients (voir la partie Validation de ce chapitre). De manière générale, les coefficients R^2 et Q^2 doivent avoir des valeurs proches de 1 (de préférence supérieures à 0,6) et leur différence doit être faible pour considérer le modèle comme robuste. Cependant, l'évaluation des coefficients doit se faire au regard de la taille de la base de données (notamment pour R^2) et de l'ordre de grandeur de l'incertitude expérimentale (RMSE et MAE). Mais d'autres paramètres sont pris en considération pour le choix du modèle comme la possibilité d'interprétation des descripteurs.

VII. Domaine d'applicabilité

Le 3ème principe de l'OCDE pour la validation des modèles QSAR/QSPR est la définition d'un domaine d'applicabilité. En effet, un modèle QSPR s'applique uniquement à des composés similaires à ceux avec lesquels le modèle a été développé. Le domaine d'applicabilité[54,55,93–97] du modèle (AD) est l'espace chimique dans lequel le modèle est fiable et peut être interpolé. Il est déterminé sur les données du jeu d'entraînement à partir des descripteurs du modèle (voir Figure 17). Il permet de déterminer si le modèle peut être utilisé pour prédire la propriété pour une nouvelle molécule

donnée. Il existe plusieurs types de méthodes pour déterminer l'AD comme les intervalles sur la valeur des descripteurs du modèle, la distribution de la densité, la distance.

La méthode *bonding box* est une méthode simple basée sur les intervalles de valeur des descripteurs du modèle. Une nouvelle molécule est considérée dans le domaine d'applicabilité si la valeur de ces descripteurs se situe dans l'intervalle respectif du maximum et du minimum de chaque descripteur. Cette méthode a pour inconvénient de ne pas prendre en compte les « espaces vides ». Ainsi, si une molécule se trouve seule dans « un coin » du rectangle délimitant le domaine d'applicabilité, elle ne sera pas détectée.

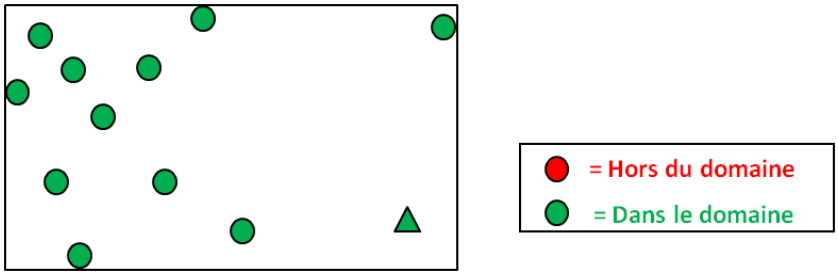


Figure 24 : Exemple du domaine d'applicabilité défini par la méthode "bonding box" (le triangle correspond à une molécule du jeu d'entraînement)

Le domaine d'applicabilité peut être défini par des méthodes basées sur la géométrie, comme la région convexe contenant toutes les molécules illustrée en Figure 25. Bien que cette méthode réduit la taille du domaine d'applicabilité, elle ne prend toujours pas en compte la présence d'espaces vide.

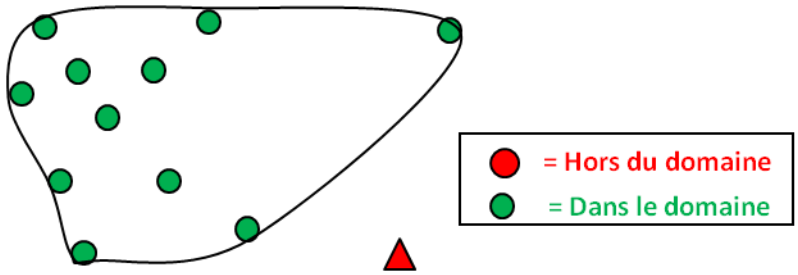


Figure 25 : Exemple du domaine d'applicabilité défini par la méthode géométrique « région convexe » (le triangle correspond à une molécule du jeu d'entraînement)

Une autre possibilité consiste à utiliser les William's plot, présentés par Gramatica[55].

Le domaine d'applicabilité est aussi défini en termes de valeur de la propriété et de type de molécule comme les nitroaliphatiques avec une chaîne carbonée de moins de 24 atomes, les nitroaromatiques à l'exclusion de ceux ayant un substituant en position ortho ou les composés organiques en général…

Dans cette thèse, nous avons choisi une méthode utilisant la distance euclidienne (équation (3. 65)) avec un seuil de 95% (pour plus de précautions) qui définit le domaine comme un cercle dont le

centre est celui de la distribution des valeurs et le diamètre est calculé pour contenir 95% des données du jeu d'entraînement (comme illustré Figure 26). En effet, la diversité des données implique la présence de valeur des descripteurs pouvant être très discriminant pour quelques molécules et ainsi créé un espace vide.

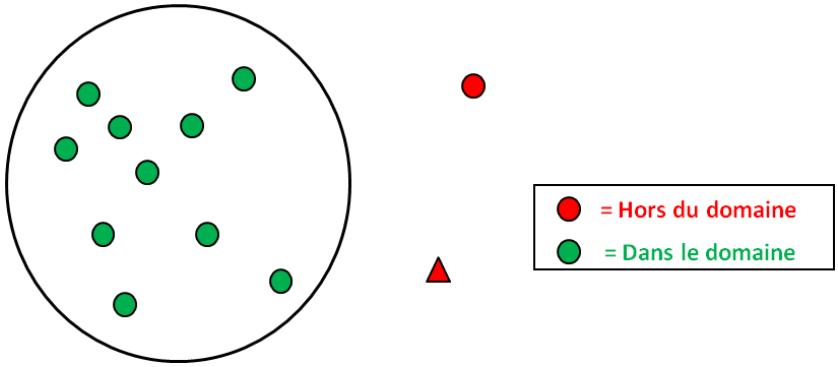


Figure 26: Domaine d'applicabilité défini par la distance euclidienne (le triangle correspond à une molécule du jeu d'entraînement)

Le logiciel *Ambit discovery*[98] a été utilisé pour déterminer le domaine d'applicabilité du modèle. Dans le cadre des distances euclidiennes, cette distance entre deux points X (x_1,x_2…x_n) et Y (y_1,y_2…y_n) est définie par la formule suivante :

(3. 65)
$$d(X,Y)=\sqrt{\sum_{i=1}^{n}(y_i-x_i)^2}$$

Les performances du modèle sont calculées à l'intérieur de son domaine d'applicabilité en supprimant les molécules des deux jeux hors du domaine d'applicabilité. Notamment, les coefficients relatifs à la mesure de la prédictivité sont recalculés en prenant en compte uniquement les molécules du jeu de validation dans le domaine d'applicabilité.

VIII. EXEMPLE DES COMPOSÉS NITROALIPHATIQUES

Dans cette partie, un exemple complet de développement d'un modèle validé et prédictif est présenté. Nous nous intéressons à la sensibilité à l'impact qui caractérise la tendance du matériau à réagir sous l'effet d'un impact. Cette grandeur, notée H_{50} (en cm), mesure la hauteur à partir de laquelle un poids de masse donnée, lâché sur un échantillon, provoque une réaction avec une probabilité de 50%. Les substances possédant des H_{50} peu élevées sont les plus sensibles à l'impact, puisque l'énergie à apporter pour les faire réagir est faible. Cette propriété sera définie plus en détail dans le chapitre 4 (paragraphe II 5) du manuscrit. Les composés nitroaliphatiques, qui sont une classe d'explosifs bien connue en raison de la présence de groupements explosophores NO_2, sont considérés dans cet exemple.

Une base de données de 50 composés nitroaliphatiques dont les valeurs expérimentales proviennent d'une source unique[99] a été utilisée afin de garantir l'homogénéité des données et répondre également au 1er principe de l'OCDE pour la validation des modèles. La BMLR a été utilisée pour sélectionner les descripteurs et développer le modèle.

Le modèle le plus simple en termes de descripteurs, parmi ceux développés[100] avec cette base de données, sera présenté ici. Les modèles pour cette propriété sont en général développés sur le log[101–105], ainsi un modèle à trois descripteurs a été obtenu à l'aide de la BMLR à partir de 66 descripteurs constitutionnels pour $logh_{50\%}$:

(3. 66)
$$\log h_{50\%} = -2{,}53\, n_N / n_{atomes} + 0{,}07\, n_{simple} - 0{,}25\, n_{NO2} + 1{,}94$$

avec n_N/n_{atomes} le nombre relatif d'atomes d'azote (t-test= -2,2), n_{simple} le nombre de liaisons simples (t-test=7,9) et n_{NO2} le nombre de groupements NO_2 (t-test=-8,4).

La Figure 15 illustre le choix de ce modèle : Q^2 diminue lorsqu'on passe au modèle à 4 descripteurs et le R^2 augmente seulement de 0,01.

D'un point de vue chimique, le nombre de groupement NO_2 peut s'expliquer par le fait que la coupure de la liaison C-NO_2 est connue comme étant la première étape du mécanisme principal de décomposition de ces composés[106–109]. Le nombre relatif d'atomes d'azote et le nombre de liaisons simples sont moins directement interprétables.

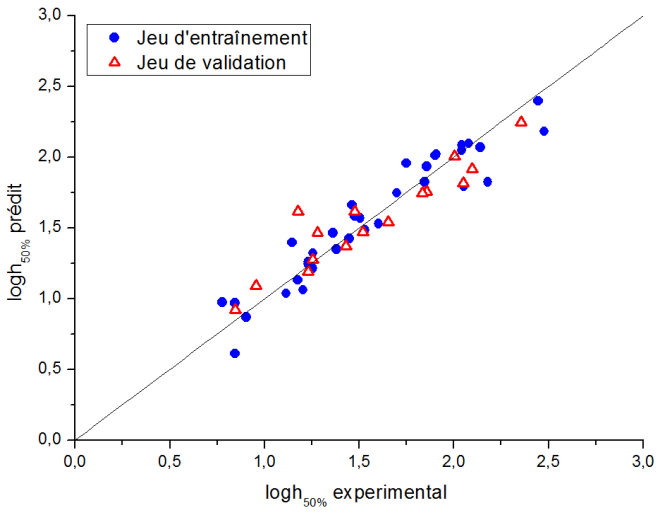


Figure 27: Valeur prédites par le modèle vs valeurs expérimentales pour la sensibilité à l'impact des nitroaliphatiques

La Figure 27 représente les valeurs prédites par le modèle en fonction des valeurs expérimentales. Le modèle est caractérisé par une bonne corrélation (R^2=0,88, RMSE=0,17) et robustesse (Q^2_{LOO}=Q^2_{5CV}=0,85 et Q^2_{10CV}=0,84). La méthode Y-scrambling valide le modèle avec des faibles valeurs

de R^2 pour les modèles fortuits (R^2_{YS}=0,09, σ_{YS}=0,07) comme illustré Figure 23. De plus, le critère de Rücker[88] est respecté : R^2 - R^2_{YS} (0,79) est plus élevé que $3\sigma_{YS}$ (0,22).

Le pouvoir prédictif est élevé (R^2_{EXT}=0,81, $RMSE_{EXT}$=0,22, Q^2_{F1}= Q^2_{F2}=0,81, Q^2_{F3}=0,83 et CCC=0,93). Les performances dans le domaine d'applicabilité (R^2_{IN}=0,78, $RMSE_{IN}$=0,23, $Q^2_{F1,IN}$ =$Q^2_{F2,IN}$ =0,78, $Q^2_{F3,IN}$ =0,82 et CCC_{IN} =0,92), duquel uniquement une molécule est exclue du jeu de validation, sont également bonnes.

Ce modèle respecte tous les principes de l'OCDE :

1) Une propriété ciblée définie (avec un protocole expérimental identifié) : la sensibilité à l'impact ;
2) Un algorithme sans équivoque : les structures sont optimisées avec une démarche précise et un niveau de calcul DFT (PBE0) et les descripteurs sont calculés et sélectionnés avec un logiciel défini clairement (BMLR) ;
3) Un domaine d'applicabilité défini ;
4) Des mesures appropriées de la qualité d'ajustement (R^2, MAE, RMSE), de la robustesse (Q^2) et du pouvoir prédictif (R^2, MAE, Q^2_{F1}, Q^2_{F2}, Q^2_{F3} et CCC);
5) Si possible, une interprétation des mécanismes sous-jacents (par le choix des descripteurs).

Ce modèle a été proposé pour son intégration dans la QSAR toolbox de l'OCDE et ECHA[110] (logiciel développé et reconnu par les instances réglementaires dans lequel des modèles QSAR sont disponibles pour l'enregistrement des substances pour REACH). Un fichier QMRF, récapitulant les informations clés sur les modèles QSAR afin de les enregistrer par les instances réglementaires, a été rempli et envoyé pour une reconnaissance réglementaire de ce modèle. On a ainsi atteint tous les objectifs : développement, validation scientifique et réglementaire du modèle.

IX. CONCLUSION

Cette partie a présenté les méthodes utilisées dans cette thèse pour le développement de modèles QSAR/QSPR selon les 5 principes de l'OCDE[70] pour la validation des modèles QSAR/QSPR. En effet, une attention particulière a été portée aux méthodes de validation des modèles avec la validation interne (validation croisée, Y-scrambling…) et la validation externe avec notamment le calcul du pouvoir prédictif.

L'exemple de la prédiction de la sensibilité à l'impact des composés nitroaliphatiques illustre la pertinence de la méthode qui sera utilisée dans ce manuscrit, en termes des performances et permet aussi de montrer que les modèles QSPR sont une alternative sérieuse pour la prédiction de propriétés physico-chimiques, en particulier dans le cadre de REACH avec une validation réglementaire (principes OCDE, QMRF).

X. RÉFÉRENCES

(1) Selassie, C. D. *History of Quantitative Structure-Activity Relationship*; Burger's Medicinal Chemistry and Drug Discovery Sixth Edition; John Wiley&Sons Inc., 2002; Vol. 1.

(2) Willett, P. Chemoinformatics: a History. *Wiley Interdisciplinary Reviews: Computational Molecular Science* **2011**, *1*, 46–56.

(3) Hansch, C.; Fujita, T. P-σ-π Analysis. A Method for the Correlation of Biological Activity and Chemical Structure. *J. Am. Chem. Soc.* **1964**, *86*, 1616–1626.

(4) Brown, A. C.; Fraser, T. R. On the Connection Between Chemical Constitution and Physiological Action; with Special Reference to the Physiological Action of the Salts of the Ammonium Bases Derived from Strychnia, Brucia, Thebaia, Codeia, Morphia, and Nicotia. *J Anat Physiol* **1868**, *2*, 224–242.

(5) Richardson, B. J. Physiological Research on Alcohols. *Med. Times Gaz.* **1868**, *2*, 703–706.

(6) Richet, C. On the Relationship Between the Toxicity and the Physical Properties of Substances. *Compt. Rendus Seances Soc. Biol.* **1893**, 775–776.

(7) Meyer, H. On the Theory of Alcohol Narcosis I. Which Property of Anesthetics Gives Them Their Narcotic Activity? *Arch. Exp. Pathol. Pharmakol.* **1899**, 109–118.

(8) Overton, C. E. *Studien Über Die Narkose Zugleich Ein Beitrag Zur Allgemeinen Pharmakologie*; Fischer, 1901.

(9) Hammett, L. P. The Effect of Structure Upon the Reactions of Organic Compounds. Benzene Derivatives. *J. Am. Chem. Soc.* **1937**, *59*, 96–103.

(10) Taft, R. W. Polar and Steric Substituent Constants for Aliphatic and o-Benzoate Groups from Rates of Esterification and Hydrolysis of Esters1. *J. Am. Chem. Soc.* **1952**, *74*, 3120–3128.

(11) Taft, R. W. The General Nature of the Proportionality of Polar Effects of Substituent Groups in Organic Chemistry. *J. Am. Chem. Soc.* **1953**, *75*, 4231–4238.

(12) Winkler, D. A. The Role of Quantitative Structure--activity Relationships (QSAR) in Biomolecular Discovery. *Brief. Bioinformatics* **2002**, *3*, 73–86.

(13) Bradbury, S. P. Quantitative Structure-activity Relationships and Ecological Risk Assessment: An Overview of Predictive Aquatic Toxicology Research. *Toxicology Letters* **1995**, *79*, 229–237.

(14) Grover, M.; Singh, B.; Bakshi, M.; Singh, S. Quantitative Structure–property Relationships in Pharmaceutical Research – Part 2. *Pharmaceutical Science & Technology Today* **2000**, *3*, 50–57.

(15) Grover, M.; Singh, B.; Bakshi, M.; Singh, S. Quantitative Structure–property Relationships in Pharmaceutical Research – Part 1. *Pharmaceutical Science & Technology Today* **2000**, *3*, 28–35.

(16) Dearden, J. C.; Rotureau, P.; Fayet, G. QSPR Prediction of Physico-chemical Properties for REACH. *SAR and QSAR in Environmental Research* **2013**, *24*, 279–318.

(17) Quintero, F. A.; Patel, S. J.; Muñoz, F.; Sam Mannan, M. Review of Existing QSAR/QSPR Models Developed for Properties Used in Hazardous Chemicals Classification System. *Ind. Eng. Chem. Res.* **2012**, *51*, 16101–16115.

(18) Leach, A. R. *Introduction to Chemoinformatics*; Springer, 2007.

(19) Dearden, J. C.; Cronin, M. T. D.; Kaiser, K. L. E. How Not to Develop a Quantitative Structure–activity or Structure–property Relationship (QSAR/QSPR). *SAR and QSAR in Environmental Research* **2009**, *20*, 241–266.

(20) Young, D.; Martin, T.; Venkatapathy, R.; Harten, P. Are the Chemical Structures in Your QSAR Correct? *QSAR & Combinatorial Science* **2008**, *27*, 1337–1345.

(21) Fourches, D.; Muratov, E.; Tropsha, A. Trust, But Verify: On the Importance of Chemical Structure Curation in Cheminformatics and QSAR Modeling Research. *J. Chem. Inf. Model.* **2010**, *50*, 1189–1204.

(22) Muehlbacher, M.; Kerdawy, A. E.; Kramer, C.; Hudson, B.; Clark, T. Conformation-Dependent QSPR Models: logPOW. *J. Chem. Inf. Model.* **2011**, *51*, 2408–2416.

(23) Royal Society of Chemistry ChemSpider - The free chemical database http://www.chemspider.com/ (accessed Jun 12, 2013).

(24) Advenced Chemistry Development, Inc. IUPAC Nomenclature of Organics Chemistry http://www.acdlabs.com/iupac/nomenclature.

(25) Daylight Chemical Information Systems, Inc. Simplified Molecular Input Line Entry System http://www.daylight.com/smiles/index.html (accessed Feb 28, 2013).

(26) Anderson, E.; Veith, G. .; Weininger, D. SMILES: A Line Notation and Computerized Interpreter for Chemical Structures. Report No. EPA/600/M-87/021. Duluth, MN 55804: U.S. EPA, Environmental Research Laboratory-Duluth. **1987**.

(27) Weininger, D.; Weininger, A.; Weininger, J. L. SMILES. 2. Algorithm for Generation of Unique SMILES Notation. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 97–101.

(28) *Scigress*; FUJITSU, 2008.

(29) Allinger, N. L.; Yuh, Y. H.; Lii, J. H. Molecular Mechanics. The MM3 Force Field for Hydrocarbons. 1. *Journal of the American Chemical Society* **1989**, *111*, 8551–8566.

(30) Goto, H.; Osawa, E. Corner Flapping - a Simple and Fast Algorithm for Exhaustive Generation of Ring Conformations. *Journal of the American Chemical Society* **1989**, *111*, 8950–8951.

(31) Goto, H.; Osawa, E. An Efficient Algorithm for Searching Low-energy Conformers of Cyclic and Acyclic Molecules. *Journal of the Chemical Society-Perkin Transactions 2* **1993**, 187–198.

(32) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Montgomery, J. A.; Vreven, T.; Kudin, K. N.; Burant, J. C.; et al. *Gaussian09*; Gaussian, Inc.: Wallingford CT, 2009.

(33) Adamo, C.; Barone, V. Toward Reliable Density Functional Methods Without Adjustable Parameters: The PBE0 Model. *Journal of Chemical Physics* **1999**, *110*, 6158–6170.

(34) Todeschini, R. *Handbook of Molecular Descriptors*; Wiley-VCH: Weinheim, New York, 2000.

(35) Talete srl *DRAGON (Software for Molecular Descriptor Calculation)*; Milano, Italy, 2012.

(36) *CodessaPro*; 2002.

(37) Varnek, A.; Fourches, D.; Horvath, D.; Klimchuk, O.; Gaudin, C.; Vayer, P.; Solov'ev, V.; Hoonakker, F.; Tetko, I. V.; Marcou, G. ISIDA - Platform for Virtual Screening Based on Fragment and Pharmacophoric Descriptors. *Current Computer - Aided Drug Design* **2008**, *4*, 191–198.

(38) Wiener, H. Structural Determination of Paraffin Boiling Points. *J. Am. Chem. Soc.* **1947**, *69*, 17–20.

(39) Randic, M. Characterization of Molecular Branching. *J. Am. Chem. Soc.* **1975**, *97*, 6609–6615.

(40) Stanton, D. T.; Dimitrov, S.; Grancharov, V.; Mekenyan, O. G. Charged Partial Surface Area (CPSA) Descriptors QSAR Applications. *SAR and QSAR in Environmental Research* **2002**, *13*, 341 – 351.

(41) Todeschini, R.; Lasagni, M.; Marengo, E. New Molecular Descriptors for 2D and 3D Structures. Theory. *Journal of Chemometrics* **1994**, *8*, 263–272.

(42) Chermette, H. Chemical Reactivity Indexes in Density Functional Theory. *Journal of Computational Chemistry* **1999**, *20*, 129–154.

(43) Geerlings, P.; De Proft, F.; Langenaeker, W. Conceptual Density Functional Theory. *Chemical Reviews* **2003**, *103*, 1793–1873.

(44) Lindgren, F.; Hansen, B.; Karcher, W.; Sjöström, M.; Eriksson, L. Model Validation by Permutation Tests: Applications to Variable Selection. *Journal of Chemometrics* **1996**, *10*, 521–532.

(45) Xu, L.; Zhang, W.-J. Comparison of Different Methods for Variable Selection. *Analytica Chimica Acta* **2001**, *446*, 475–481.

(46) Wallet, B. C.; Marchette, D. J.; Solka, J. L.; Wegman, E. J. A Genetic Algorithm for Best Subset Selection in Linear Regression. *Proceedings of the 28th Symposium on the Interface* **1996**.

(47) Ghafourian, T.; Cronin, M. T. D. The Impact of Variable Selection on the Modelling of Oestrogenicity. *SAR and QSAR in Environmental Research* **2005**, *16*, 171–190.

(48) Cho, S. J.; Hermsmeier, M. A. Genetic Algorithm Guided Selection: Variable Selection and Subset Selection. *Journal of Chemical Information and Modeling* **2002**, *42*, 927–936.

(49) Yasri, A.; Hartsough, D. Toward an Optimal Procedure for Variable Selection and QSAR Model Building. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1218–1227.

(50) Sutter, J. M.; Dixon, S. L.; Jurs, P. C. Automated Descriptor Selection for Quantitative Structure-Activity Relationships Using Generalized Simulated Annealing. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 77–84.

(51) Hasegawa, K.; Miyashita, Y.; Funatsu, K. GA Strategy for Variable Selection in QSAR Studies: GA-Based PLS Analysis of Calcium Channel Antagonists⊥. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 306–310.

(52) Hasegawa, K.; Funatsu, K. GA Strategy for Variable Selection in QSAR Studies: GAPLS and D-optimal Designs for Predictive QSAR Model. *Journal of Molecular Structure: THEOCHEM* **1998**, *425*, 255–262.

(53) Kimura, T.; Hasegawa, K.; Funatsu, K. GA Strategy for Variable Selection in QSAR Studies: GA-Based Region Selection for CoMFA Modeling. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 276–282.

(54) Bhhatarai, B.; Gramatica, P. Prediction of Aqueous Solubility, Vapor Pressure and Critical Micelle Concentration for Aquatic Partitioning of Perfluorinated Chemicals. *Environ. Sci. Technol.* **2011**, *45*, 8120–8128.

(55) Roy, P. P.; Kovarich, S.; Gramatica, P. QSAR Model Reproducibility and Applicability: A Case Study of Rate Constants of Hydroxyl Radical Reaction Models Applied to Polybrominated Diphenyl Ethers and (benzo-)triazoles. *Journal of Computational Chemistry* **2011**, *32*, 2386–2396.

(56) Tenenhaus, M.; Gauchi, J.-P.; Ménardo, C. Régression PLS et applications. *Revue de Statistique Appliquée* **1995**, *43*, 7–63.

(57) Karelson, M. *Molecular Descriptors in QSAR/QSPR*; Wiley: New York, 2000.

(58) Peduzzi, P.; Concato, J.; Feinstein, A. R.; Holford, T. R. Importance of Events Per Independent Variable in Proportional Hazards Regression Analysis II. Accuracy and Precision of Regression Estimates. *Journal of Clinical Epidemiology* **1995**, *48*, 1503–1510.

(59) Babyak, M. A. What You See May Not Be What You Get: a Brief, Nontechnical Introduction to Overfitting in Regression-type Models. *Psychosom Med* **2004**, *66*, 411–421.

(60) Hawkins, D. M. The Problem of Overfitting. *Journal of Chemical Information and Modeling* **2004**, *44*, 1–12.

(61) Puzyn, T.; Mostrag-Szlichtyng, A.; Gajewicz, A.; Skrzyński, M.; Worth, A. P. Investigating the Influence of Data Splitting on the Predictive Ability of QSAR/QSPR Models. *Structural Chemistry* **2011**, *22*, 795–804.

(62) Leonard, J. T.; Roy, K. On Selection of Training and Test Sets for the Development of Predictive QSAR Models. *Qsar & Combinatorial Science* **2006**, *25*, 235–251.

(63) Fayet, G.; Del Rio, A.; Rotureau, P.; Joubert, L.; Adamo, C. Predicting the Thermal Stability of Nitroaromatic Compounds Using Chemoinformatic Tools. *Molecular Informatics* **2011**, *30*, 623–634.

(64) Golbraikh, A.; Tropsha, A. Predictive QSAR Modeling Based on Diversity Sampling of Experimental Datasets for the Training and Test Set Selection. *Journal of Computer-Aided Molecular Design* **2002**, *16*, 357–369.

(65) Golbraikh, A.; Shen, M.; Xiao, Z. Y.; Xiao, Y. D.; Lee, K. H.; Tropsha, A. Rational Selection of Training and Test Sets for the Development of Validated QSAR Models. *Journal of Computer-Aided Molecular Design* **2003**, *17*, 241–253.

(66) Roy, K. On Some Aspects of Validation of Predictive Quantitative Structure–activity Relationship Models. *Expert Opinion on Drug Discovery* **2007**, *2*, 1567–1577.

(67) Katritzky, A. R.; Stoyanova-Slavova, I. B.; Dobchev, D. A.; Karelson, M. QSPR Modeling of Flash Points: An Update. *Journal of Molecular Graphics and Modelling* **2007**, *26*, 529–536.

(68) Schneider, G.; Wrede, P. Artificial Neural Networks for Computer-based Molecular Design. *Progress in Biophysics and Molecular Biology* **1998**, *70*, 175–222.

(69) Quinlan, J. R. Induction of Decision Trees. *Mach Learn* **1986**, *1*, 81–106.

(70) Organisation de Coopération et de Développement Economiques (OCDE) *Principles for the Validation, for Regulatory Purposes, of (Quantitative) Structure-Activity Relationship Models*; Paris, 2009.

(71) Gramatica, P. Principles of QSAR Models Validation: Internal and External. *Qsar & Combinatorial Science* **2007**, *26*, 694–701.

(72) Golbraikh, A.; Tropsha, A. Beware of Q(2)! *Journal of Molecular Graphics & Modelling* **2002**, *20*, 269–276.

(73) Efron, B.; Tibshirani, R. J. *An Introduction to the Bootstrap*; 1st ed.; Chapman and Hall/CRC, 1994.

(74) Wehrens, R.; Putter, H.; Buydens, L. M. . The Bootstrap: a Tutorial. *Chemometrics and Intelligent Laboratory Systems* **2000**, *54*, 35–52.

(75) Efron, B. Bootstrap Methods: Another Look at the Jackknife. *The Annals of Statistics* **1979**, *7*, 1–26.

(76) Joint Research Centre - JRC - European Commission http://ec.europa.eu/dgs/jrc/index.cfm (accessed Apr 18, 2013).

(77) (Q)SAR Model Reporting Format (QMRF) Inventory http://qsardb.jrc.it/qmrf/ (accessed Apr 4, 2013).

(78) Hutter, M. C. Determining the Degree of Randomness of Descriptors in Linear Regression Equations with Respect to the Data Size. *J. Chem. Inf. Model.* **2011**, *51*, 3099–3104.

(79) Sies, H. A New Parameter for Sex Education. *Nature* **1988**, *332*, 495–495.

(80) Johnson, S. R. The Trouble with QSAR (or How I Learned To Stop Worrying and Embrace Fallacy). *J. Chem. Inf. Model.* **2008**, *48*, 25–26.

(81) Tropsha, A.; Gramatica, P.; Gombar, V. K. The Importance of Being Earnest: Validation Is the Absolute Essential for Successful Application and Interpretation of QSPR Models. *Qsar & Combinatorial Science* **2003**, *22*, 69–77.

(82) Wilcox, R. E.; Huang, W.-H.; Brusniak, M.-Y. K.; Wilcox, D. M.; Pearlman, R. S.; Teeter, M. M.; DuRand, C. J.; Wiens, B. L.; Neve, K. A. CoMFA-Based Prediction of Agonist Affinities at Recombinant Wild Type Versus Serine to Alanine Point Mutated D2 Dopamine Receptors. *Journal of Medicinal Chemistry* **2000**, *43*, 3005–3019.

(83) Kubinyi, H. Comparative Molecular Field Analysis (CoMFA). In *Handbook of Chemoinformatics*; Gasteiger, J., Ed.; Wiley-VCH Verlag GmbH, 2008; pp. 1555–1574.

(84) Papa, E.; Kovarich, S.; Gramatica, P. Development, Validation and Inspection of the Applicability Domain of QSPR Models for Physicochemical Properties of Polybrominated Diphenyl Ethers. *QSAR & Combinatorial Science* **2009**, *28*, 790–796.

(85) Schüürmann, G.; Ebert, R.-U.; Chen, J.; Wang, B.; Kühne, R. External Validation and Prediction Employing the Predictive Squared Correlation Coefficient - Test Set Activity Mean Vs Training Set Activity Mean. *Journal of Chemical Information and Modeling* **2008**, *48*, 2140–2145.

(86) Wang, R.; Jiang, J. C.; Pan, Y.; Cao, H. Y.; Cui, Y. Prediction of Impact Sensitivity of Nitro Energetic Compounds by Neural Network Based on Electrotopological-state Indices. *Journal of Hazardous Materials* **2009**, *166*, 155–186.

(87) Lozano, S.; Halm-Lemeille, M.-P.; Lepailleur, A.; Rault, S.; Bureau, R. Consensus QSAR Related to Global or MOA Models: Application to Acute Toxicity for Fish. *Molecular Informatics* **2010**, *29*, 803–813.

(88) Rücker, C.; Rücker, G.; Meringer, M. y-Randomization and Its Variants in QSPR/QSAR. *J. Chem. Inf. Model.* **2007**, *47*, 2345–2357.

(89) Chirico, N.; Gramatica, P. Real External Predictivity of QSAR Models: How To Evaluate It? Comparison of Different Validation Criteria and Proposal of Using the Concordance Correlation Coefficient. *Journal of Chemical Information and Modeling* **2011**, *51*, 2320–2335.

(90) Consonni, V.; Ballabio, D.; Todeschini, R. Comments on the Definition of the Q^2 Parameter for QSAR Validation. *Journal of Chemical Information and Modeling* **2009**, *49*, 1669–1678.

(91) Lin, L. I.-K. A Concordance Correlation Coefficient to Evaluate Reproducibility. *Biometrics* **1989**, *45*, 255–268.

(92) Lin, L. I.-K. Assay Validation Using the Concordance Correlation Coefficient. *Biometrics* **1992**, *48*, 599–604.

(93) Jaworska, J.; Nikolova-Jeliazkova, N.; Aldenberg, T. QSAR Applicabilty Domain Estimation by Projection of the Training Set Descriptor Space: a Review. *Altern Lab Anim* **2005**, *33*, 445–459.

(94) Netzeva, T. I.; Worth, A.; Aldenberg, T.; Benigni, R.; Cronin, M. T. D.; Gramatica, P.; Jaworska, J. S.; Kahn, S.; Klopman, G.; Marchant, C. A.; et al. Current Status of Methods for Defining the Applicability Domain of (quantitative) Structure-activity Relationships. The Report and Recommendations of ECVAM Workshop 52. *Altern Lab Anim* **2005**, *33*, 155–173.

(95) Stanforth, R. W.; Kolossov, E.; Mirkin, B. A Measure of Domain of Applicability for QSAR Modelling Based on IntelligentK-Means Clustering. *QSAR & Combinatorial Science* **2007**, *26*, 837–844.

(96) Weaver, S.; Gleeson, M. P. The Importance of the Domain of Applicability in QSAR Modeling. *Journal of Molecular Graphics and Modelling* **2008**, *26*, 1315–1326.

(97) Baskin, I. I.; Kireeva, N.; Varnek, A. The One-Class Classification Approach to Data Description and to Models Applicability Domain. *Molecular Informatics* **2010**, *29*, 581–587.

(98) Jeliazkova, N.; Jaworska, J. *Ambit Discovery*; 2007.

(99) Storm, C. B.; Stine, J. R.; Kramer, J. F. *Sensitivity Relationships in Energetic Materials. Chemistry and Physics of Energetic Materials*; Kluwer Academic Publishers, 1990.

(100) Prana, V.; Fayet, G.; Rotureau, P.; Adamo, C. Development of Validated QSPR Models for Impact Sensitivity of Nitroaliphatic Compounds. *J. Hazard. Mater.* **2012**, *235-236*, 169–177.

(101) Kamlet, M. J. The Relationship of Impact Sensitivity with Structure of Organic High Explosives. I Polynitroaliphatic Explosives. In; Coronado, California, 1976; p. 312.

(102) Mullay, J. Relationships Between Impact Sensitivity and Molecular Electronegativity. *Propellants, Explosives, Pyrotechnics* **1987**, *12*, 60–63.

(103) Badders, N. R.; Wei, C.; Aldeeb, A. A.; Rogers, W. J.; Mannan, M. S. Predicting the Impact Sensitivities of Polynitro Compounds Using Quantum Chemical Descriptors. *Journal of Energetic Materials* **2006**, *24*, 17–33.

(104) Keshavarz, M. H.; Pouretedal, H. R. Simple Empirical Method for Prediction of Impact Sensitivity of Selected Class of Explosives. *Journal of Hazardous Materials* **2005**, *124*, 27–33.

(105) Lai, W. P.; Lian, P.; Wang, B. Z.; Ge, Z. X. New Correlations for Predicting Impact Sensitivities of Nitro Energetic Compounds. *Journal of Energetic Materials* **2010**, *28*, 45–76.

(106) Dewar, M. J. S.; Ritchie, J. P.; Alster, J. Ground-state of Molecules .65. Thermolysis of Molecules Containing No2 Groups. *Journal of Organic Chemistry* **1985**, *50*, 1031–1036.

(107) Nazin, G. M.; Manelis, G. B. Thermal-decomposition of Aliphatic Nitrocompounds. *Uspekhi Khimii* **1994**, *63*, 327–337.

(108) Nazin, G. M.; Manelis, G. B.; Nechiporenko, G. N.; Dubovitsky, F. I. Thermal Decomposition of Some Polynitrocompounds in Gas Phase. *Combustion and Flame* **1968**, *12*, 102–106.

(109) Korolev, V. L.; Pivina, T. S.; Porollo, A. A.; Petukhova, T. V.; Sheremetev, A. B.; Ivshin, V. P. Differentiation of the Molecular Structures of Nitro Compounds as the Basis for Simulation of Their Thermal Destruction Processes. *Russian Chemical Reviews* **2009**, *78*, 945–969.

(110) OECD; ECHA *QSAR Toolbox*; 2012.

CHAPITRE 4 – MODÈLES À PARTIR DES DONNÉES DE LA DATATOP

La Datatop[1] est une base de données mise en place par le TNO[2] (Organisation néerlandaise pour la recherche scientifique appliquée) en compilant des essais réalisés par différents organismes sur les peroxydes organiques. Les essais effectués correspondent à ceux demandés dans le Manuel d'épreuves et de critères pour la classification des peroxydes organiques (matières de la classe 5.2) selon les réglementations TDG[3] et CLP[4] déjà évoquées dans le chapitre 1. Dans ce chapitre, les différentes épreuves de cette base de données, seront décrites. Les propriétés caractérisées par ces épreuves sont associées à la stabilité (thermique, au choc…) de ces composés qui est directement liée à leur décomposition. L'énergie de dissociation des 105 peroxydes organiques de cette base de données sera calculée et leur corrélation avec différents descripteurs et propriétés sera étudiée. Puis les modèles QSPR développés à partir de cette base de données seront présentés.

I. PRÉSENTATION DE LA BASE DE DONNÉES

Il est difficile de trouver une base de données expérimentale pour les propriétés explosibles, c'est-à-dire caractérisant la tendance d'une substance à subir une décomposition violente et rapide produisant de la chaleur et/ou des gaz. En effet, la plupart des mesures présentées dans la littérature sont faites sur quelques peroxydes organiques[5,6]. Néanmoins une base de données contenant un grand nombre de peroxydes organiques est disponible parmi la communauté industrielle : la Datatop.

Dans un premier temps, cette base de données compilée par le TNO en 2005 et regroupant les résultats d'essais de classification des peroxydes organiques a été analysée : elle contient plus de 270 données (sur plus de 100 préparation de peroxydes organiques à des concentrations différentes) et plus de 40 colonnes (noms, concentration, conditions et résultats d'essais…). Les peroxydes de la Datatop sont rarement purs et les concentrations maximales sont très différentes d'un peroxyde à un autre. Cela s'explique par le fait que, pour des raisons de sécurité déjà évoquées dans le chapitre 1, les peroxydes ne sont pas transportés et stockés purs. Les propriétés explosives semblent varier avec la concentration. En effet, la Datatop nous permet d'observer que, pour une même substance, la valeur du résultat d'une épreuve peut être différente quand la concentration change (voir Tableau 10).

Tableau 10: Extrait de la Datatop (propriétés définies dans le Tableau 12 et le paragraphe II. « Propriétés expérimentales sélectionnées »)

Noms	C [%]	C.1	C.2	E.1 Koenen		E.2	TDAA
		(ms)	(mm/s)	(mm)	(effect)	(mm)	(°C)
tert-butyl peroxypivalate	> 67 - 77	220	0.08	2		8	25
	> 27 - 67			1.5			20
	<= 27			<1	A	<1	40
tert-butyl peroxy-3,5,5-trimethylhexanoate	> 32 - 100	1375	0.27	1		6	55
	<= 42			<1	O	<1	55
	<= 32	8400	0.27	<1	A	<1	90

Ces résultats d'épreuves n'ont pas pu être directement reliés aux propriétés explosives demandées dans REACH car ces dernières ne sont pas clairement définies dans la réglementation, contrairement aux propriétés comme la pression de vapeur, la solubilité dans l'eau ou encore le coefficient de partage n-octanol/eau. La Datatop est ensuite épurée en supprimant les mélanges d'isomères comme le 2-butanone peroxide (aussi connu par les industriels sous le nom de methyle-ethyle-cétone peroxyde ou MEKPO[7]) et les molécules avec des noms ambigus (qui ne permettent pas d'obtenir une structure unique) comme le cumyle peroxyneodecanoate. La base de données étudiée

a ensuite été agrandie avec quelques molécules présélectionnées avec le partenaire Arkema (structure uniquement, pas de valeur de propriété). Une liste de 105 peroxydes organiques uniques répartis dans les différentes familles, présentée dans le Tableau 11, est retenue pour notre étude. La Figure 28 représente ces 105 molécules dans l'espace des descripteurs (plus de 300). L'observation nous permet de visualiser leur répartition et de constater que les composés de certaines familles comme des peroxycétals, les hydroperoxydes, les peroxyacides et les peroxydicarbonates sont bien situés dans la même zone de l'espace.

Tableau 11 : Répartition par famille des peroxydes de la base de données améliorée

	Nombre de peroxydes
Peroxydes dialkyles	13
Peroxydes de diacyles	17
Hydroperoxydes	8
Peroxiacides	6
Peroxyesters	34
Peroxycétals	12
Peroxydicarbonates	13
Peroxydes de sulfonyles	1
Peroxydes de silyles	1
Tous	105

Un long travail de recherche de données a été effectué pour trouver les structures des 105 peroxydes organiques et leurs numéros CAS. La géométrie a ensuite été optimisée pour tous ces peroxydes organiques en utilisant le logiciel gaussian09[8] avec la méthode DFT (PBE0//6-31+G(d,p)).

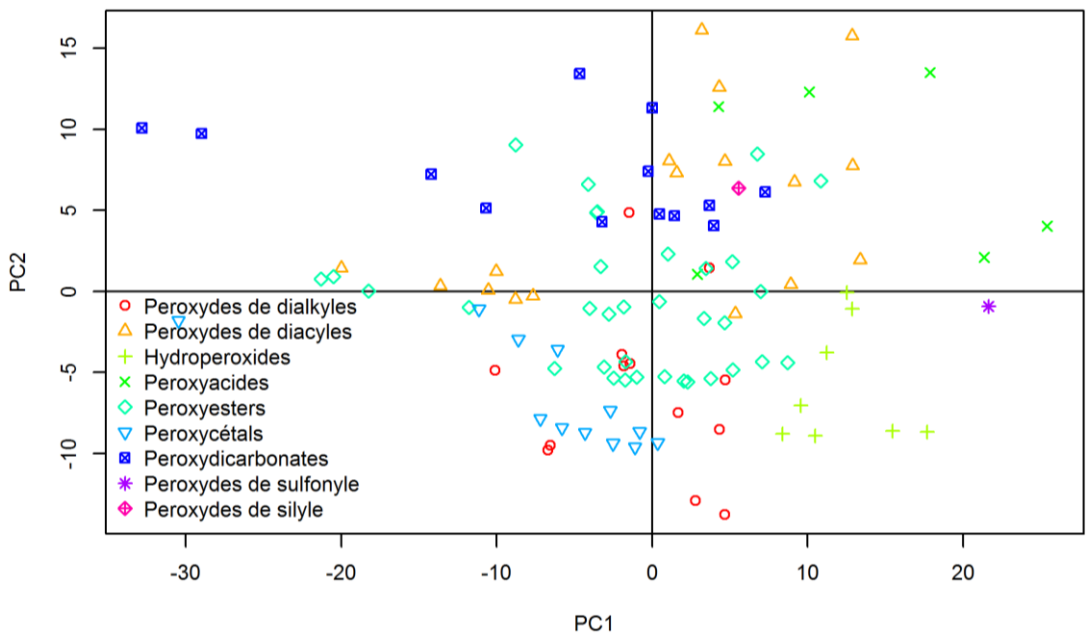


Figure 28: Représentation des 105 peroxydes organiques par famille dans l'espace des descripteurs par PCA

Des modèles QSPR ont été développés, parmi les propriétés disponibles, pour celles qui sont les plus utilisées pour la classification des peroxydes organiques (selon le diagramme en annexe I) ainsi que celles dont les mesures peuvent être effectuées à l'INERIS ont été sélectionnées (voir Tableau 12).

Tableau 12 : Liste des propriétés d'intérêts de la Datatop

Nom de l'épreuve dans la Datatop	Type d'épreuves
A.1 Tube BAM [cm]	détonation
C.1 TPT [ms]	déflagration
C.2 Deflagration test [mm/s]	déflagration
E.1 Koenen [mm]	chauffage sous confinement
E.2 DPVT [mm]	chauffage sous confinement
F.3 Trauzl [cc/10g]	pouvoir explosif
H. TDAA [°C]	température de décomposition auto-accélérée
3(a)(ii) BAM Fallhammer [J]	sensibilité à l'impact
3(b)(I) BAM Friction Apparatus [N]	sensibilité à la friction

Le Tableau 13 recense les données disponibles par propriété en fonction des concentrations (en masse) maximales ou de 100%.

Tableau 13: Nombre de données expérimentales par propriété de la Datatop avant traitement

Epreuve	Données pour des concentrations maximales *(/116)*	Données pour des concentrations de 100% *(/24)*
A.1 Tube BAM [cm]	84	16
C.1 TPT [ms]	85	15
C.2 Deflagration test [mm/s]	80	18
E.1 Koenen [mm]	108	22
E.2 DPVT [mm]	101	19
F.3 Trauzl [cc/10g]	81	17
H. TDAA [°C]	96	21
3(a)(ii) BAM Fallhammer [J]	80	16
3(b)(I) BAM Friction Apparatus	29	9

Les différentes épreuves, dont les résultats sont fournis dans la Datatop et utilisés pour le développement de modèles, sont décrites plus tard dans ce chapitre.

II. PROPRIÉTÉS EXPÉRIMENTALES SÉLECTIONNÉES

Les propriétés décrites dans ce chapitre sont mesurées selon le Manuel d'épreuves et de critères[3] de l'ONU. Les essais seront présentés dans l'ordre du tableau : de la série A à H. Néanmoins, tous les essais ne sont pas nécessaires pour classer une substance ou un mélange et ne se font pas forcement en partant de la « case 1 » pour parvenir au type. En effet, certains essais comme l'épreuve A1 sont chers, compliqués à mettre en place et ne sont pas forcément nécessaires.

1. Épreuve C1

Les épreuves de la série C répondent à la question « La déflagration s'y propage-t-elle ? ». L'épreuve de pression/temps sert à « déterminer l'aptitude d'une matière sous confinement à propager une déflagration ». Le dispositif est constitué par une bombe cylindrique en acier à laquelle sont accolés un transducteur de pression et un système de mise à feu (voir Figure 29). Un échantillon de 5 g de matière est introduit dans la bombe de manière à toucher le dispositif d'allumage. Le transducteur permet la mesure de la pression. La grandeur mesurée est le temps nécessaire pour que la pression passe de 690 kPa à 2070 kPa (pression manométrique, différence entre la pression d'un fluide et celle atmosphérique au lieu de mesurer la pression par rapport à une pression nulle).

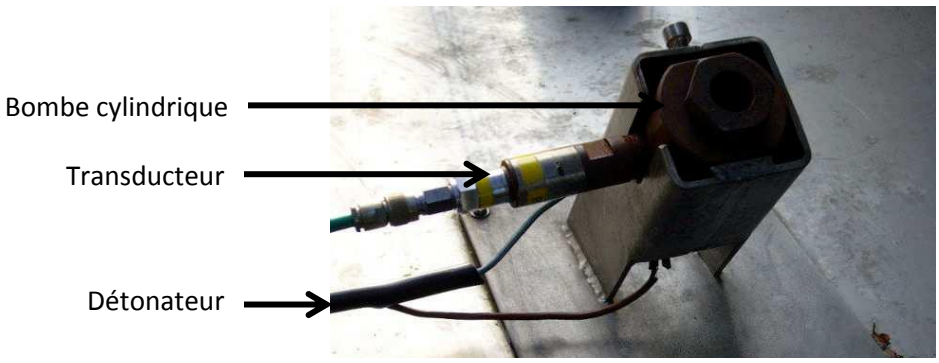


Figure 29: Montage temps/pression – appareillage INERIS

Trois essais sont effectués et le temps le plus court est retenu pour le classement. Pour répondre à la question initiale, les critères suivants sont utilisés :

- Temps de montée de 690 kPa à 2070 kPa inférieure à 30 ms : « Oui, rapidement » ;
- Temps de montée de 690 kPa à 2070 kPa égale à 30 ms ou plus : « Oui, lentement » ;
- Pas de montée de 690 kPa à 2070 kPa : « Non ».

2. Épreuve C2

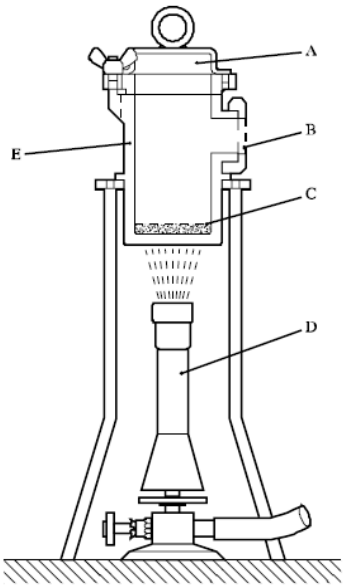
L'épreuve C2 sert à « déterminer l'aptitude d'une matière à propager une déflagration ». L'épreuve est exécutée dans un vase de Dewar avec fenêtres d'observation d'un volume d'environ 300 cm^3 et d'une hauteur de 180 à 200 mm. La vitesse de déflagration (mm/s) est le temps que prend la décomposition pour se propager entre deux repères à 50 mm et 100 mm.

Pour répondre à la question initiale, les critères suivants sont utilisés :

- Vitesse de propagation supérieure à 5,0 mm/s : « Oui, rapidement » ;
- Vitesse de propagation entre 0,35 et 5,0 mm/s : « Oui, lentement » ;
- Vitesse de propagation inférieure à 0,35 mm/s ou réaction avant d'attendre le repère inférieur : « Non ».

3. Epreuve E2

La série d'épreuves E mesure le potentiel explosif d'une matière et permet de répondre à la question « Quelle est la réaction au chauffage sous confinement ? ». L'épreuve de la bombe des Pays-Bas (illustrée Figure 30) sert à déterminer la sensibilité des matières à l'effet d'un chauffage intense sous confinement défini. La bombe en acier inoxydable contenant un échantillon (10 à 50 g) de matière est chauffée avec du butane de qualité industrielle. Plusieurs essais sont effectués et on mesure la taille du disque de lumière pour laquelle le disque de rupture en aluminium est cassé.



(A)	Disque de rupture
(B)	Disque à lumière
(C)	Matière éprouvée (10 g ou 50 g)
(D)	Bec Téclu
(E)	Bombe, diamètre intérieur de 50 mm, hauteur intérieure de 94,5 mm

Figure 30: Epreuve de la bombe des Pays-Bas (figure 25.4.2.1 du manuel d'épreuve et des critères)

La réponse à la question « Quelle est la réaction au chauffage sous confinement ? » est déterminée par la valeur du diamètre limite du disque de lumière du pour lequel le disque de rupture est cassé. La réaction peut être « violente », « modérée », « faible » ou « nulle ».

- « violente » : Rupture du disque pour un orifice de 9,0 mm ou plus et un échantillon de 10,0 g ;
- « modérée » : Pas de rupture du disque pour un orifice de 9,0 mm mais rupture pour un orifice de 3,5 mm ou plus et un échantillon de 10,0 g ;
- « faible » : Pas de rupture du disque pour un orifice de 3,5 mm ou plus et un échantillon de 10,0 g mais rupture pour un orifice de 1,0 mm ou de 2,0 mm et un échantillon de 10,0 g

  ou rupture pour un orifice de 1,0 mm et un échantillon de 50,0 g ;

- « nulle » : Pas de rupture du disque pour un orifice de 1,0 mm et un échantillon de 50,0 g.

4. Épreuve F3

Les épreuves de la série F mesurent la puissance explosive d'une matière. Dans le cas de l'épreuve de Trauzl, la matière est confinée dans une cavité ménagée dans un bloc de plomb qui est soumis à l'action d'un détonateur (voir Figure 31). La puissance explosive est exprimée en relation avec la dilatation de la cavité (en cm^3 pour 10 g).

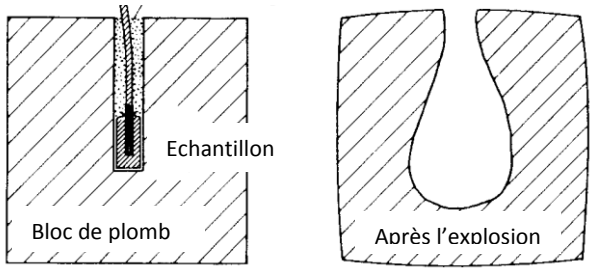


Figure 31: Schéma de l'épreuve de Trauzl9

Pour répondre à la question « Quelle est sa puissance explosive ? », la dilatation du bloc de plomb est mesurée :

- Dilatation supérieure ou égale à 25 cm^3 pour 10 g d'échantillon : « Réaction significative » ;
- Dilatation inférieure à 25 cm^3 mais égale ou supérieure à 10 cm^3 pour 10 g d'échantillon : « Réaction faible » ;
- Dilatation inférieure à 10 cm^3 pour 10 g d'échantillon : « Réaction nulle ».

5. Épreuve 3(a)(ii)

La série d'épreuve de type 3 comprend 4 types d'épreuve pour déterminer : a) la sensibilité à l'impact, b) la sensibilité au frottement, c) la stabilité de la matière à la chaleur et d) la réaction de la matière à l'inflammation. L'épreuve 3(a)(ii) correspond à l'épreuve du mouton de choc BAM. La sensibilité à l'impact caractérise la tendance du matériau à réagir sous l'effet d'un impact. Cette grandeur mesure l'énergie à partir de laquelle un poids de masse donnée (appelé mouton de choc), lâché sur un échantillon, provoque une réaction avec une probabilité de 50%. Un exemple d'appareillage utilisé à l'INERIS est illustré par la Figure 32.

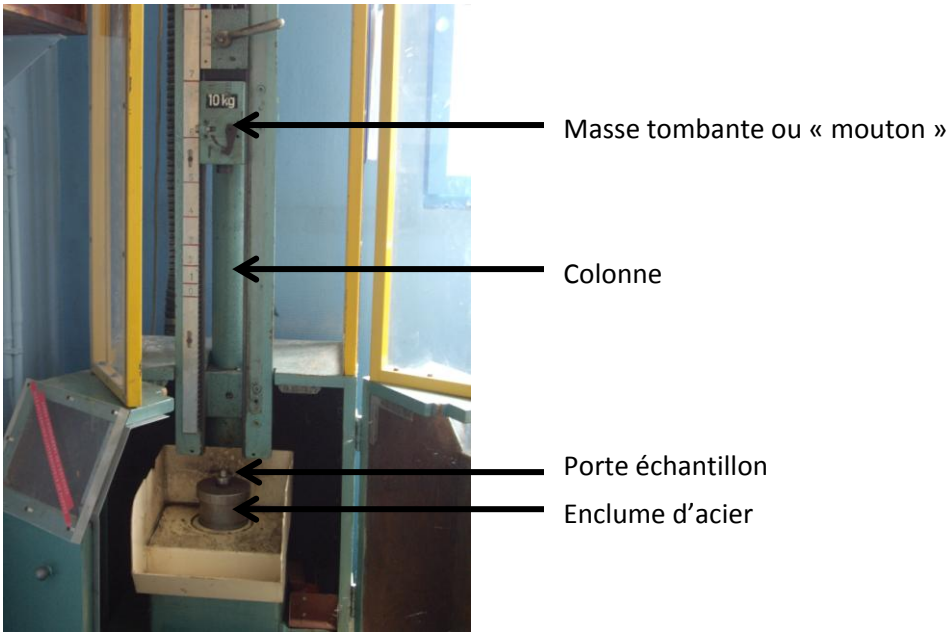


Figure 32: Mouton de Choc – appareillage INERIS

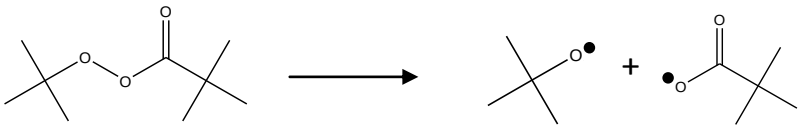
La TDAA ayant déjà été définie dans le chapitre 1 « contexte et objectifs », elle n'est pas redéfinie dans ce chapitre. Toutes ces propriétés sont liées à la décomposition des peroxydes organiques par l'intermédiaire de leur stabilité.

III. ÉNERGIE DE DISSOCIATION

Selon la littérature[10–12], la dissociation de la liaison peroxyde est la première étape de la décomposition des peroxydes organiques. En effet, Benassi a calculé l'énergie nécessaire pour la rupture des différentes liaisons de la molécule et a conclu que la liaison peroxy est la plus faible.

$$ROOR \rightarrow RO\bullet + \bullet OR$$

Par exemple, pour le tert-butyl peroxypivalate :



L'énergie de dissociation d'une liaison est l'énergie nécessaire pour couper cette liaison. Elle a été calculée pour les 105 peroxydes organiques restant. Des calculs de corrélation de cette énergie avec des propriétés de la Datatop et différents descripteurs liés à la liaison peroxy ont été effectués dans le but de repérer une relation simple (linéaire). En effet, une relation avec un ou plusieurs de ces descripteurs permettrait de confirmer facilement l'implication de la liaison peroxy. De même, une relation entre l'énergie de dissociation et une des propriétés serait un départ intéressant pour le développement de modèles, avec un descripteur interprétable chimiquement.

1. Calcul de l'énergie de dissociation

L'énergie de dissociation se calcule selon la formule suivante :

$$E_{disso} = E_{\text{radical 1}} + E_{\text{radical 2}} - E_{\text{molécule}} \quad \textbf{(4. 1)}$$

Cette énergie a été mesurée en DFT avec la fonctionnelle PBE0 et la base 6-31+G(d,p). Le Tableau 14 résume les valeurs calculées de l'énergie (E) et de l'enthalpie (H) de dissociation des 105 peroxydes organiques. La moyenne et l'écart type ont été calculés afin d'observer les valeurs par famille. Les énergies de dissociation calculées (16 à 46 kcal/mol) correspondent, en ordre de grandeur, à celles acceptées dans la littérature[13] (20 à 50 kcal/mol). La liaison peroxyde est ainsi de faible énergie (en comparaison à 83 kcal/mol pour une liaison carbone-carbone[13,14]) et peut donc se couper facilement, rendant le composé instable. On peut constater que les hydroperoxydes et les peroxyacides se distinguent avec une valeur moyenne d'énergie de dissociation (~43 kcal/mol) supérieure à celle des autres familles. Au contraire, les peroxydicarbonates ont une valeur plus faible, proche de 25 kcal/mol.

Tableau 14 : Moyennes et écarts type de l'énergie et de l'enthalpie de dissociation par famille

en kcal/mol	Nombre	Moyenne E	Ecart type E	Moyenne H	Ecart type H
Tous	105	31,42	5,83	28,41	5,95
Peroxyde de dialkyles	13	31,14	1,87	28,22	1,87
Peroxyde de diacyles	17	30,33	2,47	26,54	2,42
Hydroperoxydes	8	42,52	1,57	39,33	2,17
Peroxyacides	6	42,98	3,33	39,90	4,26
Peroxyesters	34	30,33	3,82	27,93	4,44
Peroxycétals	12	28,67	3,57	25,81	3,64
Peroxydicarbonates	13	25,35	1,28	21,81	1,19
Peroxydes de sulfonyle	1	44,59	-	40,51	-
Peroxydes de silyle	1	28,48	-	24,83	-

Les hydroperoxydes et les peroxyacides ont en commun un groupement hydroxy lié à un atome d'oxygène. On peut remarquer que ces deux familles sont situées dans la même partie (à droite) de la Figure 33 avec les hydroperoxydes en bas et les peroxyacides en haut. La PCA nous permet de voir que les composés de ces familles sont regroupés dans l'espace des descripteurs.

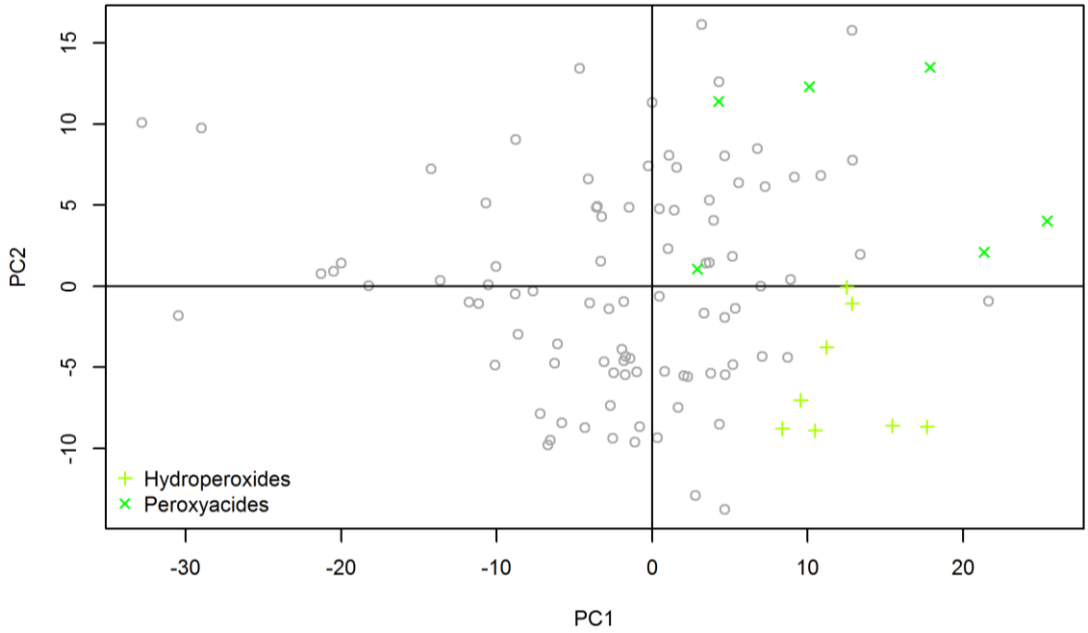


Figure 33 : Représentation des hydroperoxydes et peroxyacides dans l'espace des descripteurs par PCA

La Figure 34 démontre que la présence de ce groupement dans les hydroperoxydes augmente la valeur de leur énergie de dissociation, c'est-à-dire que le groupement hydroxy a un effet stabilisant.

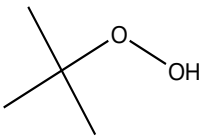
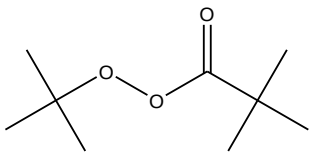
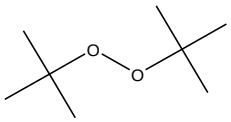
tert-butyl hydroperoxide	tert-butyl peroxypivalate	di-tert-butyl peroxide
		
E_{disso} = 43,50 kcal/mol H_{disso} = 39,52 kcal/mol	E_{disso} = 31,50 kcal/mol H_{disso} = 27,95 kcal/mol	E_{disso} = 33,59 kcal/mol H_{disso} = 30,18 kcal/mol

Figure 34 : Comparaison de l'énergie de dissociation du tert-butyl hydroperoxide en substituant H par un radical alkyl

De même, la Figure 35 confirme cette observation sur un peroxyacide : la substitution de l'hydrogène du groupe hydroxy par des groupements alkyl a un effet stabilisant.

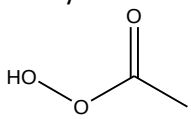
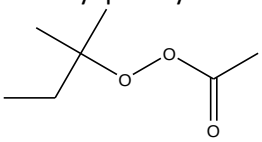
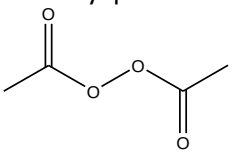
peroxyacetic acid	tert-amyl peroxyacetate	diacetyl peroxide
		
E_{disso} = 39,66 kcal/mol H_{disso} = 37,47 kcal/mol	E_{disso} = 32,87 kcal/mol H_{disso} = 29,06 kcal/mol	E_{disso} = 31,32 kcal/mol H_{disso} = 27,39 kcal/mol

Figure 35 : Comparaison de l'énergie de dissociation du peroxyacetic acid en substituant H par un radical alkyl

2. Relations de l'énergie de dissociation avec les propriétés de la Datatop

L'énergie de dissociation est une grandeur importante pour la décomposition des peroxydes et donc pour leur réaction aux différentes épreuves nécessaires à leur classement. Par conséquent, la relation de celle-ci avec les propriétés mesurées pour le classement a été étudiée. Tout d'abord, des régressions linéaires entre l'énergie de dissociation et les différentes propriétés ont été développées afin de rechercher une relation directe. Une première analyse utilisant les valeurs de propriétés disponibles mesurées pour la concentration maximale (C_{max}) contenue dans la base de données pour chaque molécule a été effectuée. Les propriétés ne sont pas toutes renseignées pour une même molécule, les molécules considérées peuvent être différentes d'une propriété à une autre. Les molécules ayant des résultats aux épreuves non exploitables dans le cadre de notre étude ont été supprimées : la valeur 9999 à l'épreuve C1 est une réponse (« pas de montée en pression au cours du temps ») ou encore la valeur 0 mm/s à l'épreuve C2 (correspond à l'absence de déflagration) mais elles ne peuvent pas être utilisées pour le développement de modèles QSPR. De même, pour les valeurs à l'épreuve 3(a)(ii) de sensibilité à l'impact « >50 » et « >40 » qui donnent une information mais ne sont pas des valeurs quantitatives. Le logarithme de la sensibilité à l'impact a été étudié car cette grandeur a déjà été modélisée avec succès dans le passé[15]. Aucune relation n'a pu être identifiée, que ce soit avec l'énergie ou l'enthalpie de dissociation.

Tableau 15 : Corrélation entre l'énergie de dissociation et les résultats d'épreuves pour Cmax

Epreuve	Nombre de molécules	R^2 (E_{disso})	R^2 (H_{disso})
C1	40 molécules	0,004	0,006
	35 molécules (valeurs <1900 ms)	0,002	0,002
C2	44 molécules	0,021	0,023
	42 molécules (valeurs <26 mm/s)	0,048	0,051
	39 molécules (valeurs <6 mm/s)	0,050	0,055
E2	64 molécules	0,014	0,016
F3	64 molécules	0,001	0,002
	61 molécules (valeurs < 40 cm^3/g)	0,003	0,005
H	61 molécules	0,192	0,216
	58 molécules (valeurs < 90)	0,158	0,160
	58 molécules (sans hydroperoxyde)	0,144	0,170
	53 molécules (sans peroxycétal)	0,297	0,300
	50 molécules (sans hydroperoxyde ni peroxycétal)	0,210	0,213
3(a)(ii)	26 molécules	0,024	0,025
Log[3(a)(ii)]	26 molécules	0,013	0,014

Le Tableau 15 résume les valeurs de coefficient de détermination R^2 obtenues. Des corrélations ont parfois été effectuées en supprimant des molécules ayant des valeurs très supérieures aux autres

molécules pour la même propriété ou des types de famille semblant se démarquer. Par exemple, la moyenne sur 44 molécules pour les valeurs expérimentales de l'épreuve C2 est de 3,90 mm/s alors que certaines de ces molécules ont des valeurs entre 100 et 6 mm/s, ce qui est extrêmement supérieur. Leur suppression réduit à 39 le nombre de molécules pour développer le modèle. De même, pour la TDAA une première corrélation est observée lorsqu'on prend en compte toutes les molécules et une deuxième en ne considérant plus les peroxycétals. Les « meilleures » corrélations sont obtenues pour la TDAA : $R^2(E_{disso}) > 0,1$.

L'analyse de la Datatop (voir Tableau 10) nous permet de dire que la valeur des propriétés peut varier lorsque la concentration change. Cette observation a aussi été faite dans la littérature par Wehrstedt[16] avec le test *Mini Closed Pressure Vessel Test (MCPVT)* pour la caractérisation de l'explosibilité . Des modèles ont été développés en considérant uniquement les molécules ayant des concentrations élevées et homogènes (concentration maximales supérieures ou égales à 75%).

Tableau 16 : Corrélation entre l'énergie de dissociation et les résultats d'épreuves pour Cmax ≥ 75%

Epreuve	Nombre de molécules	R^2 (E_{disso})	R^2 (H_{disso})
C1	20 molécules	0,017	0,011
	18 molécules (valeurs <1900)	0,118	0,110
C2	27 molécules	0,074	0,077
	26 molécules (valeurs <26 mm/s)	0,077	0,077
	23 molécules (valeurs <6 mm/s)	0,327	0,335
	24 molécules (sans hydroperoxyde)	0,072	0,073
	20 molécules (sans hydroperoxyde et valeurs <6 mm/s)	0,302	0,304
E2	36 molécules	0,022	0,022
F3	33 molécules	0,014	0,018
H	35 molécules	0,291	0,288
	32 molécules (sans hydroperoxyde)	0,131	0,130
3(a)(ii)	15 molécules	0,235	0,239
	14 molécules (valeurs <40 cm)	0,127	0,130
	12 molécules (valeurs <10 cm)	0,388	0,379
Log[3(a)(ii)]	15 molécules	0,177	0,181
	12 molécules (H_{50} < 10 cm)	0,356	0,346

Le Tableau 16 résume les résultats obtenus. Les corrélations sont plus élevées que précédemment, ce qui confirme l'idée que la concentration est une variable importante. La sensibilité à l'impact donne cette fois des coefficients R^2 légèrement plus élevés que celle de la TDAA mais pour un nombre de molécules deux fois moins grand (35< molécules pour la TDAA et 15 pour la sensibilité à l'impact). La TDAA est encore la propriété présentant les meilleures corrélations. Cependant, elle dépend de l'emballage du peroxyde lors de la mesure de l'épreuve et pas uniquement de la

structure : le développement de modèle type QSPR semble compliqué puisque ces modèles sont basés sur la structure des molécules uniquement.
L'énergie de dissociation, bien que caractéristique de la décomposition des peroxydes, ne présente aucune relation linéaire avec les propriétés de la Datatop. Ainsi une étude de la relation entre les descripteurs de la liaison peroxy et l'énergie de dissociation a été effectuée.

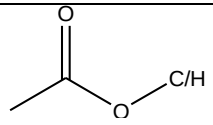
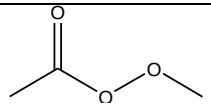
3. Relation de l'énergie de dissociation avec des descripteurs liés à la chimie des peroxydes

Le Tableau 17 résume les performances des régressions linéaires qui ont été obtenues entre l'énergie de dissociation et différents descripteurs pouvant être reliés à la réactivité des peroxydes, notamment à la liaison OO mais aussi aux descripteurs de la DFT conceptuelle (chapitre 2).
Lorsqu'on considère tous les peroxydes organiques, les meilleures régressions linéaires sont obtenues avec le nombre de groupements OOH (n_{OOH}, R^2=0,56) et la moyenne des charges de la liaison peroxyde (Q_{OO}, R^2=0,46). Ces corrélations ne sont pas très élevées contrairement à ce qui avait pu être observé pour les composés nitroaromatiques[17] où l'indice d'électrophilicité ainsi que l'affinité électronique ont pu être corrélés à l'énergie de décomposition expérimentale. Ces travaux avaient été effectués sur un jeu de 22 molécules, ainsi le calcul des corrélations précédent en classant les peroxydes organiques par famille (jeu de taille plus réduite et focalisée sur un type précis de peroxydes) a été effectué dans le but d'obtenir de meilleures performances.

Pour le descripteur Qoo, les performances diminuent drastiquement (proches de zéro) sauf pour la famille des hydroperoxydes (R^2=0,33 pour 8 molécules) et la famille des peroxyacides pour laquelle la corrélation augmente (R^2=0,59 pour 6 molécules). Ces deux familles ont comme points communs : un petit nombre de molécules, une valeur élevée de l'énergie de dissociation (Tableau 14) et la présence de liaison hydroperoxy (n_{OOH}). Ce dernier descripteur est celui qui présente, après Q_{oo}, la meilleure corrélation lorsqu'on considère toutes les molécules (R^2=0,56). La valeur de n_{OOH} est différente de zéro uniquement pour les hydroperoxydes (R^2=0,34) et les peroxyacides (R^2=0,48).

Les corrélations obtenues avec les hydroperoxydes sont peu élevées : la meilleure est R^2=0,40 pour la distance entre les deux atomes O de la liaison peroxy (d_{OO}). La famille de peroxydes de dialkyles (17 molécules) ne présentent aucune corrélation avec tous ces descripteurs. Le coefficient de détermination le plus élevé a été obtenu pour l'angle de torsion (R^2=0,23). De même, les peroxydicarbonates (13 molécules) atteignent leur meilleures corrélations avec l'électronégativité (R^2=0,26) et le moment dipolaire (R^2=0,26).

Tableau 17 : Corrélation entre E_{disso} et quelques descripteurs

Définition	Tous	Peroxyde de dialkyles	Peroxyde de diacyles	Hydroperoxyde	Peroxyacide	Peroxyester	Peroxycétal	Peroxydicarbonate
Nombre de molécules	105	13	17	8	6	34	12	13
Fréquence de l'élongation de la liaison peroxy en cm^{-1}	0,01	0,08	0,20	0,17	0,03	0,10	0,18	0,23
Distance entre les deux atomes O de la liaison peroxy en Å	0,02	0,09	0,12	0,40	0,54	0,03	0,15	0,07
Charge (Mulliken) moyenne des deux atomes O de la liaison peroxy	0,46	0,01	0,00	0,33	0,59	0,09	0,02	0,12
Valeur minimale de l'angle ROO de la liaison peroxy et l'atome voisin	0,14	0,01	0,53	0,06	0,26	0,01	0,28	0,02
Valeur maximale de l'angle ROO de la liaison peroxy et l'atome voisin	0,00	0,01	0,64	0,02	0,39	0,01	0,61	0,09
Angle de torsion ROOR (valeur entre -180° et 180°)	0,00	0,23	0,00	0,18	0,00	0,06	0,01	0,06
Angle de torsion ROOR (valeur entre 0° et 360° - valeur positive)	0,00	0,23	0,00	0,18	0,00	0,06	0,01	0,06
Valeur absolue de l'angle de torsion ROOR	0,12	0,17	0,18	0,29	0,70	0,00	0,77	0,02
Nombre de groupements 	0,01	0,01	0,06	X	X	0,04	0,29	X
Nombre de groupements 	0,00	0,00	0,32	X	0,48	0,02	X	0,03
Nombre de groupements peroxy	0,01	0,05	0,06	0,34	0,48	0,02	0,20	X
Nombre de groupements hydroperoxy	0,56	X	X	0,34	0,48	X	X	X
Nombre de groupements peroxy sans considérer les hydroperoxy	0,17	0,05	0,06	X	X	0,02	0,20	X
Energie de l'orbitale HOMO en u.a	0,00	0,18	0,67	0,02	0,25	0,02	0,04	0,15
Energie de l'orbitale LUMO en u.a	0,00	0,01	0,64	0,08	0,03	0,03	0,51	0,00

Définition	Tous	Peroxyde de dialkyles	Peroxyde de diacyles	Hydroperoxyde	Peroxyacide	Peroxyester	Peroxycétal	Peroxydicarbonate
Dureté : $\eta = \varepsilon_{\mathrm{HOMO}} - \varepsilon_{\mathrm{LUMO}}$ en u.a	0,00	0,14	0,71	0,03	0,06	0,02	0,30	0,09
Electronégativité : $\chi_M = -\frac{\varepsilon_{\mathrm{HOMO}} + \varepsilon_{\mathrm{LUMO}}}{2} = -\mu$ en u.a	0,00	0,01	0,36	0,22	0,00	0,04	0,08	0,26
Indice d'électrophilicité $\omega = \frac{\mu^2}{2\eta}$ en u.a	0,00	0,00	0,60	0,11	0,03	0,03	0,34	0,09
Moment dipolaire	0,00	0,00	0,07	0,22	0,50	0,03	0,38	0,26
Polarisabilité principale : $\alpha = \frac{1}{3}\left(\alpha_{xx} + \alpha_{yy} + \alpha_{zz}\right)$ en Å^3	0,13	0,11	0,00	0,06	0,89	0,01	0,00	0,01
Polarisabilité anisotropique : $\Delta\alpha = \left(\frac{1}{2}\right)^2\left[\left(\alpha_{xx} - \alpha_{yy}\right)^2 + \left(\alpha_{xx} - \alpha_{ZZ}\right)^2 + 6\left(\alpha_{xy}^2 + \alpha_{xz}^2 + \alpha_{yz}^2\right)\right]^{1/2}$	0,04	0,00	0,07	0,19	0,78	0,00	0,10	0,01

La base de données contient également un peroxyde de sulfonyle est un peroxyde de silyle.

Les peroxydes de diacyles présentent de bonnes corrélation avec l'angle ROO (R^2=0,53 et R^2=0,64) ainsi qu'avec les descripteurs liés au énergies des orbitales HOMO (E_{HOMO}) et LUMO (E_{LUMO}) : R^2=0,67 pour E_{HOMO}, R^2=0,64 pour E_{LUMO} et R^2=0,71 pour la dureté. Les peroxyacides sont ceux qui présentent le plus de corrélations élevées : R^2=0,70 avec les valeurs absolues de l'angle de torsion ROOR, R^2=0,59 pour Q_{oo} et 0,54 pour d_{OO}. La famille des peroxycétals présente une corrélation élevée pour : l'angle ROO maximal (R^2=0,61) et l'énergie de l'orbitale LUMO (R^2=0,51). La famille des peroxyesters (34 molécules) ne montre aucune corrélation avec ces descripteurs. La corrélation la plus élevée est obtenue avec la fréquence de l'élongation de la liaison peroxy (R^2=0,10). Toutes ces corrélations ne permettent pas de faire des prédictions. Des modèles multilinéaires ont donc été développés afin d'obtenir des modèles performants.

IV. DÉVELOPPEMENT DE MODÈLES QSPR

Cette partie présente les résultats obtenus lors du développement de modèles QSPR en utilisant les données de la Datatop. Les modèles ont été développés selon la méthode présentée dans le chapitre 3 : optimisation de la géométrie par DFT, puis calcul des descripteurs (plus de 300) avec CodessaPro[18] et enfin sélection des descripteurs par BMLR[19]. Dans un premier temps les modèles seront développés sur toutes les molécules disponibles pour chaque propriété. Dans un second temps, les modèles seront développés pour la famille des peroxyesters uniquement car elle est la plus présente dans la Datatop (Tableau 11) et il y a suffisamment de molécules pour entraîner des modèles (34 molécules). Même s'ils ne seront pas validés par un jeu de validation, ils seront soumis aux différentes méthodes de validation interne (validation croisée, Y-scrambling).

1. Toutes les familles de peroxydes organiques confondues

Plusieurs jeux de molécules ont été testés : les molécules ayant une concentration supérieure ou égale à 75% et toutes les molécules en considérant leur concentration maximale disponible. Le Tableau 18 récapitule les relations obtenues et leurs performances.

Tableau 18: Performances des modèles développés avec les données de la Datatop

Epreuve	Nombre de molécules	Nombre de descripteurs	R^2	MAE	MAE(%)	Q^2_{LOO}
C.1	40	2	0,32	529	2734%	0,21
	20 (C≥75%)	2	0,51	334	371%	0,24
	20 (C≥75%)	3	0,70	272	674%	0,47
C.2	44	3	0,94	2,28	X	0,72
	38 (et C.2<6 mm/s)	5	0,80	0,48	X	0,69
	40 (et C.2<4 mm/s)	3	0,62	0,40	X	0,56
	27 (C≥75%)	5	0,99	0,45	X	0,98
	24 (C≥75% et C2<6 mm/s)	2	0,45	0,39	X	0,32
F.3	64	2	0,19	8,0	55%	0,06
	62 (F.3<50 cc/10g)	2	0,14	6,1	50%	0,05
	33 (C≥75%)	3	0,57	3,9	24%	0,44
H	61	3	0,63	14	X	0,59
	35 (C≥75%)	3	0,67	11	X	0,59
3(a)(ii)	26	4	0,61	6,2	50%	0,40
	15 (C≥75%)	2	0,48	6,5	86%	0,10
	15 (C≥75%)	5	0,95	2,3	28%	0,85

On observe que les modèles ne présentent pas de bonnes performances, même après suppression de certaines données en fonction des valeurs expérimentales (suppression des valeurs aberrantes comme expliqué dans le paragraphe « Relations de l'énergie de dissociation avec les propriétés de la Datatop »). Par exemple, le modèle pour la propriété C2 avec l'ensemble des données (44 molécules) présente un bon R^2 mais Q^2_{LOO} est faible, en particulier au regard de la valeur élevée du coefficient R^2.

En effet, ces deux coefficients doivent avoir des valeurs proches, en plus d'une valeur « élevée », pour que le modèle soit considéré comme robuste (voir le chapitre 3 : « Principe et méthodes des modèles QSPR »). Dans le cas de la suppression de ces valeurs aberrantes, les performances en R^2 et Q^2 diminuent mais la différence entre les deux coefficients se réduit. De plus, le calcul de la MAE montre que l'erreur diminue avec la suppression de 4 molécules. Il faut cependant noter que le nombre de descripteurs dans le modèle n'est pas le même : la valeur des coefficients sont à prendre au regard du nombre de descripteurs puisque l'augmentation de ce dernier implique une amélioration des performances d'ajustement mais pas nécessairement celles en prédiction.

Les modèles développés avec uniquement les peroxydes organiques ayant une concentration supérieure ou égale à 75% ont de meilleures performances de manière générale (notamment en MAE). Cela confirme l'influence de la concentration et donc la nécessité de développer des modèles pour des molécules à des concentrations similaires.

2. Peroxyesters uniquement

Des modèles ont été développés par famille pour observer si la diversité de la base de données n'était pas un obstacle à l'obtention de modèle prédictif. Seule la famille des peroxyesters a été considérée dans ce paragraphe car elle seule contient plus de 20 molécules renseignée par propriété. La Figure 36 illustre le fait que cette famille couvre encore une grande partie de l'espace chimique.

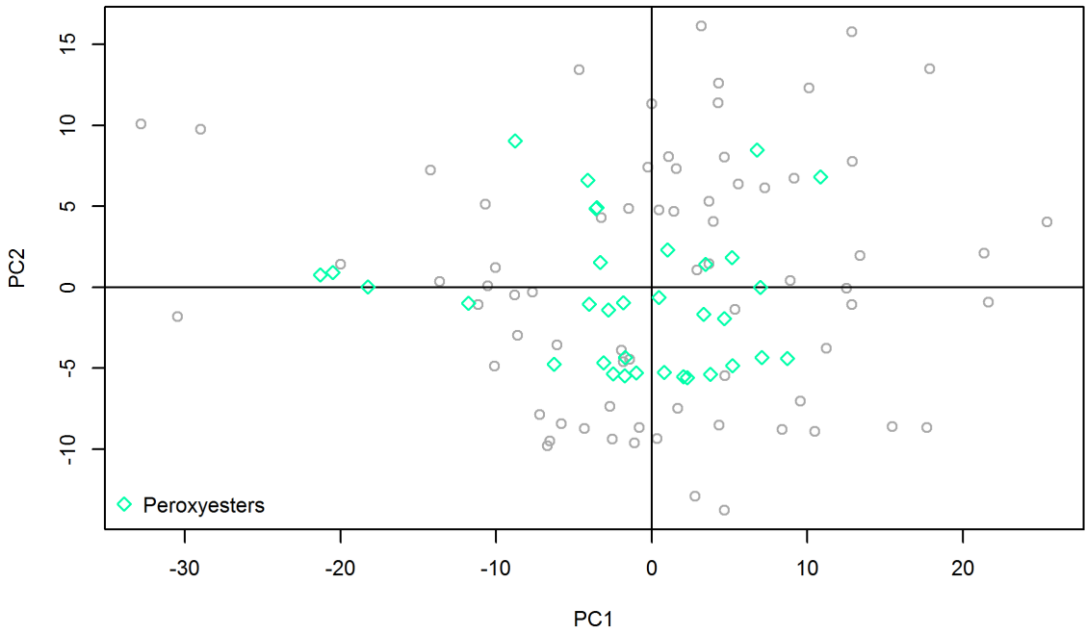


Figure 36 : Représentation des peroxyesters dans l'espace des descripteurs par PCA

La concentration maximale a été considérée de manière à avoir assez de molécules pour entraîner des modèles. Là encore, pour chaque propriété, les molécules ayant des valeurs expérimentales

aberrantes par rapport à l'ensemble des molécules considérées ont été supprimées pour le développement des modèles.

Tableau 19 : Performances des modèles développés avec les peroxyesters de la Datatop

Epreuve	Nombre de molécules	Nombre de descripteurs	R^2	MAE	MAE %	Q^2_{LOO}	Q^2_{5cv}	Q^2_{10cv}	R^2_{YS}	σ_{YS}
C.1 (ms)	16	3	0,82	229	889%	0,71	0,66	0,67	0,20	0,13
C.2 (mm/s)	20	3	0,73	0,76	419%	0,58	0,57	0,61	0,16	0,11
F.3 (cc/10g)	18	4	0,88	2	14%	0,81	0,79	0,81	0,23	0,14
H (°C)	22	4	0,93	5	X	0,87	0,76	0,86	0,19	0,11
3(a)(ii) (J)	6	-	-	-	-	-	-	-	-	-

Quelle que soit la propriété modélisée, les résultats (voir Tableau 18 et Tableau 19) sont meilleurs lorsqu'on ne considère que les peroxyesters. Cependant, il n'y a pas assez de molécules dans la Datatop pour faire une validation externe avec un jeu de validation. La validation croisée donne des performances plutôt intéressantes, par rapport aux valeurs de R^2, même si pour l'épreuve C2 elle n'est pas concluante. La validation par Y-scrambling est plus complexe : selon le critère de Rücker[20], le modèle est validé (R^2-$R^2_{YS} > 2{,}3\sigma_{YS}$) mais l'observation de la répartition des 1000 points obtenus par rapport à la corrélation entre les valeurs expérimentales « vraies » et celles « mélangées » n'est pas toujours convaincante (voir Figure 38, par exemple). Au regard du Tableau 19 uniquement, les modèles pour les propriétés des épreuves F3 et H ont de bonnes performances en ajustement et en Y-scrambling (passent le critère de Rücker). Cependant, l'observation des figures représentant les performances des modèles fortuits, seul le modèle pour la TDAA (épreuve F3) est acceptable.

a) Epreuve C1

À partir de 16 molécules, les résultats de l'épreuve C1 (en ms) permettent de développer un modèle à 3 descripteurs :

$$C1 = -20{,}7\,OB + 79099\,BO_{O,min} - 934\,{}^2CIC_{avg} - 14767 \quad \textbf{(4. 2)}$$

Avec OB la balance en oxygène (t-test= -4,99), $BO_{O,min}$ l'ordre de liaison minimal (>0,1) d'un atome d'oxygène (t-test= 3,07) et ${}^2CIC_{avg}$ l'indice moyen d'information complémentaire d'ordre 2 (t-test=-3,75). Le descripteur le plus important, la balance en oxygène définie par la réglementation TDG, est un critère de pré-sélection des composés explosibles qui se calcule avec l'équation (4. 3).

$$OB = \frac{-1600}{mw}\left[2n_C + \frac{n_H}{2} - n_O\right] \quad \textbf{(4. 3)}$$

Avec mw la masse moléculaire, n_C le nombre d'atomes de carbones, n_H celui d'atomes d'hydrogène et n_O celui d'atomes d'oxygène.

La Figure 37 représente l'application de cette équation sur le jeu d'entraînement ainsi que les résultats de la procédure d'Y-scrambling.

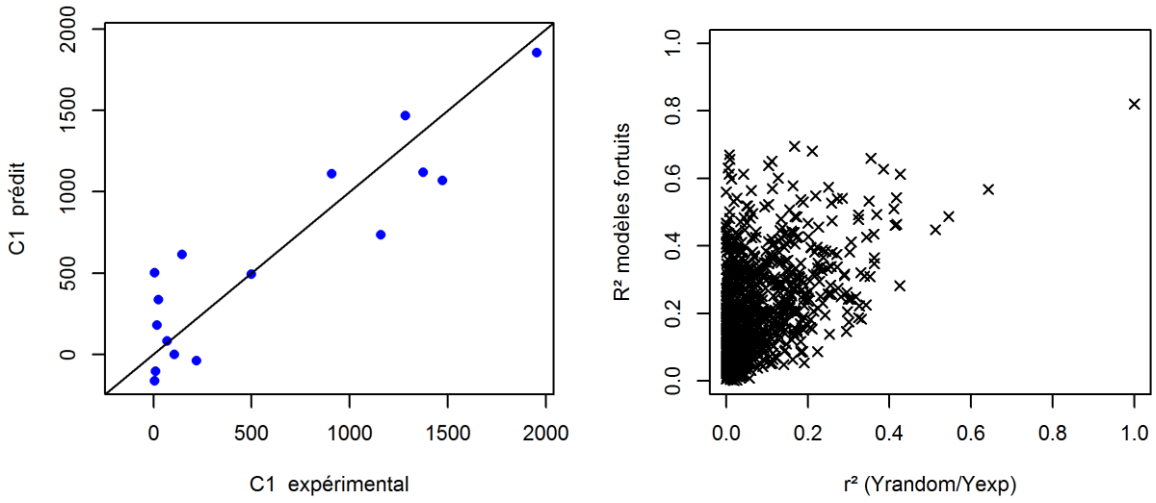


Figure 37 : 1) Valeurs expérimentales vs valeurs prédites par le modèle (4. 2) pour les peroxyesters 2) Résultats du Y-scrambling

L'équation (4. 2) présente de mauvaises performances en ajustement, on ne s'attend pas à une bonne prédictivité puisque l'observation de la Figure 37 montre que la régression est influencée par une molécule (en haut à droite) et que les modèles fortuits présentant les coefficients de détermination d'une valeur supérieure à 0,6.

b) Epreuve C2

À partir de 20 molécules, les résultats de l'épreuve C2 (en mm/s) permettent de développer un modèle à 3 descripteurs :

$$C2 = 73{,}1\ {}^{2}IC_{avg} - 7{,}46R_{C,min} + 7{,}49\ S_{YZ} - 7{,}18 \quad \textbf{(4. 4)}$$

Avec ${}^{2}IC_{avg}$ l'indice moyen d'information d'ordre 2 (t-test=3,63), $R_{C,min}$ l'indice minimal de réactivité de l'atome C (t-test=-4,86) et S_{YZ} un descripteur géométrique qui correspond à la projection de l'ombre de la molécule selon le plan YZ (t-test=3,56).

La Figure 38 représente l'application de cette équation sur le jeu d'entraînement ainsi que les résultats de la procédure d'Y-scrambling avec la représentation des coefficients de détermination des modèles fortuits, certains d'entre eux présentent une valeur supérieure à 0,5. Ce modèle n'est pas utilisable en termes de prédictions. Il n'est pas robuste (Q^2=0,58) et l'équation ne présente pas de bonnes performances en ajustement avec une erreur supérieure à 100%.

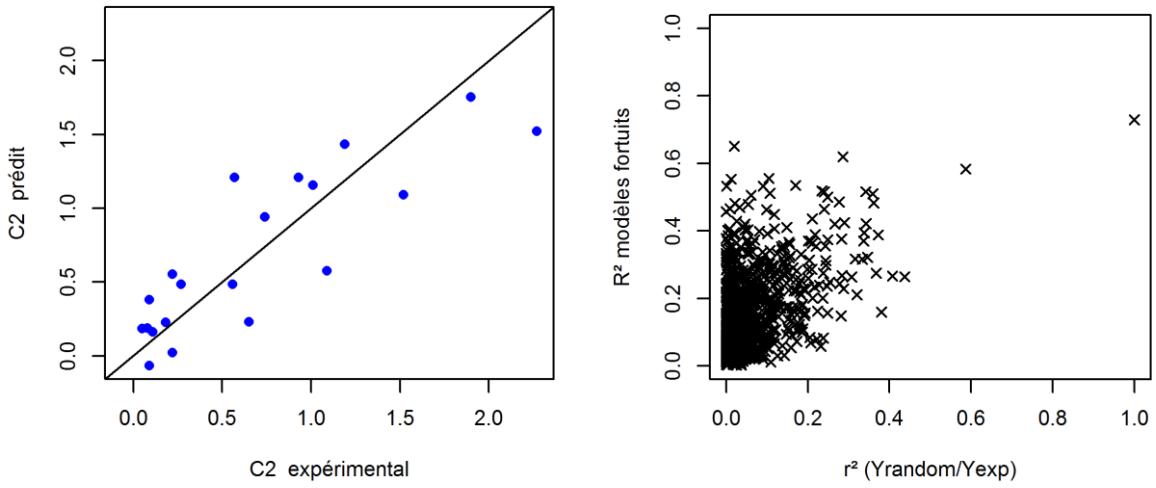


Figure 38: 1) Valeurs expérimentales vs valeurs prédites par le modèle (4. 4) pour les peroxyesters 2) Résultats du Y-scrambling

c) Epreuve F3

À partir de 18 molécules, les résultats de l'épreuve F3 (en cc/10 g), permettent de développer un modèle à 4 descripteurs :

$$F3 = 0,033 D_{ROOR} - 1084\ BO_{H,max} - 264\ P_{Q_{max},Q_{min}} + 10,2\ \mu_h + 1094 \quad \textbf{(4. 5)}$$

Avec D_{ROOR} l'angle de torsion de la liaison peroxy (t-test=5,11), $BO_{H,max}$ l'ordre de liaison maximale pour un atome H (t-test=-5,10), $P_{Q_{max},Q_{min}}$ la différence entre la charge partielle maximale et la charge partielle minimale (t-test=-6,18) et μ_h la composante hybride du moment dipolaire de la molécule (t-test=5,78).

Le descripteur D_{ROOR} est directement lié à la liaison peroxy. Les descripteurs $P_{Q_{max},Q_{min}}$ et μ_h sont reliés à la polarité de la molécule.

Ce modèle possède un risque élevé d'avoir été obtenu par chance. En effet, la Figure 39.b représente les résultats de la procédure d'Y-scrambling avec la représentation des coefficients de détermination des modèles fortuits. Ces modèles ont des performances élevées puisque certains présentent une valeur supérieure $R^2 > 0,7$. Aucune validation externe n'a pu être effectuée. Néanmoins, au vu des performances en validation interne, on s'attend déjà à des mauvaises prédictions. Cette équation ne doit pas être utilisée pour prédire les résultats de l'épreuve F3.

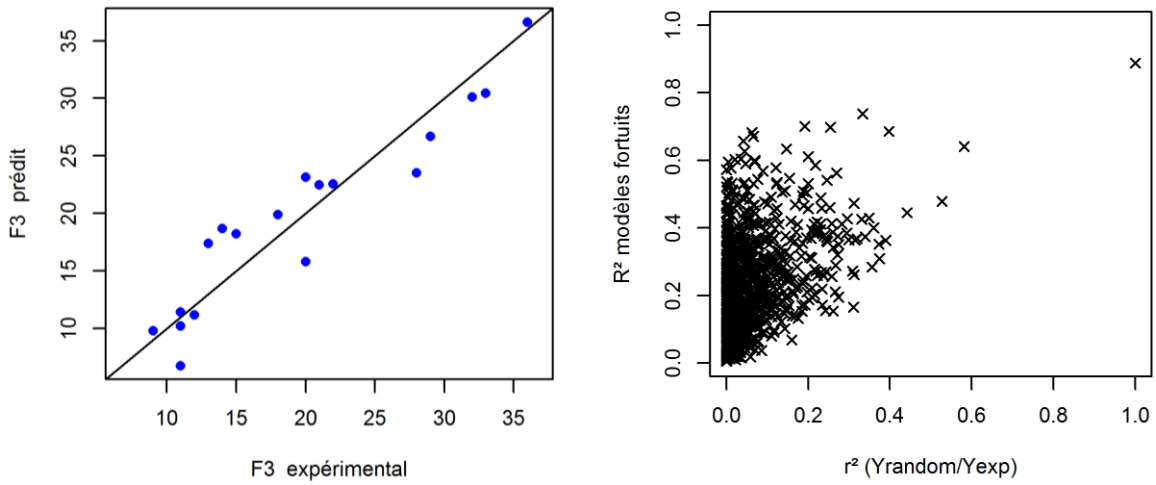


Figure 39 : 1) Valeurs expérimentales vs valeurs prédites par le modèle (4. 5) pour les peroxyesters 2) Résultats du Y-scrambling

d) Epreuve H

À partir de 22 molécules, les résultats de l'épreuve H (TDAA en °C) permettent de développer un modèle à 4 descripteurs :

$$TDAA = 152F_{OO}^{-} + 3,68RPCG + 0,0075W - 19,9E_{LUMO} - 6,34 \quad \textbf{(4. 6)}$$

Avec F_{OO}^{-} la fonction de Fukui localisée sur l'orbitale HOMO de la liaison peroxy (t-test=9,72), RPCG la surface chargée positive (t-test=6,22), W l'indice de Wiener (t-test=4,38) et E_{LUMO} l'énergie de l'orbitale LUMO (t-test=-3,56).

Les descripteurs quantiques F_{OO}^{-} et E_{LUMO} sont tous les deux reliés à la réactivité de la molécule. Plus particulièrement, pour F_{OO}^{-}, à celle de la liaison peroxy. L'indice de Wiener[19,21] est un descripteur topologique qui caractérise la compacité de la molécule.

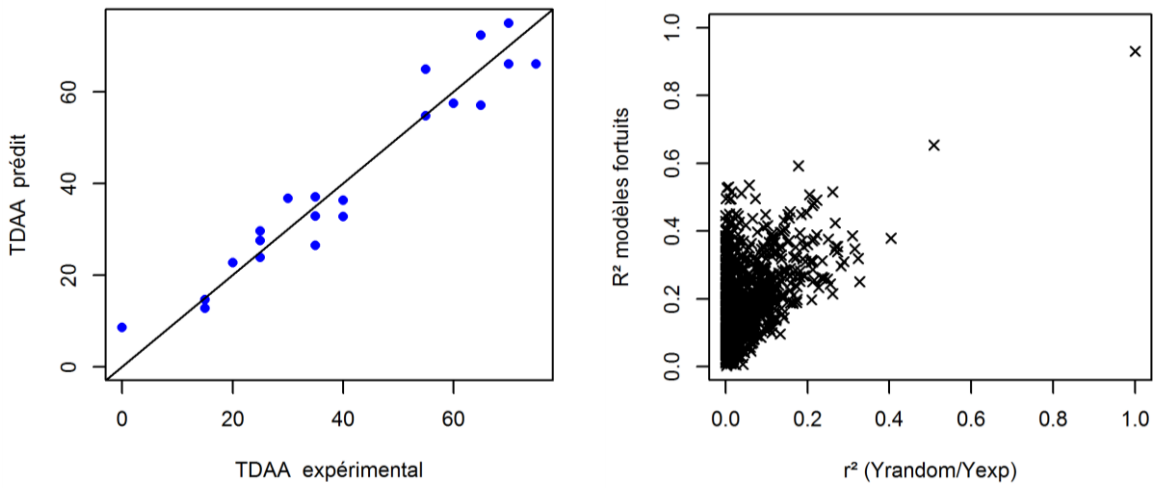


Figure 40: 1) Valeurs expérimentales vs valeurs prédites par le modèle (4. 6) pour les peroxyesters 2) Résultats du Y-scrambling.

Les performances en ajustement sont bonnes pour ce modèles qui présente une erreur de 5°C et une robustesse Q^2=0,87 proche de la valeur de R^2=0,93. La procédure de Y-scrambling (Figure 40.2) donne des résultats moins catastrophiques que les modèles développés sur les résultats des épreuves précédentes. Cette propriété est celle pour laquelle les résultats sont les plus encourageants avec une équation à 4 descripteurs (pour 22 peroxyesters) dont l'erreur est de 5°C.

V. CONCLUSION

L'énergie de dissociation est un paramètre qui a permis l'identification de familles : les peroxyesters et les peroxyacides ayant une valeur élevée de l'ordre de 43 kcal/mol et celle des peroxydicarbonates avec une énergie de dissociation moyenne de 25 kcal/mol. Cependant, aucune relation n'a peu être observée entre cette grandeur et d'autres descripteurs de la liaison peroxy ou bien même avec une propriété expérimentale.

Les modèles QSPR développés avec cette base de données ne sont pas concluants. Cela peut s'expliquer par le fait qu'elle n'est pas homogène : les mesures, bien quelles ait été obtenues selon les protocoles du Manuel d'épreuves et de critères, n'ont pas toutes été faites dans un seul laboratoire. Les données n'ont pas été toutes mesurées sur le même appareil et dans les mêmes conditions. De plus, les concentrations sont très différentes d'un peroxyde organique à un autre, hors il a été vu dans le chapitre 1 « Contexte et objectif » que la dilution permettait de stabiliser le produit et de rendre le transport et le stockage plus facile et plus sûr. L'influence de la concentration dans le développement de modèles QSPR a été identifiée (Tableau 18).

Les modèles QSPR développés pour la famille des peroxyesters uniquement présente une amélioration des performances. Cela peut s'expliquer par un nombre moins important de molécules à considérer lors de l'entraînement (ajustement plus facile à effectuer) mais aussi par une réduction de la diversité de la base de données permettant une description plus facile de celles-ci. Néanmoins, ces modèles restent inutilisables, tant par le manque de validation externe que par leur faibles performances (notamment en validation croisée). Cependant, le modèle pour la TDAA des peroxyesters est encourageant (MAE=5°C, R^2=0,93) tant en validation croisée (Q^2=0,87) qu'en Y-scrambling (Figure 40)

Une base de données expérimentale a été entièrement développée dans le cadre du projet PREDIMOL pour le développement de modèles QSPR pour les peroxydes organiques. Les modèles développés sont présentés dans le chapitre suivant.

VI. RÉFÉRENCE

(1) Datatop, TNO Defence, Security and Safety; Energetic Materials Research Group, Rijswijk, The Netherlands **2005**.

(2) TNO Defence, Security and Safety; Energetic Materials Research Group http://www.tno.nl (accessed May 30, 2013).

(3) *Recommandations relatives au transport des marchandises dangereuses - Manuel d'épreuves et de critères ST/SG/AC.10/11/Rev.5*; Nations Unies, 2010.

(4) Règlement CLP. Règlement (CE) n° 1272/2008 du Parlement Européen et du Conseil du 16 Décembre 2008 relatif à la classification, à l'étiquetage et l'emballage des substances et des mélanges, modifiant et abrogeant les directives 67/548/CEE et 1999/45/CE et modifiant le règlement CE n° 1907/2006.

(5) Whitmore, M. W.; Baker, G. P. Investigation of the use of a closed pressure vessel test for estimating condensed phase explosive properties of organic compounds. *Journal of Loss Prevention in the Process Industries* **1999**, *12*, 207–216.

(6) Yu, Y.; Hasegawa, K. Derivation of the self-accelerating decomposition temperature for self-reactive substances using isothermal calorimetry. *Journal of Hazardous Materials* **1996**, *45*, 193–205.

(7) Milas, N. A.; Golubović, A. Studies in Organic Peroxides. XXV. Preparation, Separation and Identification of Peroxides Derived from Methyl Ethyl Ketone and Hydrogen Peroxide. *J. Am. Chem. Soc.* **1959**, *81*, 5824–5826.

(8) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Montgomery, J. A.; Vreven, T.; Kudin, K. N.; Burant, J. C.; Millam, J. M.; Iyengar, S. S.; Tomasi, J.; Barone, V.; Mennucci, B.; Cossi, M.; Scalmani, G.; Rega, N.; Petersson, G. A.; Nakatsuji, H.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ispida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Klene, M.; Li, X.; Knox, J. E.; Hratchian, H. P.; Cross, J. B.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Ayala, P. Y.; Morokuma, K.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Zakrzewski, V. G.; Dapprich, S.; Daniels, A. D.; Strani, M. C.; Farkas, O.; Malick, D. K.; Rabuck, A. D.; Raghavachari, K.; Foresman, J. B.; Ortiz, J. V.; Cui, Q.; Baboul, A. G.; Clifford, S.; Cioslowski, J.; Stefanov, B. B.; Liu, G.; Liashenko, A.; Piskorz, P.; Komaromi, V.; Martin, R. L.; Fox, D. J.; Keith, T.; Al-Laham, M. A.; Peng, C. Y.; Nanayakkara, A.; Challacombe, M.; Gill, P. M. W.; Johnson, B.; Chen, W.; Wong, M. W.; Gonzalez, C.; Pople, J. A. *Gaussian09*; Gaussian, Inc.: Wallingford CT, 2009.

(9) Meyer, R. *Explosives*; 5th, completely rev. ed.; Wiley-VCH: Weinheim, 2002.

(10) Swern, D.; Swern, D. *Organic Peroxides. Volume 1*; Wiley-Interscience: New York, 1970.

(11) Benassi, R.; Folli, U.; Sbardellati, S.; Taddei, F. Conformational properties and homolytic bond cleavage of organic peroxides. I: an empirical approach based upon molecular mechanics and ab initio calculations. *J. Comput. Chem.* **1993**, *14*, 379–391.

(12) Benassi, R.; Taddei, F. Homolytic bond-dissociation in peroxides, peroxyacids, peroxyesters and related radicals: ab-initio MO calculations. *Tetrahedron* **1994**, *50*, 4795–4810.

(13) Duh, Y.-S.; Hui wu, X.; Kao, C.-S. Hazard ratings for organic peroxides. *Proc. Safety Prog.* **2008**, *27*, 89–99.

(14) Zumdahl, S. S. *Chimie générale*; De Boeck Université; Les Ed. CEC: Paris; [S.l], 1998.

(15) Prana, V.; Fayet, G.; Rotureau, P.; Adamo, C. Development of validated QSPR models for impact sensitivity of nitroaliphatic compounds. *J. Hazard. Mater.* **2012**, *235-236*, 169–177.

(16) Wehrstedt, K. .; Knorr, A.; Schuurman, P. The “Mini Closed Pressure Vessel Test (MCPVT)” as a screening or classification test for explosive properties of organic peroxides. *Journal of Loss Prevention in the Process Industries* **2003**, *16*, 523–531.

(17) Fayet, G.; Joubert, L.; Rotureau, P.; Adamo, C. On the use of descriptors arising from the conceptual density functional theory for the prediction of chemicals explosibility. *Chemical Physics Letters* **2009**, *467*, 407–411.

(18) *CodessaPro*; 2002.

(19) Karelson, M. *Molecular Descriptors in QSAR/QSPR*; Wiley: New York, 2000.

(20) Rücker, C.; Rücker, G.; Meringer, M. y-Randomization and its Variants in QSPR/QSAR. *J. Chem. Inf. Model.* **2007**, *47*, 2345–2357.

(21) Wiener, H. Structural Determination of Paraffin Boiling Points. *J. Am. Chem. Soc.* **1947**, *69*, 17–20.

CHAPITRE 5 - MODÈLES DÉVELOPPÉS À PARTIR D'UNE BASE DE DONNÉES OBTENUE DANS PREDIMOL

La base de données Datatop, utilisée dans le chapitre précédent présente un grand nombre de données mais celles-ci n'ont pas été mesurées dans des conditions expérimentales homogènes : les peroxydes ont des concentrations et des diluants différents. Pour pallier ce problème, des mesures de DSC (*Differential Scanning Calorimetry* soit calorimétrie différentielle à balayage), de densité, de point d'éclair et de sensibilité à l'impact ont été réalisées à l'INERIS et chez Arkema dans le cadre du projet ANR PREDIMOL. Ces données expérimentales ont été enregistrées dans des conditions homogènes et la concentration des substances a été mesurée pour chaque échantillon testé. Les mesures ont été effectuées sur des substances avec la concentration la plus élevée possible. Au final, une base de données de 38 peroxydes organiques a été construite afin de développer des modèles QSPR présentés dans ce chapitre.

I. DONNÉES EXPÉRIMENTALES OBTENUES DANS PREDIMOL

Il est difficile de trouver dans la littérature une base de données suffisamment grande pour les propriétés liées à la stabilité thermique des peroxydes organiques. Or, la taille de la base de données est importante car une base de données trop petite rend difficile l'obtention de modèles prédictifs. Les études disponibles dans la littérature sont habituellement réalisées pour un ou deux peroxydes à des concentrations différentes ou avec un contaminant[1–4]. Ando présente une large base de données (environ 800 composés) pour les propriétés liées à la stabilité thermique mais il n'y a parmi eux que neuf peroxydes organiques. La base de données expérimentale la plus grande pour ces propriétés des peroxydes est celle proposée par Lu et Mannan[5] avec seulement 16 peroxydes organiques.

1. Construction de la base de données

Pour pallier la difficulté d'obtenir des données expérimentales fiables, des mesures de caractérisation expérimentale ont été réalisées par l'INERIS et Arkema dans le cadre du projet PREDIMOL afin d'obtenir une base de données robuste.

À notre connaissance, il n'existe pas de règle officielle à propos du nombre de molécules nécessaire pour développer un modèle. Peduzzi[6] estime que pour avoir des estimations correctes, le nombre minimal observations par descripteurs doit être compris entre 10 et 15. Dans le cas d'un modèle à 3 ou 4 descripteurs, une base de données d'environ 30 ou 40 molécules serait donc suffisante. En 2002, Golbraikh et Tropsha[7] considèrent que le jeu de validation doit être composé d'un minimum de 5 molécules, représentant l'intervalle complet des descripteurs et de la propriété. En 2011, Puzyn[8] recommande un minimum de 10 composés dans le jeu de validation afin que les performances ne soient pas trop fortement perturbées par le mode de partage de la base de données. Ainsi une base de données de 30 molécules semble être un minimum pour effectuer une validation externe, dans le cas d'un partage des données avec un tiers des données dans le jeu de validation. Par extension, un minimum de 20 molécules est nécessaire pour une validation interne uniquement.

Les 38 peroxydes organiques de cette base de données ont été sélectionnés en fonction de leur dangerosité, de leur disponibilité en Europe et de leur transportabilité. Elle est constituée de mesures de DSC effectuées sur 31 peroxydes organiques par le partenaire Arkema afin d'obtenir des valeurs de chaleur de décomposition, température de début de décomposition, température maximale du pic de décomposition, de point d'éclair et de densité. Des mesures complémentaire de DSC ont été effectuées à l'INERIS sur 7 peroxydes organiques, fournies par Akzo Nobel, afin d'élargir la base de données. La liste des peroxydes de la base de données est disponible en annexe II.

Ces données expérimentales ont été effectuées dans des conditions homogènes afin de limiter au mieux les problèmes liés aux bases de données. Ainsi, la concentration des substances a été mesurée pour chaque échantillon testé et les mesures ont été effectuées avec des substances dont la concentration est la plus élevée possible. L'homogénéité de la base de données a été vérifiée. En effet, des mesures effectuées chez Arkema ont été reproduites à l'INERIS (sur les mêmes échantillons) et les résultats concordent (variation de 11% pour le tert-amylperoxy2-ethylhexyl carbonate et de 2% pour le tert-amyl hydroperoxide). Le Tableau 20 résume le type de familles auxquelles les peroxydes de la base de données obtenue appartiennent.

Tableau 20 : Répartition par famille des peroxydes de la base de données construite dans le cadre du projet PREDIMOL

	Sur 38 peroxydes "PREDIMOL"	31 peroxydes Arkema	7 peroxydes Akzo Nobel
Peroxydes de dialkyles	7	6	1
Peroxydes de diacyles	5	4	1
Hydroperoxydes	4	3	1
Peroxyesters	13	11	2
Peroxycétals	5	5	0
Peroxydicarbonates	4	2	2

À la suite, une campagne de mesure de sensibilité à l'impact a été menée à l'INERIS pour 36 peroxydes organiques. Les mesures ont été effectuées sur les mêmes échantillons que ceux analysés en DSC. Le Tableau 21 récapitule le nombre de données disponible par propriété.

Tableau 21 : Récapitulatif des données obtenues dans PREDIMOL

Propriété	Nombre de peroxydes
Chaleur de décomposition	38
Température onset	38
Température maximale du pic de décomposition	38
Sensibilité à l'impact	36
Densité	30
Point d'éclair	24

2. Calorimétrie différentielle à balayage (DSC)

Lors d'une analyse DSC, un échantillon de quelques milligrammes est chauffé dans un creuset. Le flux thermique absorbé par l'échantillon est mesuré en comparant le flux thermique échangé par la cellule contenant l'échantillon à celui échangé par une cellule jumelle vide, placée dans les mêmes conditions. L'appareil impose à l'échantillon une montée en température définie au préalable (voir Figure 41).

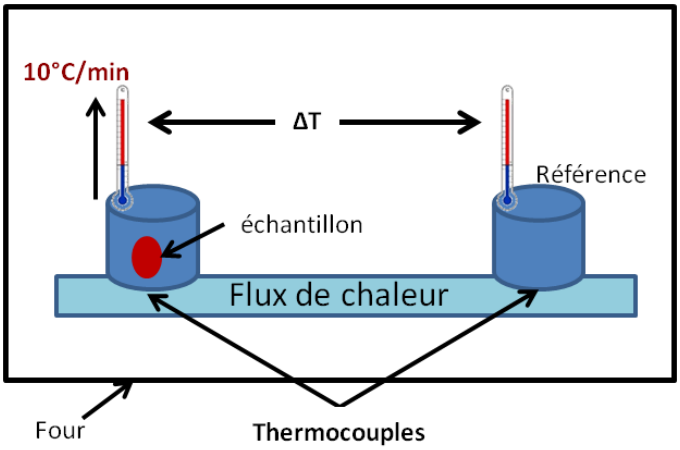


Figure 41: Schéma représentant le principe d'un calorimètre différentiel à balayage

La DSC permet d'acquérir une courbe (appelée thermogramme) représentant l'activité thermique (fusion, réaction, décomposition) d'une substance en utilisant des échantillons de faible volume.

Cette technique d'analyse donne accès en particulier aux informations suivantes : chaleur de décomposition (avec une incertitude de 5 à 10% [9]), température de début et fin de décomposition, température du pic de décomposition… La Figure 42 montre un exemple de thermogramme obtenu.

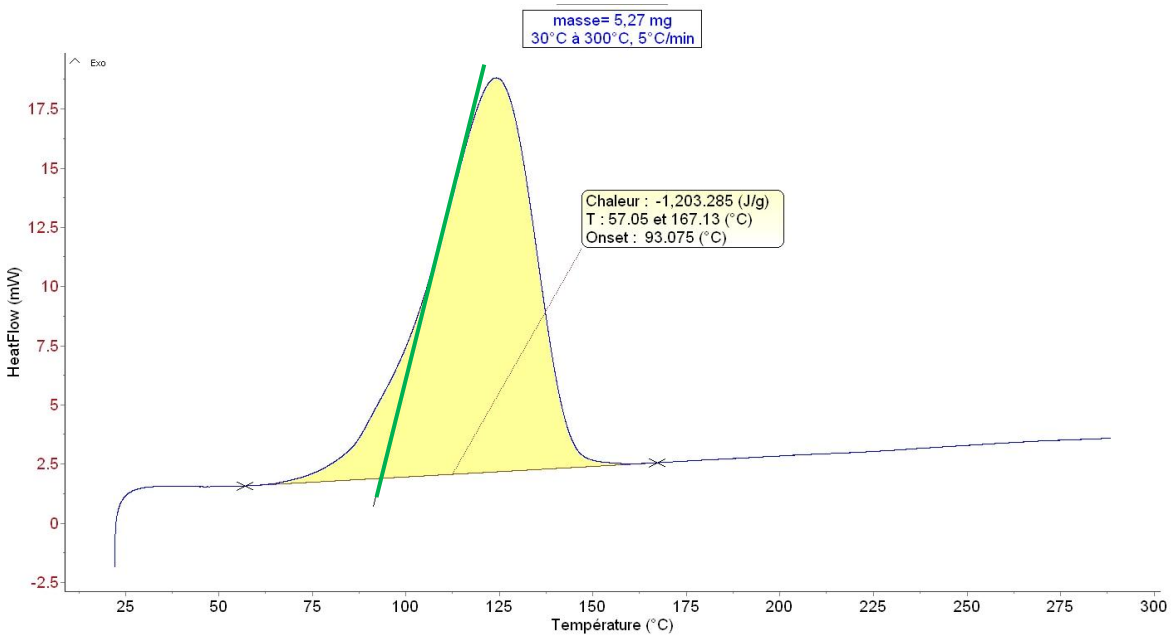


Figure 42: Thermogramme DSC pour le tert-butyl peroxydiethylacetate

Les données ont été mesurées dans les conditions expérimentales suivantes : rampe de chauffe à 5°C/min dans une gamme de température allant de la température ambiante à 250/300 °C.

La chaleur de décomposition correspond à l'intégration de l'aire sous la courbe du pic, c'est à dire la surface en jaune. La température onset est la température à laquelle la tangente à la courbe de montée du pic de décomposition (droite verte sur la Figure 42) coupe la ligne de base (en bleue).

3. Les 38 peroxydes

La liste des 38 peroxydes et leur structure topologique sont disponibles en annexe II.

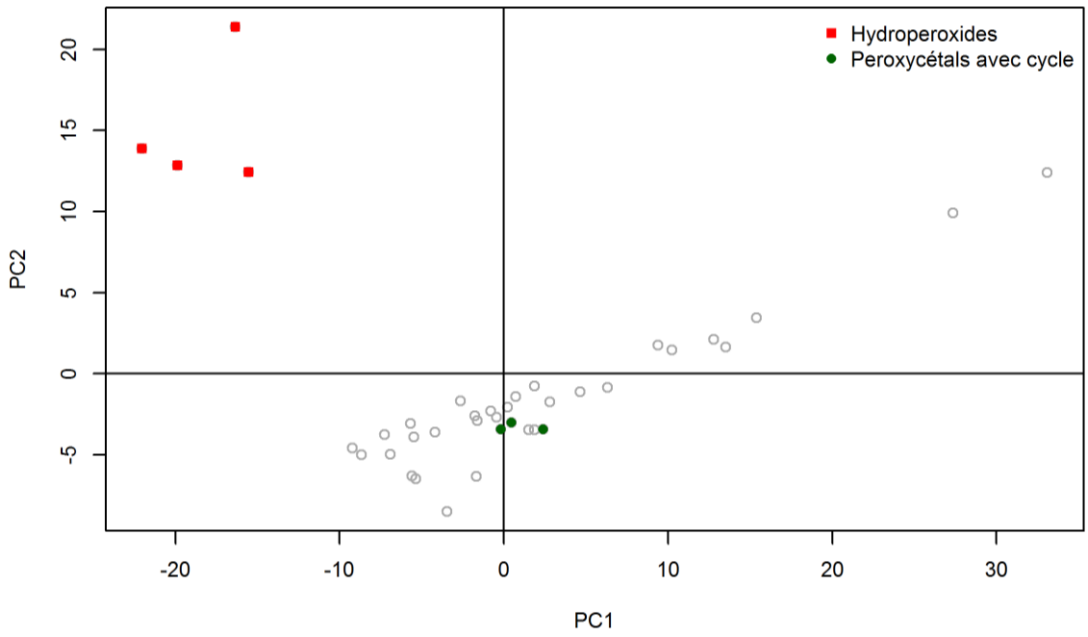


Figure 43: PCA des données PREDIMOL

L'observation de cette liste permet d'identifier deux familles particulières en termes de structure : les peroxycétals avec cycle (3 peroxydes organiques/38) et les hydroperoxydes (4 peroxydes organiques/38). Une analyse en composantes principales (PCA en Figure 43) permet de représenter les différentes molécules dans l'espace des descripteurs. Dans cette représentation, les hydroperoxydes se distinguent très nettement des autres molécules.

Comme dans le chapitre précédent, la corrélation entre les énergies de dissociation calculées et les propriétés renseignées dans la base de données a été calculée dans le but de trouver des relations simples entre les propriétés et la décomposition des peroxydes. Cependant, ces corrélations présentent toujours des valeurs peu élevées (voir Tableau 22), bien que cette fois les concentrations soient connues et aient été choisies pour être les concentrations maximales transportables.

Tableau 22: Corrélation (R^2) entre les propriétés de la base de données PREDIMOL et l'énergie de dissociation (énergie, énergie libre et enthalpie en kcal/mol)

	E_{disso}	G_{disso}	H_{disso}
Tonset (°C)	0,068	0,199	0,091
Tpic (°C)	0,022	0,155	0,036
ΔH (J/g)	0,243	0,080	0,238
Densité (g/cm^3)	0,001	0,002	0,007
Point d'éclair (°C)	0,011	0,030	0,027

Contrairement à ce qu'on pourrait attendre, nous n'avons pas identifié de lien direct entre cette énergie et les propriétés mesurées. Ainsi des modèles plus complexes (multilinéaires) ont été développés pour ces propriétés.

II. PRÉDICTION DE LA STABILITÉ THERMIQUE DES PEROXYDES ORGANIQUES

Comme il a été vu dans le chapitre 1 (contexte et objectifs), les peroxydes organiques sont des composés dangereux et très réactifs. La stabilité thermique est une caractéristique importante que nous avons décidé de modéliser par des modèles QSPR.

1. Modèles QSPR existants

Il existe plusieurs modèles pour la prédiction des propriétés liées à la stabilité thermique de différentes familles de molécules comme les composés nitroaromatiques[10–13], les nitramines[14–16], les liquides ioniques[17] et les polymères [17,18]. Cependant, à notre connaissance, seulement un article existe pour la prédiction des propriétés dangereuses des peroxydes organiques par approche QSPR. Lu et Mannan[5] ont développé des modèles pour la prédiction de la chaleur de décomposition et la température de début de décomposition pour les peroxydes organiques en utilisant une base de données de 16 molécules. Ils ont utilisé deux méthodes pour l'entraînement des modèles : MLR et PLS[19] (régression des moindres carrés partiels). Cependant, ces modèles n'ont pas été validés par un jeu de validation. Les peroxydes de notre base de données n'ayant pas été utilisés pour le développement de ces modèles (23 molécules) l'ont été pour leur validation externe et le calcul de la prédictivité (R^2_{ext}, RMSEP et MAEP).

Tableau 23: Performances des modèles de Lu et Mannan[5]

Propriété	Méthode	Nombre de descripteurs	R^2	Q^2	RMSE	MAE%	R^2_{ext}	*RMSEP*	*MAEP*	*MAEP%*
T_{onset}	MLR	4	0,916	0,108	7	3,64%	0,67	71	62	62,27%
	PLS	13	0,957	0,859	13	3,21%	X	X	X	X
ΔH	MLR	4	0,921	-0,811	43	18,50%	0,06	264	163	55,37%
	PLS	5	0,913	0,828	59	20,88%	0,05	279	164	56,05%

Les performances, résumées dans le Tableau 23, en validation croisée (Q^2=0,108 et 0,811) et externe (MAEP=62°C et 163 kcal/mol) des modèles obtenus par MLR sont très mauvaises (voir Figure 44). Les modèles obtenus par PLS présentent de meilleures performances mais le pouvoir prédictif qui a pu être calculé (seulement pour ΔH) n'est pas élevé. Le modèle PLS pour la température onset n'a pas pu être validé par le jeu de validation externe car certains descripteurs ne sont pas clairement définis dans l'article (« angle ROO » par exemple car les molécules ne sont pas symétriques).

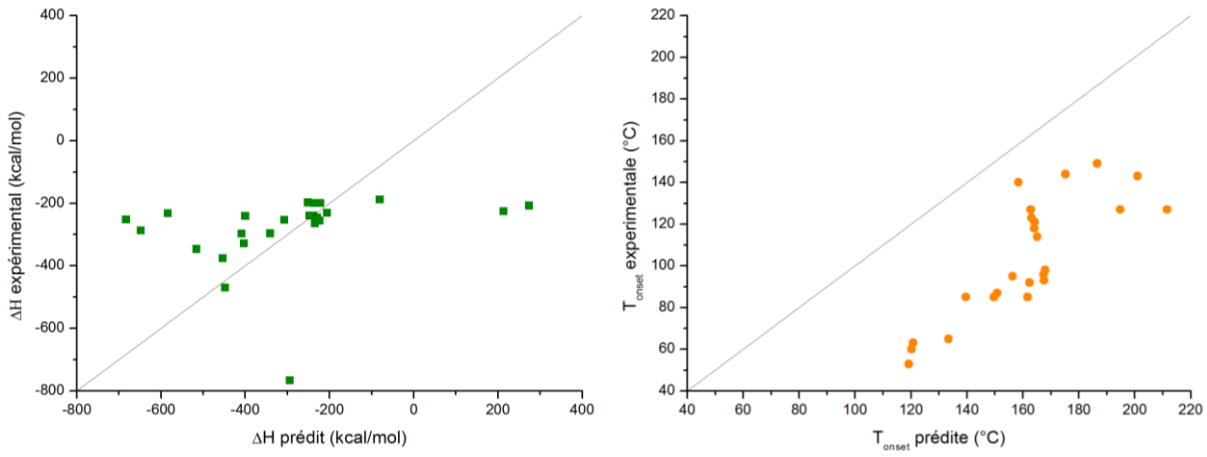


Figure 44 : Représentation des valeurs prédites avec les modèles MLR de Lu pour la chaleur de décomposition et la température onset.

Ainsi, nous nous proposons de développer des modèles validés et prédictifs pour ces deux propriétés mais aussi pour la température maximale du pic de décomposition qui est aussi disponible dans la base de données de PREDIMOL.

2. Modèle pour la chaleur de décomposition

La chaleur de décomposition (en J/g) est l'énergie dégagée lors de la décomposition du produit (surface en jaune sur la Figure 42). Pour la prédiction de cette propriété, seuls 37 peroxydes organiques seront utilisés pour le développement et la validation des modèles. En effet, le 2,5-dimethyl-2,5-dihydroperoxyhexane a une valeur très différente des autres molécules (voir Figure 45) et influence trop fortement les régressions.

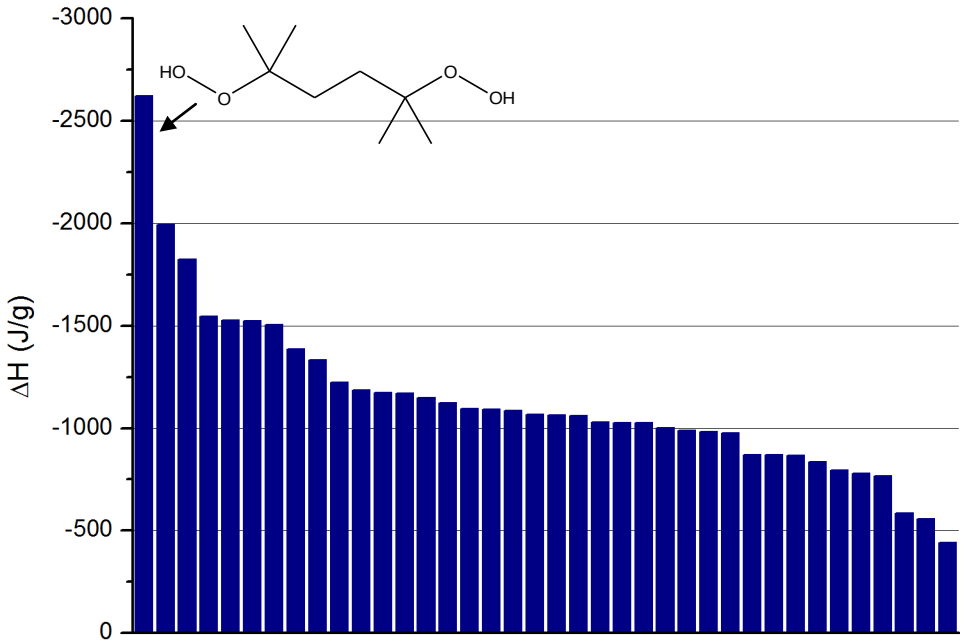


Figure 45: Diagramme des valeurs expérimentales pour la chaleur de décomposition

Des modèles ont été développés pour la chaleur de décomposition avec et sans le descripteur « concentration » parmi plus de 300 descripteurs calculés par le logiciel Codessa. Le Tableau 24 rapporte les performances des modèles obtenus.

Tableau 24: Performances des modèles pour la chaleur de décomposition ΔH (J/g)

Modèles	Nombre de descripteurs	R^2	Q^2	R^2YS	σYS	MAE	R^2_{ext}	MAEP	R^2_{in}	$MAEP_{in}$
ΔH (J/g)	4	0,90	0,82	0,17	0,10	6,29%	0,32	20,74%	0,32	20,74%
ΔH (J/g) – sans C	3	0,81	0,73	0,13	0,10	10,80%	0,05	29,21%	0,13	26,83%

Le meilleur modèle obtenu, par la méthode BMLR expliquée dans le chapitre 3, pour la chaleur de décomposition est un modèle à quatre descripteurs :

(5. 1) $$\Delta H = -874C + 3359Q_{OO,nbo} + 366S_{OO}^{-} + 87S_{rot} - 2475$$

Avec C la concentration du peroxyde (t-test=-4,8), $Q_{OO,nbo}$ la moyenne des charges NBO des atomes d'oxygène de la liaison peroxyde (t-test=-6,0), S_{OO}^{-} la mollesse locale sur l'orbitale HOMO des atomes d'oxygène de la liaison peroxyde (t-test=6,0) et S_{rot} l'entropie de rotation à 300 K (t-test=-3,9). Ce modèle ne présente aucune molécule du jeu de validation hors du domaine d'applicabilité.

Tout d'abord, il est intéressant de remarquer que les deux descripteurs les plus importants de cette équation, $Q_{OO,nbo}$ et S_{OO}^{-}, sont directement liés à la liaison peroxy. Ensuite, la présence du descripteur C (3ème descripteur le plus important avec une valeur de t-test=-4,8) est un fait remarquable qui confirme l'importance (déjà observée dans la Datatop) de ce paramètre pour la stabilité thermique des peroxydes organiques.

D'autre part, le meilleur modèle obtenu pour la chaleur de décomposition, en ne considérant pas la concentration, est un modèle à 3 descripteurs :

(5. 2) $$\Delta H = 592Q_{OO,nbo} + 263,2S_{OO}^{-} - 240\,^2BIC_{avg} + 1768$$

Avec $Q_{OO,nbo}$ la moyenne des charges NBO des atomes d'oxygène de la liaison peroxyde (t-test=-6,9), S_{OO}^{-} la mollesse locale sur l'orbitale HOMO des atomes d'oxygène de la liaison peroxyde (t-test=3,4) et $^2BIC_{avg}$ l'indice moyen d'information liante d'ordre 2 (t-test=-5,7).

Les deux descripteurs $Q_{OO,nbo}$ et S_{OO}^{-} sont encore présents dans l'équation mais le modèle obtenu présente des performances beaucoup moins bonnes. L'effet de la concentration sur la valeur des propriétés est vérifié. C'est pourquoi des modèles ont été développés pour la chaleur de décomposition divisée par la concentration, afin d'introduire cette information dans la propriété.

3. Modèle pour la chaleur de décomposition divisée par la concentration : ΔH/C

En divisant la chaleur de décomposition par la concentration, un modèle à 4 descripteurs est obtenu (équation (5. 3) et Figure 46) :

(5. 3) $$\Delta H / C = 54,96\,{}^{1}\kappa - 990,4 n_{OO} + 12934 d_{OO} + 2631 Q_{OO} - 19371$$

Avec ${}^{1}\kappa$ l'indice de Kier shape d'ordre 1 (t-test= 12,7), n_{OO} le nombre de groupe peroxy (t-test=-14,9), d_{OO} la distance entre les atomes d'oxygène de la liaison coupée (t-test=4,5) et Q_{OO} la moyenne des charges de Mulliken sur les atomes d'oxygène de la liaison rompue (t-test=7,8).

Tableau 25 : Performances du modèle (5. 3) pour ΔH/C

R^2	RMSE	MAE	MAE(%)	Q^2	$Q^2$5cv	$Q^2$10cv	$Q^2$7cv	R^2_{YS}	σ_{YS}
0,97	99	67	5,66%	0,94	0,95	0,94	0,94	0,17	0,10
R^2_{in}	$RMSEP_{in}$	$MAEP_{in}$	$MAEP_{in}$ (%)	Q^2F1_{in}	Q^2F2i_{in}	Q^2F3_{in}	CCC_{in}		
0,81	301	173	14,47%	0,74	0,74	0,77	0,82		

Le modèle donne trois descripteurs très intéressants en termes d'interprétation (n_{OO} , d_{OO} et Q_{OO}) car ils sont directement liés à la liaison peroxy. Il présente aussi de très bonnes performances (voir Tableau 25) en ajustement (R^2=0,97, MAE=5,66%), en robustesse (Q^2=0,94 valeur élevée proche de R^2) et en prédiction (R^2_{in}=0,81 et $MAEP_{in}$=14,47%). La procédure de Y-scrambling respecte les critères de Rücker ($R^2 - R^2_{YS} > 2.3\sigma_{YS}$) avec R^2-R^2_{YS}= 0,97-0,17=0,80 et 2,3σ_{YS}=0,23.

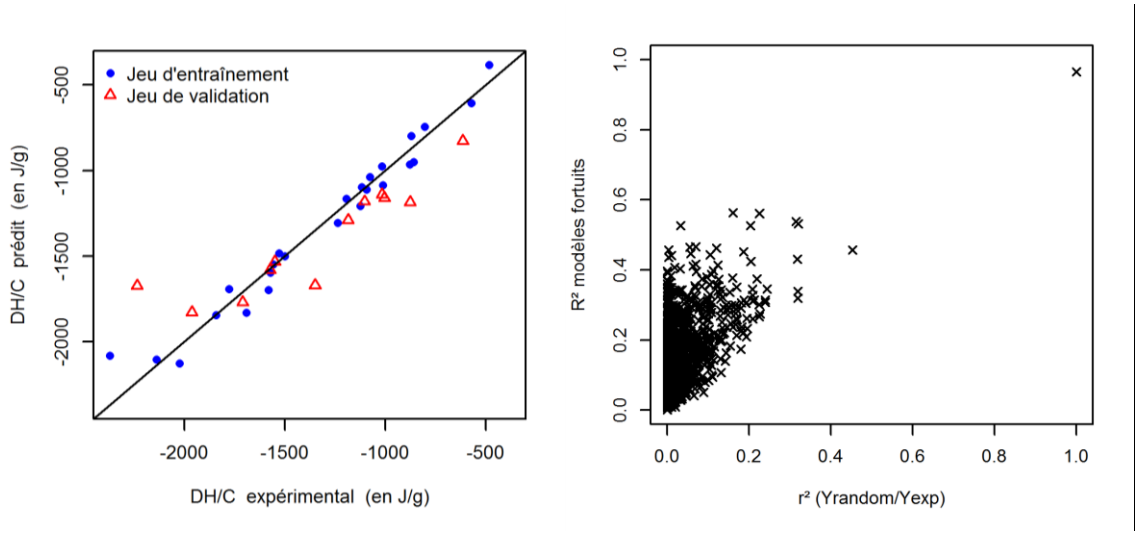


Figure 46 : Représentation des données expérimentales vs données prédites en utilisant l'équation (5. 3) et représentation des résultats de la procédure de Y-scrambling

Ce modèle ne présente aucune molécule du jeu de validation hors du domaine d'applicabilité. Parmi les descripteurs, l'indice de Kier shape d'ordre 1 encode le degré relatif de cyclicité[20,21]. Les trois autres descripteurs sont directement reliés à la liaison peroxy et le descripteur le plus important est le nombre de liaisons peroxy dans la molécule. Cela confirme le mécanisme de décomposition qui

commence par la rupture de la liaison peroxy. Ce modèle est en accord avec les observations faites à partir de la Datatop sur l'influence de la concentration sur les propriétés liées à la stabilité thermique et avec les mesures expérimentales faites dans PREDIMOL par Arkema, montrant une corrélation entre la concentration et la chaleur de décomposition. Ce modèle présente aussi une meilleure robustesse que le modèle MLR proposé par Lu[5] (Q^2=0,77 contre Q^2=-0,81 pour Lu).

4. Modèle pour la température onset

La température onset (°C) est la température à laquelle la tangente à la courbe de montée du pic de décomposition (droite verte sur la Figure 42) coupe la ligne de base (en bleue). De même que pour la chaleur de décomposition, la visualisation du diagramme des valeurs expérimentales de température onset permet de supprimer une molécule de la base de données : la 3,3,5,7,7 pentamethyl 1,2,4 trioxepane.

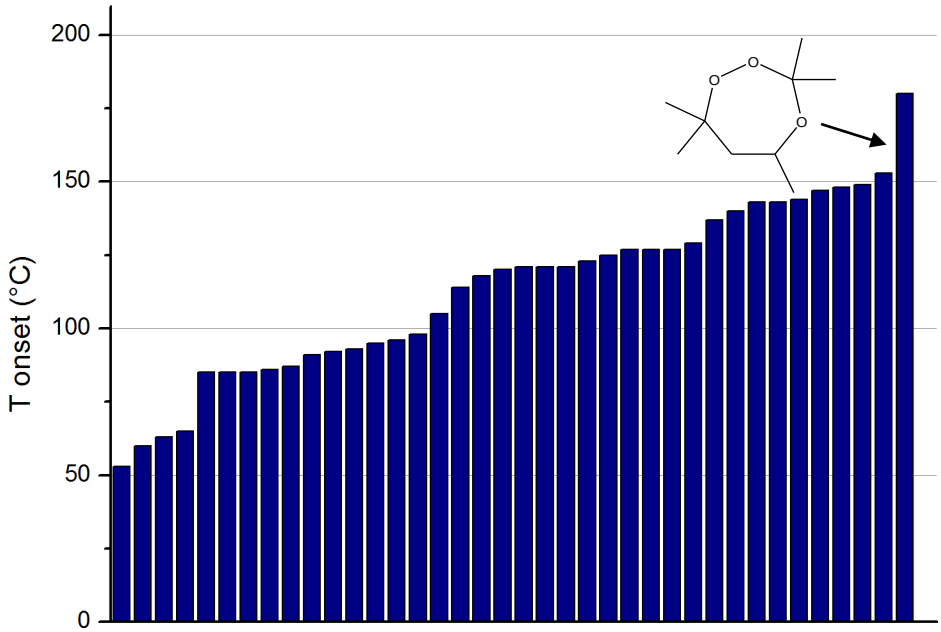


Figure 47 : Diagramme des valeurs expérimentales pour la température onset

Un modèle à 3 descripteurs est obtenu (équation (5. 4) et Figure 48) :

(5. 4) $$T_{onset} = 144F_{OO}^{-} + 29,3n_{OO} - 19,6gap_{HOMO-LUMO} + 194$$

Avec n_{OO} le nombre de liaisons peroxy dans la molécule (t-test=3,6), F_{OO}^{-} la moyenne des fonctions de Fukui localisées sur les atomes de la liaison peroxy (t-test=7,4) et $gap_{HOMO-LUMO}$ la différence d'énergie entre les orbitales HOMO et LUMO en eV (t-test=-4.4).

Le Tableau 26 résume les performances du modèle (5. 4) qui présente une molécule du jeu de validation hors du domaine d'applicabilité: 2,5-dimethyl-2,5-di(tert-butylperoxy)hexane.

Tableau 26 : Performances du modèle (5. 4) pour T_{onset}

R^2	RMSE	MAE	MAE(%)	Q^2	$Q^2$5cv	$Q^2$10cv	$Q^2$7cv	R^2_{YS}	σ_{YS}
0,84	12	9	8,43%	0,77	0,68	0,78	0,78	0,13	0,09
R^2_{ext}	RMSEP	MAEP	MAEP(%)	Q^2F1	Q^2F2	Q^2F3	CCC		
0,80	16	11	9,97%	0,74	0,74	0,78	0,87		
R^2_{in}	$RMSEP_{in}$	$MAEP_{in}$	$MAEP_{in}$ (%)	Q^2F1_{in}	Q^2F2_{in}	Q^2F3_{in}	CCC_{in}		
0,83	14	10	9,33%	0,78	0,78	0,83	0,90		

Le coefficient de détermination est légèrement plus faible part rapport au modèle de Lu et Mannan (R^2=0,84 contre 0,92) mais la robustesse est extrêmement meilleure (Q^2=0,77 au lieu de 0,12).

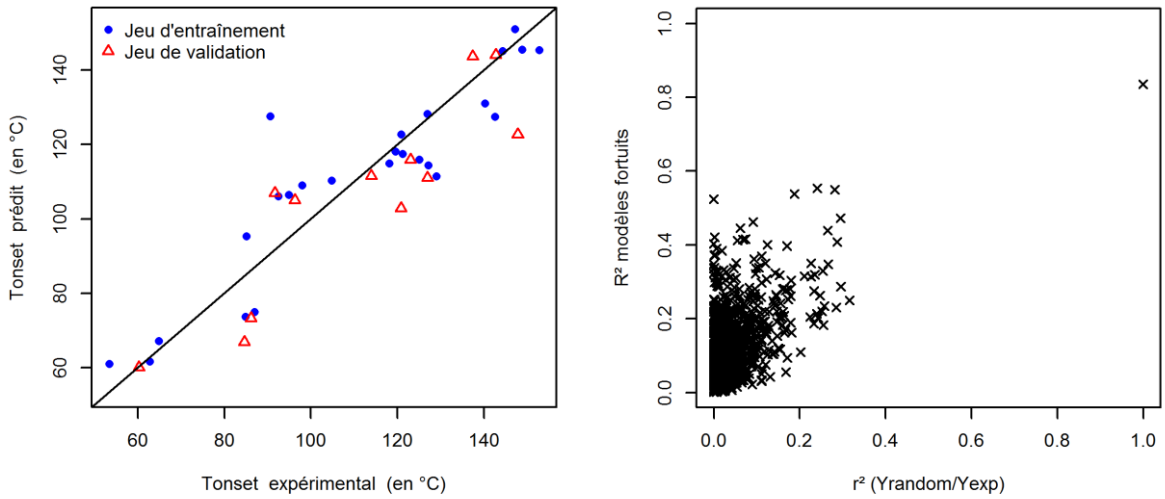


Figure 48 : Représentation des données expérimentales vs données prédites en utilisant l'équation (5. 4) et représentation des résultats de la procédure de Y-scrambling

Le descripteur F_{OO}^{-} caractérise la réactivité de la liaison peroxy, en particulier les attaques électrophiles. La différence d'énergie entre les orbitales HOMO et LUMO est un nombre positif qui est un indicateur de la stabilité de la molécule : plus l'écart est grand, plus la molécule est stable. Deux descripteurs sont directement liés à la liaison peroxy : F_{OO}^{-} et n_{OO}. Le nombre de liaisons peroxy est déjà présent dans le modèle (5. 3) mais cette fois le descripteur le plus important est F_{OO}^{-}.

5. Modèle pour la température maximale du pic de décomposition

Tout comme les propriétés précédentes, la visualisation du diagramme (Figure 47) des valeurs expérimentales de température maximale du pic de décomposition (Tpic en °C) permet de supprimer une molécule de la base de données : la molécule 3,3,5,7,7 pentamethyl 1,2,4 trioxepane qui est la même que celle supprimée pour le développement de modèles pour la température onset.

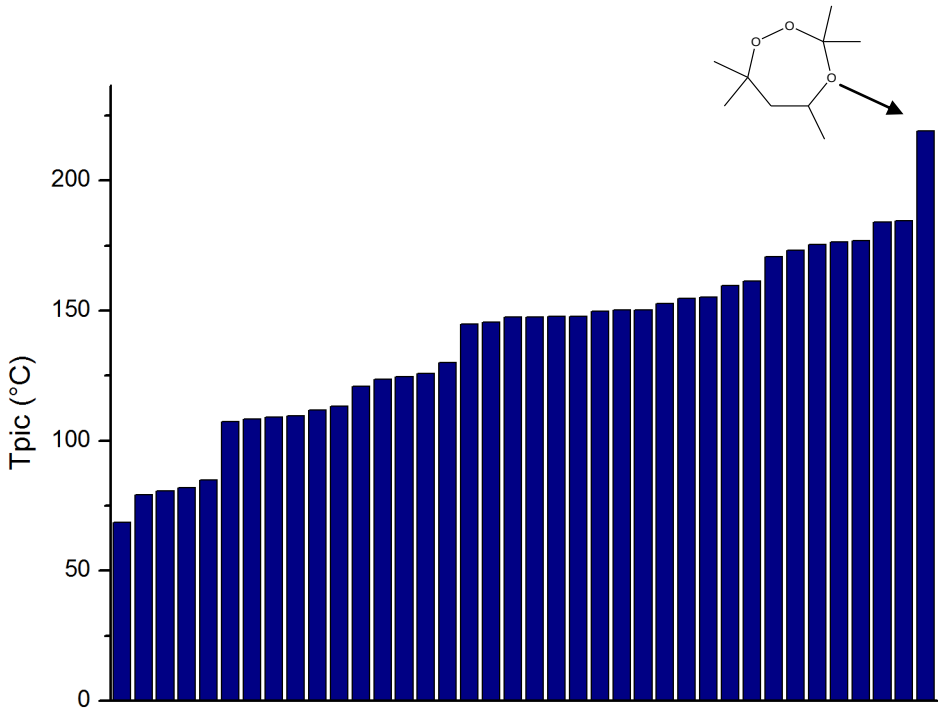


Figure 49 : Diagramme des valeurs expérimentales pour la température maximale du pic de décomposition

Un modèle à 3 descripteurs est obtenu (Figure 49) :

(5. 5)
$$T_{pic} = 83F_{OO}^{-} + 34,42E_{HOMO} - 398R_{avg,O} + 394$$

Avec F_{OO}^{-} la moyenne des fonctions de Fukui localisées sur les atomes de la liaison peroxy (t-test=13), E_{HOMO} l'énergie de l'orbitale HOMO (t-test=4,6) et $R_{avg,O}$ l'indice moyen de réactivité des atomes d'oxygène (t-test=-3,1).

Ce modèle présente de bonnes performances (MAE=7,52%, Q^2=0,81, $MAEP_{in}$=9,11%) qui sont rapportées dans le Tableau 27 et deux molécules du jeu de validation hors du domaine d'applicabilité : tert-amyl hydroperoxide et 2,5-dimethyl-2,5-dihydroperoxyhexane.

Tableau 27: Performances du modèle (5. 5) pour T_{pic}

R^2	RMSE	MAE	MAE(%)	Q^2	$Q^2$5cv	$Q^2$10cv	$Q^2$7cv	R^2_{YS}	σ_{YS}
0,86	13	10	7,52%	0,81	0,83	0,81	0,81	0,13	0,09
R^2_{ext}	RMSEP	MAEP	MAEP(%)	Q^2F1	Q^2F2	Q^2F3	CCC		
0,33	36	20	14,45%	0,09	0,09	0,80	0,57		
R^2_{in}	$RMSEP_{in}$	$MAEP_{in}$	$MAEP_{in}$ (%)	Q^2F1_{in}	Q^2F2i_{in}	Q^2F3_{in}	CCC_{in}		
0,87	17	11	9,11%	0,83	0,82	0,97	0,91		

Le descripteur F_{OO}^{-} est présent dans le modèle pour la prédiction de la température onset et le descripteur E_{HOMO} est relié au gap d'énergie entre les orbitales HOMO et LUMO (descripteur du modèle (5. 4)).

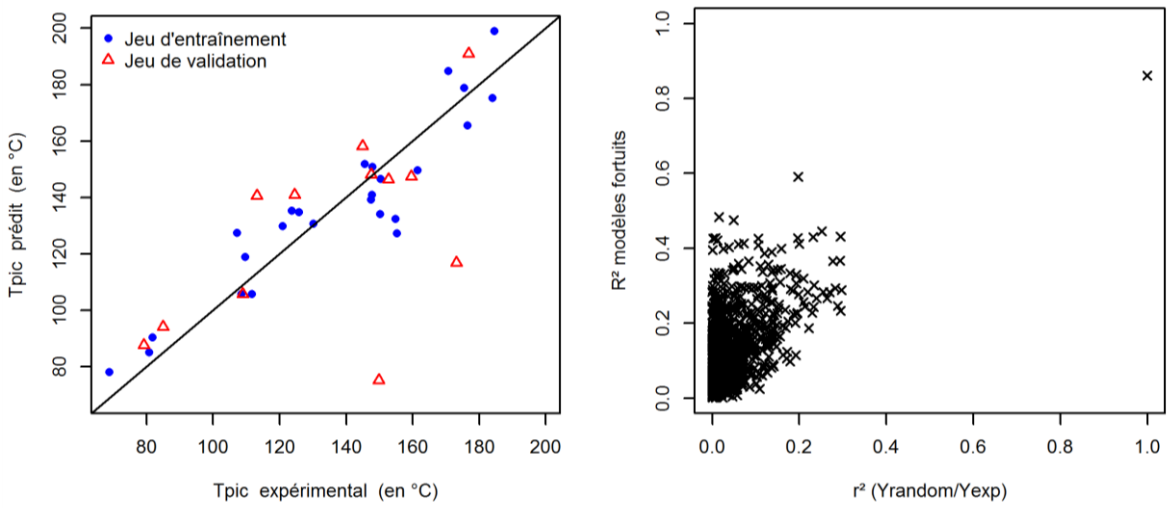


Figure 50 : Représentation des données expérimentales vs données prédites en utilisant l'équation (5. 5) et représentation des résultats de la procédure d'Y-scrambling

Ces deux modèles ont des descripteurs qui sont identiques ou liés entre eux. Cela s'explique par le fait que ces propriétés sont, par définition, corrélées l'une à l'autre.

6. Un modèle unique pour la prédiction de deux températures

Comme la température onset et celle maximale du pic de décomposition sont corrélées, les descripteurs obtenus pour la température onset ont été utilisés pour obtenir par régression multilinéaire un modèle pour la température maximale du pic. Et inversement, pour les descripteurs de la température maximale du pic avec la température onset. Ces régressions ont été faites dans le but de déterminer si l'utilisation des mêmes descripteurs est possible pour les deux propriétés et, de ce fait, faciliter la prédiction des propriétés (en termes de calcul des descripteurs nécessaires pour l'application du modèle).

a) MLR pour la température onset

Dans un premier temps, ce sont les descripteurs de l'équation (5. 4) qui ont été utilisés pour obtenir une équation afin de prédire la température onset :

(5. 6)
$$T_{\text{onset}} = 114F_{OO}^{-} + 36,92E_{HOMO} - 859R_{avg,O} + 386$$

Avec F_{OO}^{-} la moyenne des fonctions de Fukui localisées sur les atomes de la liaison peroxy (t-test=3,98), E_{HOMO} l'énergie de l'orbitale HOMO (t-test=7,18) et $R_{avg,O}$ l'indice moyen de réactivité des atomes d'oxygène (t-test=-1,91).

Tableau 28 : Performances du modèle (5. 6) pour T_{onset} à partir des descripteurs de l'équation (5. 4)

R^2	RMSE	MAE	MAE(%)	Q^2	$Q^2$5cv	$Q^2$10cv	$Q^2$7cv	R^2_{YS}	σ_{YS}
0,82	13	9	8,86%	0,77	0,77	0,74	0,76	0,04	0,05
R^2_{in}	$RMSEP_{in}$	$MAEP_{in}$	$MAEP_{in}$ (%)	Q^2F1_{in}	Q^2F2i_{in}	Q^2F3_{in}	CCC_{in}		
0,74	17	11	11,59%	0,73	0,73	0,95	0,85		

Le Tableau 28 résume les performances du modèle obtenu qui sont quasiment aussi bonnes que celle du modèle obtenu initialement (équation (5. 4)) avec une valeur de MAE identique, la MAEP qui augmente d'un seul degré Celsius et la RMSEP de 3°C. La variation importante de la valeur du coefficient R^2 (R^2_{in}=0,83 pour le modèle (5. 4) et R^2_{in}=0,74 pour le modèle (5. 6)) est due à la petite taille de la base de données expérimentale. Aucune molécule du jeu de validation ne se situe en dehors du domaine d'applicabilité.

b) MLR pour la température maximale du pic

De même, les descripteurs du modèle (5. 5) ont été utilisés pour la prédiction de la température maximale du pic. L'équation (5. 7), dont les performances sont résumées dans le Tableau 29, est obtenue :

(5. 7) $$T_{\text{pic}} = 161F_{OO}^{-} + 35,6n_{OO} - 22,85gap_{\text{HOMO-LUMO}} + 235$$

Avec F_{OO}^{-} la moyenne des fonctions de Fukui localisées sur les atomes de la liaison peroxy (t-test=7,22), n_{OO} le nombre de liaisons peroxy dans la molécule (t-test=4,24) et $gap_{\text{HOMO-LUMO}}$ la différence d'énergie entre les orbitales HOMO et LUMO en eV (t-test=-5,15).

Tableau 29 : Performances du modèle (5. 7) pour Tpic à partir des descripteurs de l'équation (5. 5)

R^2	RMSE	MAE	MAE(%)	Q^2	$Q^2$5cv	$Q^2$10cv	$Q^2$7cv	R^2_{YS}	σ_{YS}
0,82	15	11	8,45%	0,66	0,65	0,67	0,70	0,04	0,05
R^2_{ext}	RMSEP	MAEP	MAEP(%)	Q^2F1	Q^2F2	Q^2F3	CCC		
0,88	14	10	7,06%	0,87	0,87	0,97	0,94		
R^2_{in}	$RMSEP_{in}$	$MAEP_{in}$	$MAEP_{in}$ (%)	Q^2F1_{in}	Q^2F2i_{in}	Q^2F3_{in}	CCC_{in}		
0,88	15	10	7,43%	0,87	0,87	0,97	0,94		

Le 1,1-di-(tert-butylperoxy)-3,3,5-trimethylcyclohexane est en dehors du domaine d'applicabilité. Les performances du modèle (5. 7) (voir le Tableau 29) sont, là encore, très proches de celles du modèle déjà développé pour la température maximale du pic de décomposition (équation (5. 5)).

Ces nouveaux modèles ((5. 6) et (5. 7)) ont l'intérêt de diminuer le nombre de descripteurs à calculer puisque que les mêmes sont utilisés pour les deux propriétés, tout en conservant une prédictivité très bonne. Le nombre de descripteurs étant le même dans les modèles pour la prédiction de la température onset (5. 4) et la température maximale du pic (5. 5), les modèles basés sur les descripteurs obtenus lors du développement de modèles pour la température onset ((5. 4) et (5. 7)) sont préférés. En effet, la présence du descripteur n_{OO} permet une interprétation du mécanisme de décomposition des peroxydes organiques.

III. INFLUENCE DE LA CONFORMATION

Dans cette partie, les modèles développés précédemment sont appliqués aux conformations sélectionnées, pour deux peroxydes, par le programme Callisto (voir le chapitre 2 IV) afin d'observer l'influence de la conformation. La première molécule étudiée est un peroxyde flexible qui appartient au jeu de validation pour le développement des modèles pour ΔH/C et Tpic : il s'agit du di-(4-tert-butylcyclohexyl) peroxydicarbonate représenté dans la Figure 51.

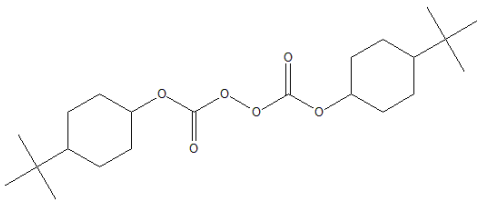


Figure 51 : Structure du di-(4-tert-butylcyclohexyl) peroxydicarbonate

On s'intéressera aux conformations restantes après clustering et avant l'analyse de population de Boltzmann afin de conserver la diversité des conformations initiales. La génération automatique de conformations avec Scigress[22] donne accès à 908 conformères pour le di-(4-tert-butylcyclohexyl) peroxydicarbonate. Ils peuvent être réduit à 2 clusters si on considère un nombre de clusters faible et la valeur maximale du profil silhouette (S_{max}=0,28 pour 2 clusters, voir Figure 52).

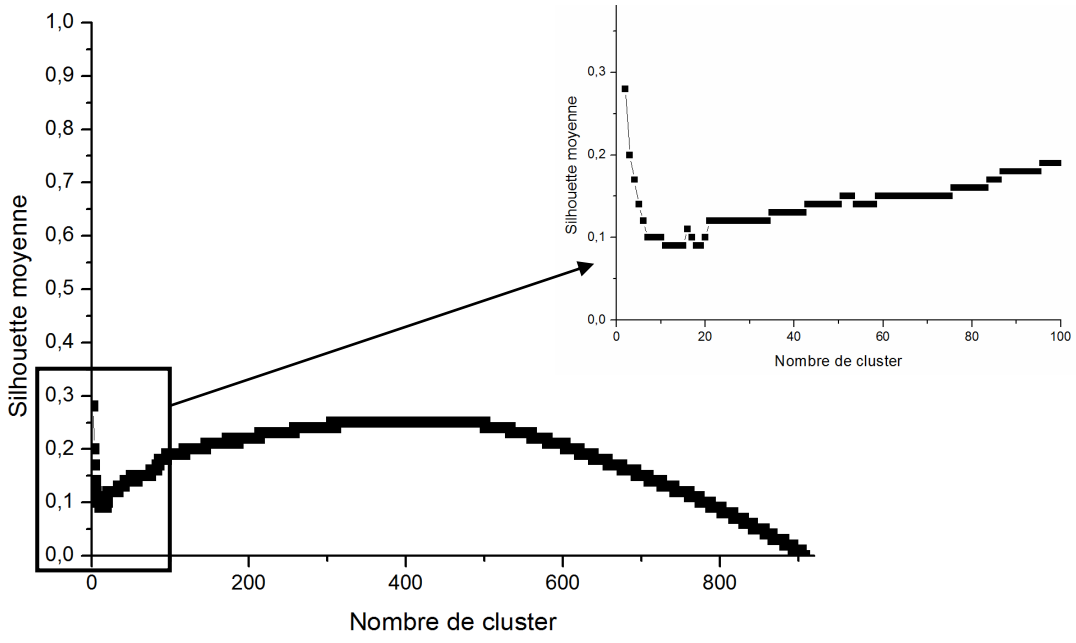
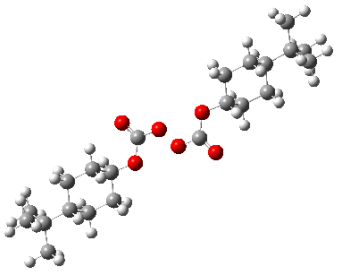
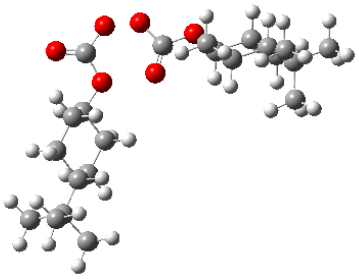


Figure 52 : Valeur du profil silhouette en fonction du nombre de clusters obtenus pour le di-(4-tert-butylcyclohexyl) peroxydicarbonate

Les descripteurs ont été calculés pour ces deux conformations (sans optimisation DFT supplémentaire) puis les équations (5. 3) et (5. 7) ont été utilisées. Les prédictions obtenues pour ces différentes conformations, rapportées dans le Tableau 30, ne sont pas identiques.

Tableau 30 : Prédictions pour les conformations du di-(4-tert-butylcyclohexyl) peroxydicarbonate avec E l'énergie (calculée en MM3 par Scigress) en kcal/mol et RMSD la valeur de RMSD en Å entre les conformations n et n-1

	Structure 3D	ΔH/C (J/g)	Tpic (°C)
Conf n°1 E=34.81		-793,4	125,7
Conf n°2 E=37.00 RMSD=2.81		-567,8	144,6
Conformation obtenue suivant le même protocole que pour l'obtention des structures des molécules du jeu d'entraînement (Conf n°1 optimisée en DFT comme expliqué dans le chapitre 3)		-826,9	87,9

Un autre peroxyde ayant un nombre de conformation faible après la sélection par le programme Callisto, le tert-butyl peroxy-3,5,5-trimethylhexanoate, a été étudié. L'analyse du profil silhouette (S_{max}=0,29 pour 174 à 181 clusters) ne permet pas la sélection d'un petit nombre optimal de clusters.

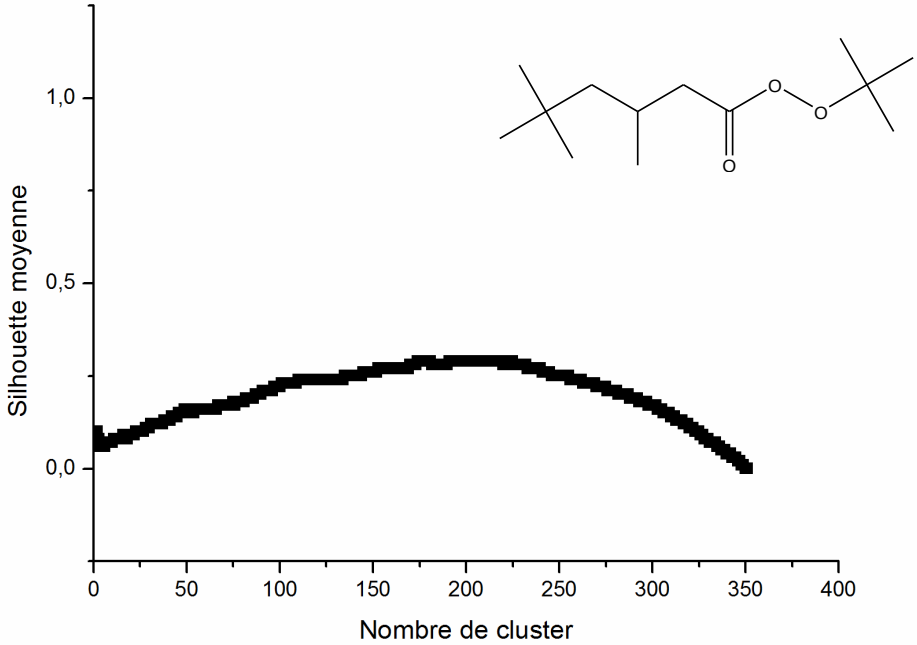
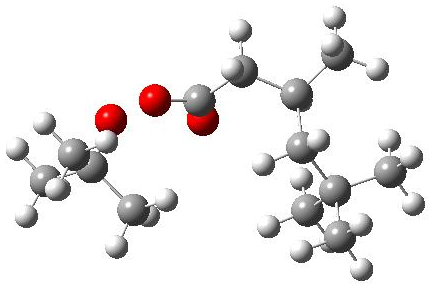
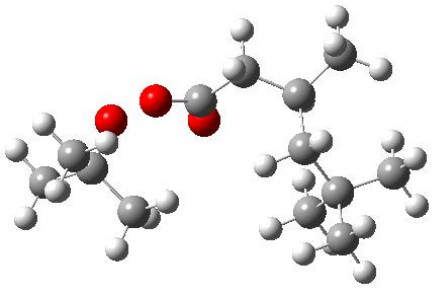
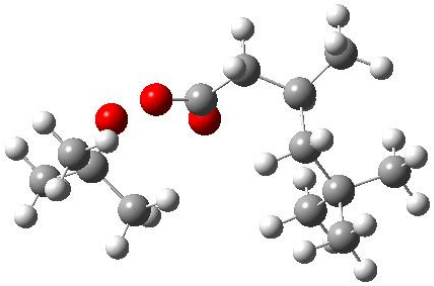
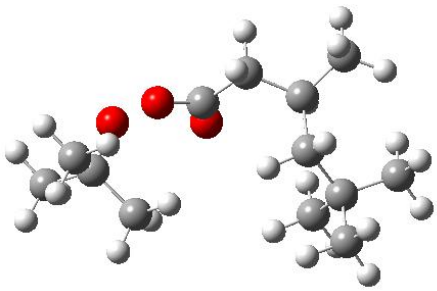
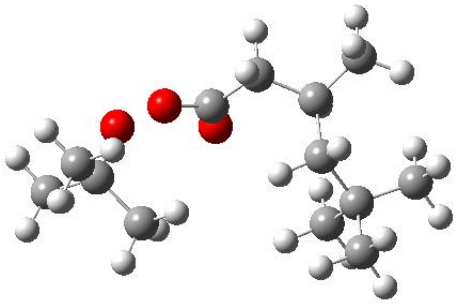
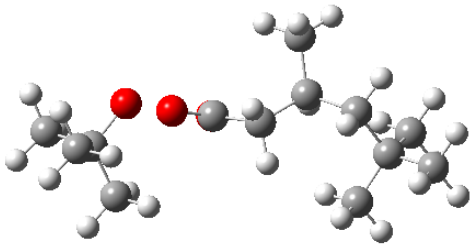
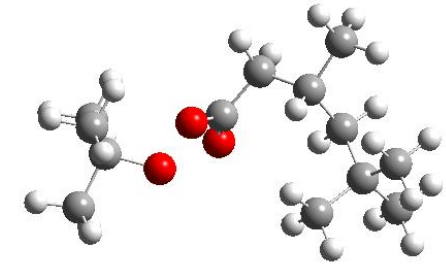
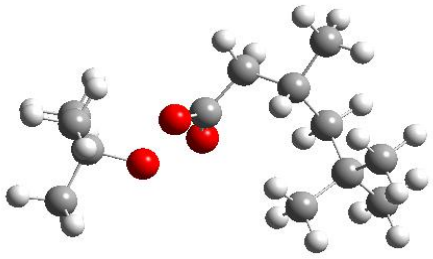


Figure 53 : Valeur du profil silhouette en fonction du nombre de clusters obtenus pour le tert-butyl peroxy-3,5,5-trimethylhexanoate

Pour ce peroxyde le choix des conformations sélectionnées est basé sur les options par défaut du programme Callisto (voir Chapitre 2.IV) : clustering agglomératif average link avec une valeur seuil de RMSD de 1,50 Å. Une liste de 8 conformations (obtenue à partir de 351 conformères), dont les valeurs prédite sont disponibles Tableau 31, est obtenue (S=0,07).

Tableau 31: Prédictions pour les conformations du tert-butyl peroxy-3,5,5-trimethylhexanoate avec E l'énergie (calculée en MM3 par Scigress) en kcal/mol et RMSD la valeur de RMSD en Å entre les conformations n et n-1

	Structure 3D	ΔH/C (J/g)	Tonset (°C)
Conf n°1 E= 31.60		-1185,2	111,7
Conf n°2 E= 31.61 RMSD=1.07		-1185,6	111,7
Conf n°3 E= 31.61 RMSD=1.51		-1185,2	111,7
Conf n°4 E= 31.63 RMSD=1.52		-1185,2	111,7

	Structure 3D	ΔH/C (J/g)	Tonset (°C)
Conf n°5 E= 31.63 RMSD=1.53		-1184,9	111,7
Conf n°6 E= 31.89 RMSD=1.64		-1225,4	106,2
Conf n°7 E= 32.54 RMSD=1.59		-1116,7	100,4
Conf n°8 E= 32.57 RMSD=1.08		-1116,9	100,3
Conformation obtenue suivant le même protocole que pour l'obtention des structures des molécules du jeu d'entraînement (Conf n°1 optimisée en DFT comme expliqué dans le chapitre 3)		-1185,7	111,6

Les conformations 1 à 5 sont proches en géométrie ainsi que leurs valeurs de chaleur de décomposition et de température onset. En revanche, la conformation n°6, qui se distingue bien des précédentes en termes de géométrie, a des valeurs plus faibles. De même, les conformations n°7 et 8 ont des valeurs similaires entre elles.

Dans ces exemples, on observe bien l'influence de la conformation sur les valeurs prédites. Ainsi, une analyse conformationnelle ou *a minima* une optimisation de la géométrie conduisant à la localisation du minimum global d'énergie est nécessaire avant l'utilisation de modèles. La géométrie utilisée doit être obtenue de la même manière pour les molécules du jeu d'entraînement et celles du jeu de validation afin d'être cohérent dans les prédictions.

Nous ne sommes, pour le moment, pas allé plus loin avec les peroxydes organiques. La poursuite de cette analyse est le développement de modèles avec différents jeux de conformations (considération soit de la structure minimale MM3, soit de la structure minimale DFT après clustering, soit de toutes les structures, soit enfin de toutes les structures en fonction du poids des conformations[23]) afin de comparer la prédictivité et le temps de calcul nécessaire.

IV. INFLUENCE DE LA MÉTHODE DE PARTAGE

Le partage de la base de données en jeu d'entraînement et de validation est un paramètre important pour le développement des modèles[8,24]. Jusqu'à présent, les modèles obtenus ont tous été développés en suivant la même méthode de partage (basée sur la valeur de la propriété[25]). Dans cette partie, l'influence du choix des jeux lors du développement a été étudiée en utilisant une autre méthode de séparation des jeux que l'on nommera « manuelle ». Le principe de ce découpage sera expliqué et les modèles développés à partir des nouveaux jeux pour la chaleur de décomposition, la température onset et la température maximale du pic de décomposition seront présentés.

1. Description d'une méthode de partage alternative

La répartition « manuelle » (avec toujours 1/3 des molécules dans le jeu de validation) des molécules des jeux d'entraînement et de validation s'effectue par l'intermédiaire d'une analyse en composante principale. Les molécules sont représentées dans l'espace des descripteurs et sont sélectionnées de manière à avoir une distribution homogène des deux jeux. La distribution par rapport aux valeurs expérimentales de la propriété est aussi vérifiée.

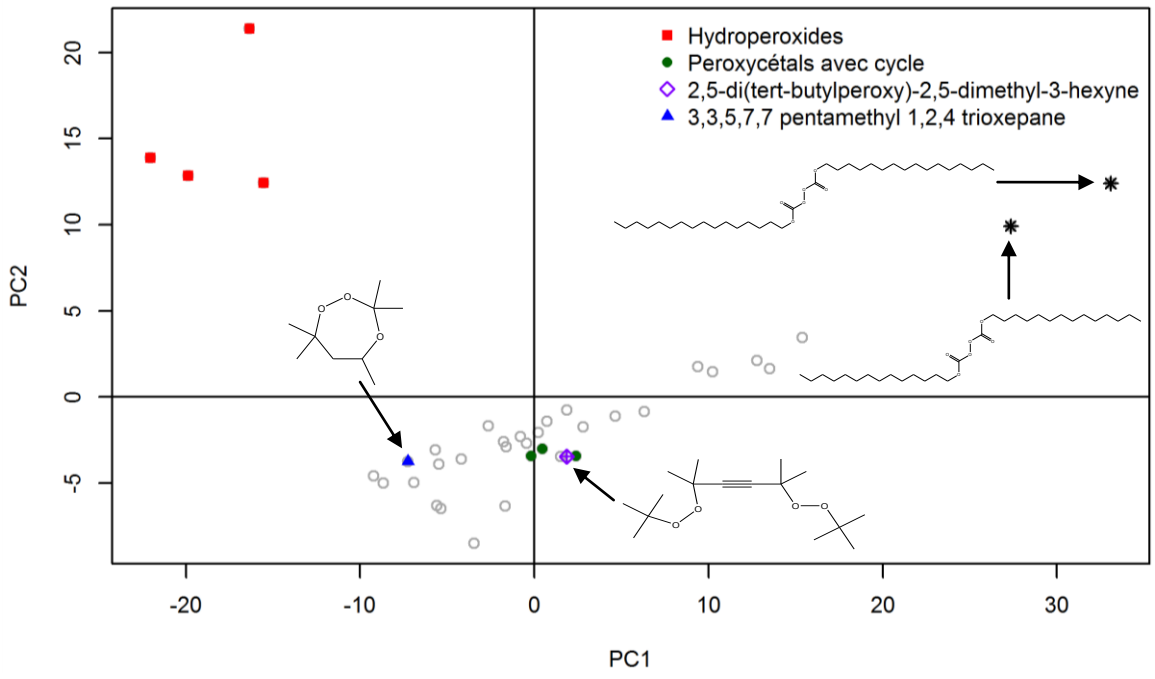


Figure 54 : PCA des données PREDIMOL

La PCA combinée avec l'observation des structures (disponible en annexe II) permet d'identifier quelques molécules "particulières" (comme illustrée en Figure 54), notamment 4 hydroperoxydes, 3 peroxycétals avec cycle, une molécule avec une liaison triple (2,5-di(tert-butylperoxy)-2,5-dimethyl-3-hexyne) et une molécule cyclique (3,3,5,7,7 pentamethyl 1,2,4 trioxepane). Ces deux dernières sont mises dans le jeu d'entraînement afin de couvrir la plus grande diversité de structures possibles soit un domaine d'applicabilité plus grand. Pour les familles de molécules comme les hydroperoxydes et les peroxycétals avec cycle, au moins une molécule de chaque groupe doit être présente dans le jeu d'entraînement ainsi que dans le jeu de validation afin quelles soient représentées dans les deux jeux. La PCA permet aussi d'identifier deux molécules clairement isolées, en haut à droite : dicetyl peroxydicarbonate et dimyristyl peroxydicarbonate (Figure 54). Une de ces deux molécules sera donc dans le jeu de validation et l'autre dans le jeu d'entraînement. Pour la sélection des molécules du jeu de validation parmi celles restantes, la valeur de la propriété est utilisée de la même manière que pour la méthode basée sur la propriété : les molécules sont ordonnées par valeur de propriété puis une sur trois est sélectionnée. Il s'agit d'obtenir une distribution de la propriété la plus homogène possible en plus de celle des structures afin d'obtenir des modèles plus avec des meilleures performances[8] par rapport au partage basé uniquement sur les structures ou uniquement sur la valeur de la propriété.

2. Modèle pour la chaleur de décomposition

Ainsi pour la chaleur de décomposition divisée par la concentration, un nouveau modèle à trois descripteurs est obtenu (équation (5. 8) et Figure 55) :

(5. 8) $$\Delta H / C = -1298 n_{OO} + 159\,{}^{0}\chi + 112\,\mathrm{WNSA3} - 1308$$

Avec n_{OO} le nombre de groupes peroxy (t-test=-12,1), ${}^{0}\chi$ l'indice Kier&Hall d'ordre 0 (t-test=-10,8) et WNSA3 la surface pondérée par la surface partiellement chargée négative (t-test=11,5).

Notons que le descripteur n_{OO} est encore une fois présent. Il est donc important pour cette propriété, d'autant plus qu'il est celui ayant la valeur de t-test la plus élevée. Le Tableau 32 présente les très bonnes performances du modèle (5. 8) qui ne présente aucune molécule du jeu de validation hors du domaine d'applicabilité.

Tableau 32 : Performances du modèle (5. 8)

R^2	RMSE	MAE	MAE(%)	Q^2	$Q^2$5cv	$Q^2$10cv	$Q^2$7cv	R^2_{YS}	σ_{YS}
0,91	145	102	7,93%	0,87	0,89	0,87	0,90	0,13	0,09
R^2_{in}	$RMSEP_{in}$	$MAEP_{in}$	$MAEP_{in}$ (%)	Q^2F1_{in}	Q^2F2i_{in}	Q^2F3_{in}	CCC_{in}		
0,89	224	145	12,51%	0,87	0,87	0,84	0,94		

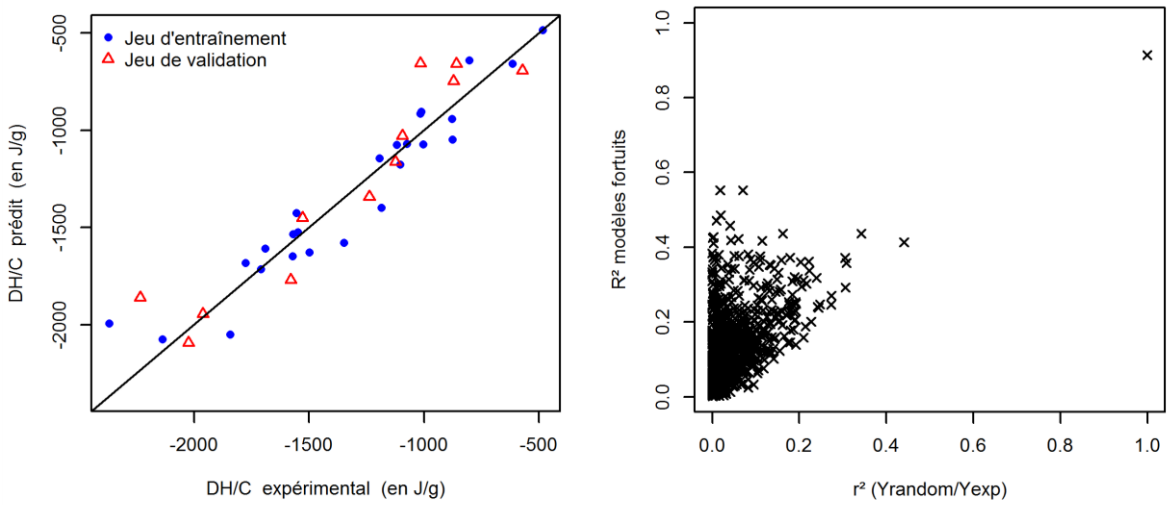


Figure 55 : Représentation des données expérimentales vs données prédites en utilisant l'équation (5. 8) et résultats de la procédure de Y-scrambling

Le modèle « manuel » nécessite moins de descripteurs et a une meilleure prédictivité que le modèle (5. 3) obtenu précédemment avec le partage 1/3-2/3. Cependant, les descripteurs présentés par le modèle (5. 3) sont plus facilement interprétables chimiquement avec 3 descripteurs liés à la liaison peroxy (présence, distance et charge). Ces modèles sont tous les deux intéressants mais le modèle « manuel » semble meilleur dans le cadre de la prédiction de propriétés.

3. Modèle pour la température onset

Suite à l'utilisation de cette seconde méthode de partage de la base de données en deux jeux, une nouvelle équation est obtenue :

(5. 9) $$T_{onset} = 60F_{OO}^{-} + 1079d_{OO} - 15,4gap_{HOMO\text{-}LUMO} - 1327$$

Avec F_{OO}^{-} la moyenne des fonctions de Fukui localisées sur les atomes de la liaison peroxy (t-test=2,97), d_{OO} la distance de la liaison peroxy (t-test=3,74) et $gap_{HOMO\text{-}LUMO}$ la différence d'énergie entre les orbitales HOMO et LUMO en eV (t-test=-3,44).

Le Tableau 33 donne les performances du modèle (5. 9) qui ne présente aucune molécule du jeu de validation hors du domaine d'applicabilité.

Tableau 33:Performances du modèle (5. 9)

R^2	RMSE	MAE	MAE(%)	Q^2	$Q^2$5cv	$Q^2$10cv	$Q^2$7cv	R^2_{YS}	σ_{YS}
0,84	13	9	8,50%	0,80	0,79	0,81	0,81	0,12	0,09
R^2_{in}	$RMSEP_{in}$	$MAEP_{in}$	$MAEP_{in}$ (%)	Q^2F1_{in}	Q^2F2i_{in}	Q^2F3_{in}	CCC_{in}		
0,78	13	9	8,66%	0,77	0,77	0,97	0,85		

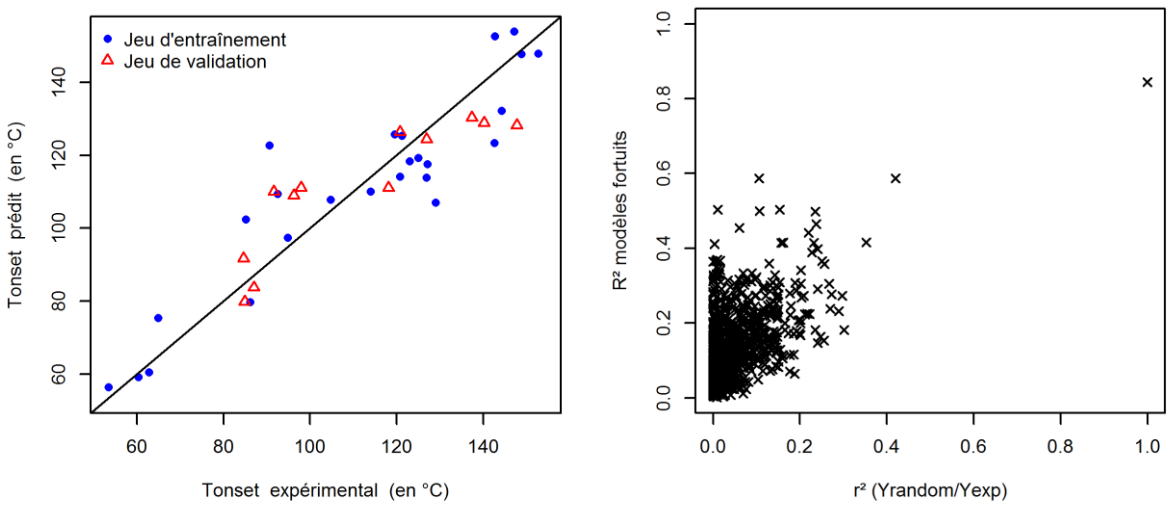


Figure 56 : Représentation des données expérimentales vs données prédites en utilisant l'équation (5. 9) et résultats de la procédure de Y-scrambling

Par rapport au modèle (5. 4), un seul descripteur change : d_{OO} remplace n_{OO}. Ce descripteur est plus complexe à calculer mais reste directement relié à la liaison peroxy.

4. Modèle pour la température maximale du pic

La seconde méthode de partage utilisée donne l'équation suivante :

(5. 10) $$T_{pic} = 3{,}14\Delta H_{disso} + 95F_{OO}^{-} + 22n_{COOC} - 25{,}7gap_{HOMO\text{-}LUMO} + 200$$

Avec ΔH_{disso} l'énergie de dissociation par la coupure homolytique de la liaison OO en kcal/mol (t-test=5,14), F_{OO}^{-} la moyenne des fonctions de Fukui localisées sur les atomes de la liaison peroxy (t-test=7,35), n_{COOC} le nombre de liaisons peroxy hors liaisons hydroperoxy (t-test=4,06) et $gap_{HOMO-LUMO}$ la différence d'énergie entre les orbitales HOMO et LUMO en eV (t-test=-7,30).

Tableau 34: Performances du modèle (5. 10)

R^2	RMSE	MAE	MAE(%)	Q^2	$Q^2$5cv	$Q^2$10cv	$Q^2$7cv	R^2_{YS}	σ_{YS}
0,94	9	6	4,83%	0,91	0,91	0,91	0,91	0,17	0,10
R^2_{ext}	RMSEP	MAEP	MAEP(%)	Q^2F1	Q^2F2	Q^2F3	CCC		
0,60	25	15	12,06%	0,54	0,54	0,92	0,72		
R^2_{in}	$RMSEP_{in}$	$MAEP_{in}$	$MAEP_{in}$ (%)	Q^2F1_{in}	Q^2F2i_{in}	Q^2F3_{in}	CCC_{in}		
0,80	19	11	8,30%	0,79	0,79	0,98	0,88		

Le Tableau 34 donne les performances du modèle (5. 10) qui présente trois molécules du jeu de validation en dehors du domaine d'applicabilité : dibenzoyl peroxide, tert-amyl hydroperoxide et le tert-butyl hydroperoxide. Ce modèle présente de bonnes performances avec une erreur MAE=4,83% en ajustement, Q^2=0,91 en robustesse et une erreur $MAEP_{in}$=8,30% en prédictivité dans son domaine d'applicabilité.

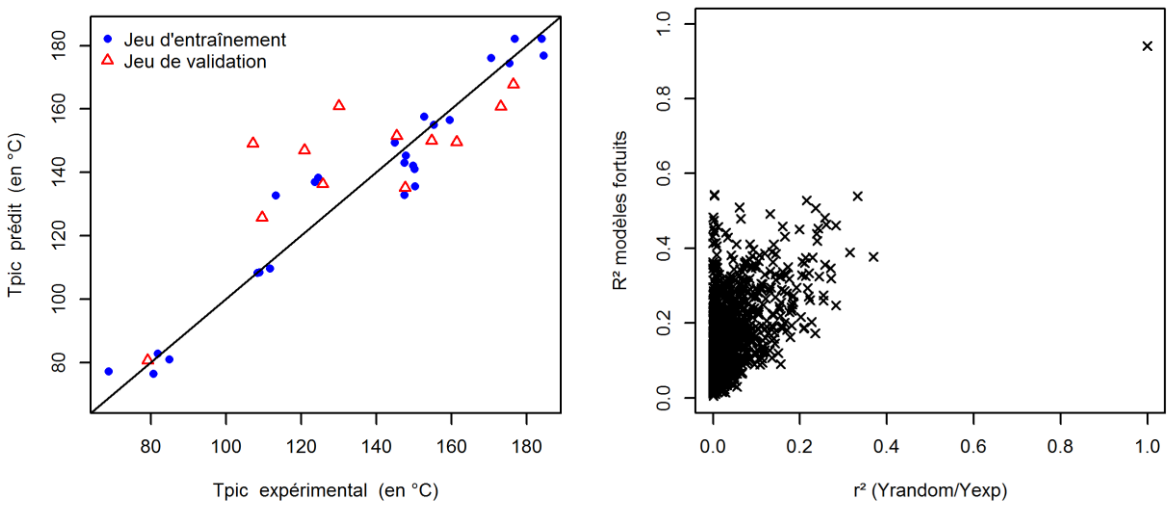


Figure 57 : Représentation des données expérimentales vs données prédites en utilisant l'équation (5. 10) et résultats de la procédure de Y-scrambling

Les descripteurs $gap_{HOMO-LUMO}$ et F_{OO}^{-}, déjà présents dans le modèle développé pour la température onset sur le jeu obtenu par le partage manuel, le sont aussi dans ce modèle. Le descripteur F_{OO}^{-} illustre l'importance de la réactivité de la liaison peroxy, tout comme le descripteur ΔH_{disso}. Celui-ci est une information importante sur le lien entre ΔH/C et l'énergie de dissociation calculée de la liaison peroxy. Cela confirme le mécanisme supposé de décomposition des peroxydes organiques, à savoir la coupure homolytique de la liaison peroxy.

5. Comparaison des résultats

Cette « nouvelle » façon de déterminer les jeux d'entraînement et de validation donne accès à de nouveaux modèles ayant des performances supérieures ou identiques en prédictivité (R^2_{in} et $MAEP_{in}$%) par rapport aux modèles précédents comme le montre le Tableau 35, qui résume les performances des modèles développés à partir des deux méthodes de répartition (par valeur de propriété ou par PCA) pour les trois propriétés.

Tableau 35 : Résumé des performances des 6 modèles développés

Modèles	Nombre de descripteurs	R^2	Q^2	MAE%	R^2_{ext}	MAEP%	R^2_{in}	$MAEP_{in}$ %
ΔH/C	4	0,97	0,94	5,66%	0,81	14,47%	0,81	14,47%
ΔH/C – Manuel	3	0,91	0,87	7,93%	0,89	12,51%	0,89	12,51%
Tonset	3	0,84	0,77	8,43%	0,80	9,97%	0,83	9,33%
Tonset - Manuel	3	0,84	0,80	8,50%	0,78	8,66%	0,78	8,66%
Tpic	3	0,86	0,81	7,52%	0,33	14,45%	0,87	9,11%
Tpic - Manuel	4	0,94	0,91	4,83%	0,60	12,06%	0,80	8,30%

L'amélioration des performances s'explique par le fait que la distribution des données dans les jeux d'entraînement et de validation soit la meilleure possible en considérant la valeur de la propriété mais aussi la répartition des molécules dans l'espace chimique.

Les descripteurs obtenus sont proches en interprétation quand ils ne sont pas identiques, comme par exemple pour les modèles (5. 4) et (5. 9). Le descripteur F_{OO}^{-} intervient dans les 4 modèles concernant la température onset et la température maximale du pic, indiquant que la réactivité de la zone de la liaison peroxy semble importante pour ces deux propriétés. Pour la chaleur de décomposition, le descripteur commun aux deux modèles est n_{OO}, le nombre de liaisons peroxy qui semble avoir une grande importance pour cette dernière propriété. Une observation similaire avait été faite pour la prédiction de la chaleur de décomposition des composés nitroaromatiques[26] où le nombre de groupement NO_2 était un paramètre important.

V. VERS LA SIMPLIFICATION DES MODÈLES

Les modèles recherchés sont des modèles qui, en plus d'avoir des descripteurs interprétables, sont faciles à utiliser pour les industriels, parfois au prix d'une diminution des performances, notamment dans le cadre des procédures d'enregistrement de REACH des substances chimiques. Pour cela, des modèles plus simples, c'est-à-dire contenant des descripteurs ne nécessitant pas la détermination de la structure géométrique ou des calculs quantiques, ont aussi été développés avec les deux méthodes précédentes.

1. Modèles pour la chaleur de décomposition

Des modèles ont été développés pour la prédiction de la chaleur de décomposition divisée par la concentration en partant de a) 83 descripteurs constitutionnels et topologiques puis de b) 45 descripteurs constitutionnels uniquement.

a) Modèles avec des descripteurs constitutionnels et topologiques

Pour commencer, un modèle a été développé sur le jeu d'entraînement obtenu par un partage 1/3-2/3 (comme expliqué dans le chapitre 3 au paragraphe V.1.a) avec les descripteurs constitutionnels et topologiques. Le modèle suivant est obtenu :

(5. 11) $$\Delta H / C = -604 n_{OO} - 386 n_{OOH} - 4134 n_{O,r} + 63,7\,{}^{1}\mathrm{BIC} - 1220$$

Avec n_{OO} le nombre de liaisons peroxy (t-test=-10,0), n_{OOH} le nombre de liaisons hydroperoxy (t-test=4,32), $n_{O,r}$ le nombre d'oxygènes relatif (t-test=-5,41) et ^{1}BIC l'indice d'ordre 1 d'information relatif aux liaisons (t-test=8,83).

Les performances de ce modèle, qui ne présente aucune molécule du jeu de validation hors du domaine d'applicabilité, sont disponibles dans le Tableau 36.

Tableau 36 : Performances du modèle (5. 11)

R^2	RMSE	MAE	MAE(%)	Q^2	$Q^2$5cv	$Q^2$10cv	$Q^2$7cv	R^2_{YS}	σ_{YS}
0,95	117	79	5,89%	0,91	0,91	0,90	0,90	0,17	0,10
R^2_{in}	$RMSEP_{in}$	$MAEP_{in}$	$MAEP_{in}$ (%)	Q^2F1_{in}	Q^2F2i_{in}	Q^2F3_{in}	CCC_{in}		
0,68	345	207	15,72%	0,66	0,66	0,70	0,79		

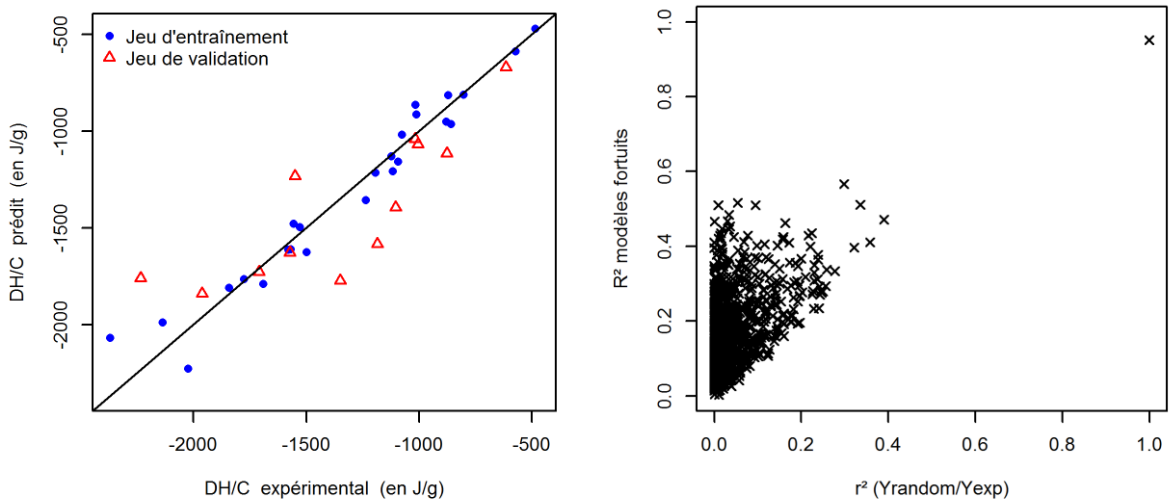


Figure 58 : Représentation des données expérimentales vs données prédites en utilisant l'équation (5. 11) et résultats de la procédure de Y-scrambling

Le descripteur n_{OO} est, encore une fois, présent avec la valeur de t-test la plus élevée. Ce descripteur est donc important pour la prédiction de la chaleur de décomposition divisée par la concentration.

Un second modèle a été développé sur le jeu « manuel » dont l'équation est la suivante :

(5. 12) $$\Delta H / C = -660 n_{OO} - 700 n_{OOH} - 60 \mathrm{BO}_{100} + 7,26\, {}^{1}\mathrm{BIC} - 244$$

Avec n_{OO} le nombre de liaisons peroxy (t-test=-9,09), n_{OOH} le nombre de liaisons hydroperoxy (t-test=-5.75), BO_{100} la balance en oxygène selon Kamlet[27] (t-test=-6,04) et ^{1}BIC l'indice d'ordre 1 d'information relatif aux liaisons (t-test=7,22).

Le Tableau 37 présente les performances de ce modèle pour lequel une seule molécule du jeu de validation est hors du domaine d'applicabilité : le tert-butyl hydroperoxide.

Tableau 37 : Performances du modèle (5. 12)

R^2	RMSE	MAE	MAE(%)	Q^2	$Q^2$5cv	$Q^2$10cv	$Q^2$7cv	R^2_{YS}	σ_{YS}
0,92	147	107	8,88%	0,83	0,83	0,84	0,85	0,17	0,10
R^2_{ext}	RMSEP	MAEP	MAEP(%)	Q^2F1	Q^2F2	Q^2F3	CCC		
0,85	270	136	8,58%	0,83	0,83	0,79	0,91		
R^2_{in}	$RMSEP_{in}$	$MAEP_{in}$	$MAEP_{in}$ (%)	Q^2F1_{in}	Q^2F2i_{in}	Q^2F3_{in}	CCC_{in}		
0,89	268	123	8,08%	0,83	0,83	0,81	0,89		

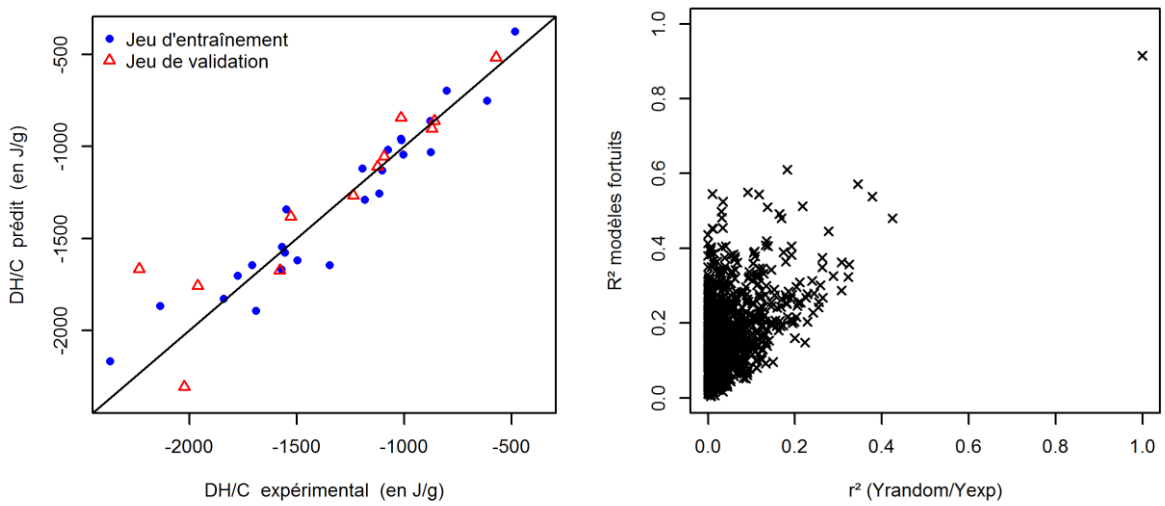


Figure 59 : Représentation des données expérimentales vs données prédites en utilisant l'équation (5. 12) et résultats de la procédure de Y-scrambling

Le modèle présente les mêmes descripteurs que le modèle (5. 11) précédent à l'exception de $n_{O,r}$ qui est remplacé par OB_{100}. Les performances en prédiction de ce modèle sont plus élevées, ce qui confirme l'observation selon laquelle le découpage manuel présente des modèles plus prédictifs.

b) Modèles avec des descripteurs constitutionnels uniquement

Les modèles développés avec moins de descripteurs (constitutionnels uniquement) mais beaucoup plus simples à calculer sont présentés ici. Tout d'abord, avec le découpage 1/3-2/3 l'équation suivante est obtenue :

(5. 13) $$\Delta H / C = 10301 n_{C,r} + 16,3 n_H - 311 n_{cycle} - 4918$$

Avec $n_{C,r}$ le nombre relatif d'atomes de carbone (t-test=3,71), n_H le nombre d'atomes d'hydrogène (t-test=3,39) et n_{cycle} le nombre de cycles dans la molécule (t-test=-1,70).

Le Tableau 38 présente les performances du modèle (5. 13) pour lequel deux molécules du jeu de validation sont hors du domaine d'applicabilité : dibenzoyl peroxide et di-(4-tert-butylcyclohexyl) peroxydicarbonate.

Tableau 38 : Performances du modèle (5. 13)

R^2	RMSE	MAE	MAE(%)	Q^2	$Q^2$5cv	$Q^2$10cv	$Q^2$7cv	R^2_{YS}	σ_{YS}
0,74	267	185	13,99%	0,65	0,57	0,62	0,61	0,13	0,09
R^2_{ext}	RMSEP	MAEP	MAEP(%)	Q^2F1	Q^2F2	Q^2F3	CCC		
0,01	711	476	41,74%	-0,64	-0,65	0,66	0,07		
R^2_{in}	$RMSEP_{in}$	$MAEP_{in}$	$MAEP_{in}$ (%)	Q^2F1_{in}	Q^2F2i_{in}	Q^2F3_{in}	CCC_{in}		
0,21	496	350	27,36%	0,23	0,20	0,90	0,30		

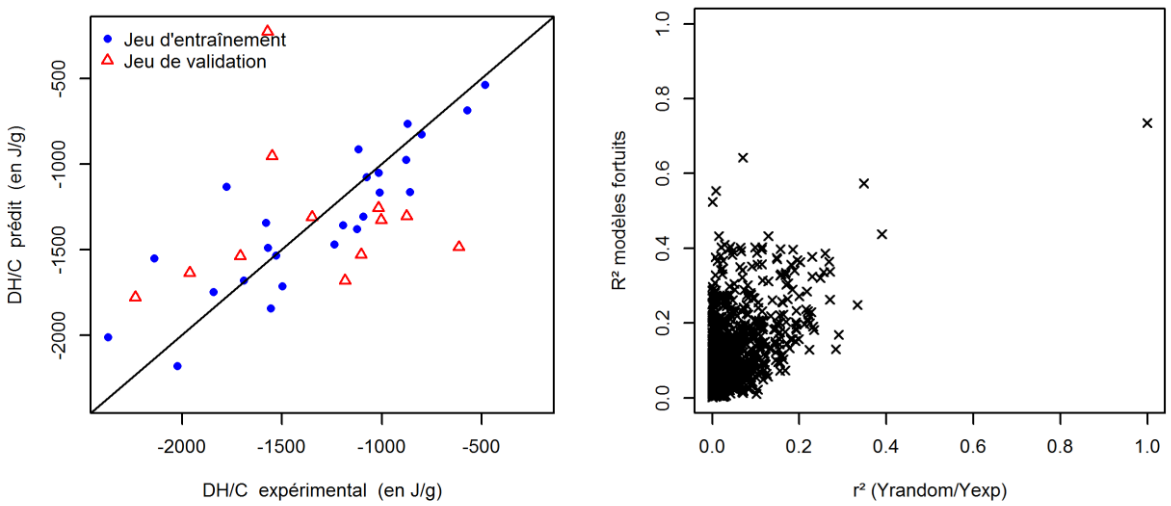


Figure 60 : Représentation des données expérimentales vs données prédites en utilisant l'équation (5. 13) et résultats de la procédure de Y-scrambling

Les descripteurs obtenus n'ont pas beaucoup de sens par rapport à la propriété étudiée (aucun en relation avec la liaison peroxy ou les atomes d'oxygène en général) et les performances sont mauvaises. Ce modèle n'a pas d'intérêt en termes d'interprétation ou de prédiction. En revanche, le modèle développé sur le jeu manuel présente des descripteurs intéressants et de bonnes performances.

(5. 14) $$\Delta H / C = -663 n_{OO} - 699 n_{OOH} - 4,79\,\mathrm{BO} + 11 n_{\mathrm{simple}} - 2036$$

Avec n_{OO} le nombre de liaisons peroxy (t-test=-8,13), n_{OOH} le nombre de liaisons hydroperoxy (t-test=-5,29), BO la balance en oxygène selon la réglementation TDM[28] (t-test=-4,12) et n_{simple} le nombre de liaisons simples (t-test=5,36).

Les performances de ce modèle (5. 14), qui présente une molécule du jeu de validation hors du domaine d'applicabilité (tert-butyl hydroperoxide), sont disponibles dans le Tableau 39.

Tableau 39: Performances du modèle (5. 14)

R^2	RMSE	MAE	MAE(%)	Q^2	$Q^2$5cv	$Q^2$10cv	$Q^2$7cv	R^2_{YS}	σ_{YS}
0,89	166	118	10,93%	0,78	0,80	0,81	0,81	0,17	0,10
R^2_{ext}	RMSEP	MAEP	MAEP(%)	Q^2F1	Q^2F2	Q^2F3	CCC		
0,79	306	171	11,98%	0,78	0,78	0,94	0,89		
R^2_{in}	$RMSEP_{in}$	$MAEP_{in}$	$MAEP_{in}$ (%)	Q^2F1_{in}	Q^2F2i_{in}	Q^2F3_{in}	CCC_{in}		
0,80	311	162	11,85%	0,77	0,77	0,95	0,86		

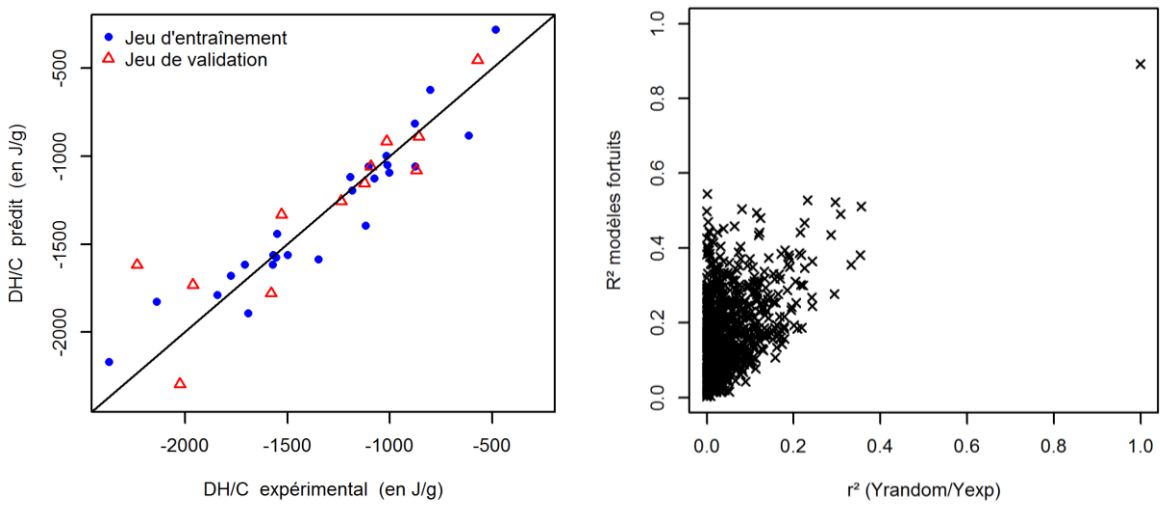


Figure 61 : Représentation des données expérimentales vs données prédites en utilisant l'équation (5. 14) et résultats de la procédure de Y-scrambling

Pour ce modèle, nous avons avancé un peu plus vers la simplicité. Le domaine d'applicabilité a aussi été défini par la méthode « bonding box » (voir chapitre 3 VII). Tout d'abord, le modèle est applicable à toutes les familles de peroxydes à l'exception des peroxydes de sulfonyle et de silyle. Ensuite, pour pouvoir appliquer ce modèle, les peroxydes organiques ne doivent pas avoir plus de 2 liaisons peroxy et peuvent avoir au maximum une liaison hydroperoxy. Les peroxydes considérés ont un nombre de liaisons simples compris entre 15 et 103 et une valeur de balance en oxygène comprise entre -132 et -269. Ainsi, dans notre exemple, aucune molécule du jeu de validation n'est en dehors du domaine d'applicabilité. La Figure 62 représente les molécules de notre base de données dans l'espace des descripteurs du modèle (5. 14).

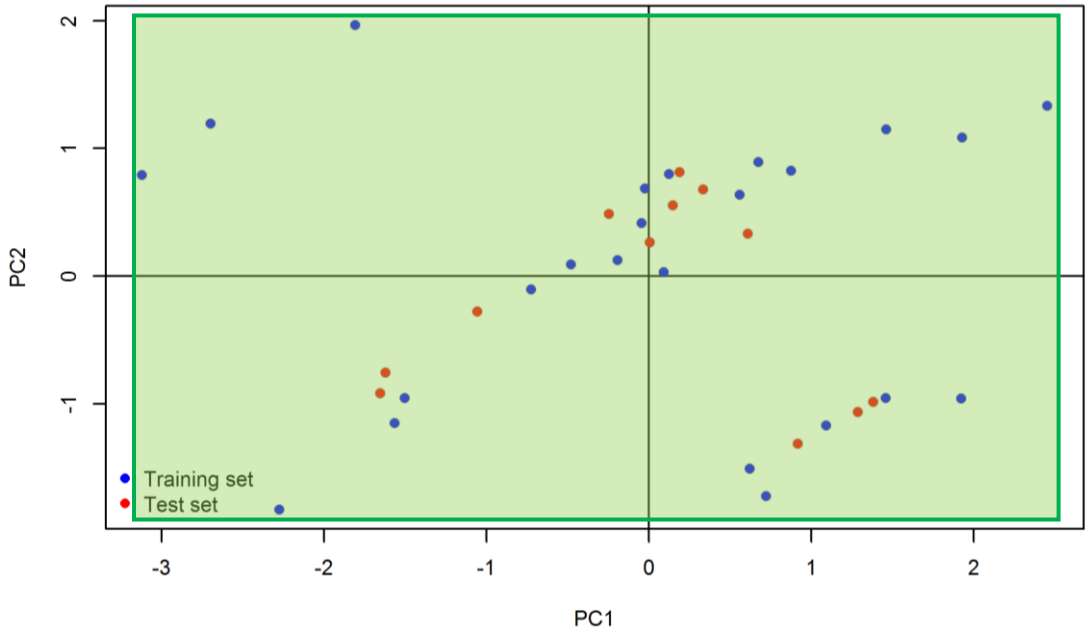


Figure 62 : PCA dans l'espace des descripteurs du modèle (5. 14)

c) Comparaison des modèles

Le Tableau 40 récapitule les performances de tous les modèles développés pour la chaleur de décomposition divisée par la concentration. Les trois modèles surlignés en bleu ont été développés en utilisant le partage manuel, tandis que les trois autres l'ont été avec le partage 1/3-2/3. Les modèles « descripteurs 1D/2D » correspondent au modèles développés à partir des descripteurs constitutionnels et topologiques et les modèles « descripteurs constitutionnels » à partir des descripteurs constitutionnels uniquement. Les modèles notés en violet et en gras sont ceux qui sont prédictifs et validés selon les principes de l'OCDE.

Tableau 40 : Modèles obtenus pour ΔH/C

Modèles ΔH/C	Nombre de descripteurs	R^2	Q^2	MAE%	R^2_{ext}	MAEP%	R^2_{in}	MAEP $_{in}$ %
Tous les descripteurs	4	0,97	0,94	5,66%	0,81	14,47%	0,81	14,47%
Tous les descripteurs - Manuel	3	0,91	0,87	7,93%	0,89	12,51%	0,89	12,51%
Descripteurs 1/2D	4	0,95	0,91	5,89%	0,68	15,72%	0,68	15,72%
Descripteurs 1/2D - Manuel	4	0,92	0,83	8,88%	0,85	8,58%	0,89	8,08%
Descripteurs constitutionnels	3	0,74	0,65	13,99%	0,01	41,74%	0,21	27,36%
Descripteurs constitutionnels - Manuel	4	0,89	0,78	10,93%	0,79	11,98%	0,80	11,85%

Le premier modèle, obtenu à partir de tous les descripteurs avec le découpage 1/3-2/3, est très intéressant, bien que sa prédictivité ne soit pas la meilleure, car trois de ces descripteurs sont directement liés à la liaison peroxy. De plus, il ne présente aucune molécule du jeu de validation hors du domaine d'applicabilité et, bien qu'élevée, l'erreur sur les prédictions reste de l'ordre de grandeur de l'incertitude expérimentale.

Finalement, le meilleur modèle pour cette propriété est celui développé sur le jeu obtenu avec le découpage manuel avec les descripteurs constitutionnels et topologiques. En effet, il présente les meilleures performances en prédictions (R^2_{in}, $RMSEP_{in}$) ainsi que des descripteurs intéressants (n_{OO}, n_{OOH} et BO_{100}).

La simplification des descripteurs permet de présenter des modèles avec des performances similaires, avec des valeurs de MAE proches, sauf dans le cas de l'équation (5. 13) obtenue à partir du découpage 1/3-2/3 et des descripteurs constitutionnels. Cette équation est la seule ne contenant pas la variable n_{OO}, ce qui appuie l'observation de son importance pour obtenir de bonnes prédictions.

2. Modèles pour la température onset

Quatre modèles ont été développés pour la température onset en suivant la même méthode utilisée précédemment avec la chaleur de décomposition divisée par la concentration.

a) Modèles avec des descripteurs constitutionnels et topologiques

Le développement sur le jeu obtenu par le découpage 1/3-2/3 avec des descripteurs constitutionnels et topologiques permet l'obtention de l'équation à deux paramètres suivante :

(5. 15) $$T_{onset} = -4,06BO_{100} - 2,67\phi + 54$$

Avec BO_{100} la balance en oxygène selon Kamlet[27] (t-test=-2,95) et ϕ l'indice de flexibilité de Kier (t-test=-4,23).

Le Tableau 41 donne les performances du modèle (5. 15) qui ne présente aucune molécule du jeu de validation hors du domaine d'applicabilité.

Tableau 41: Performances du modèle (5. 15)

R^2	RMSE	MAE	MAE(%)	Q^2	$Q^2$5cv	$Q^2$10cv	$Q^2$7cv	R^2_{YS}	σ_{YS}
0,51	21	17	16,31%	0,36	0,24	0,42	0,24	0,09	0,08
R^2_{in}	$RMSEP_{in}$	$MAEP_{in}$	$MAEP_{in}$ (%)	Q^2F1_{in}	Q^2F2i_{in}	Q^2F3_{in}	CCC_{in}		
0,54	20	15	13,71%	0,53	0,53	0,91	0,69		

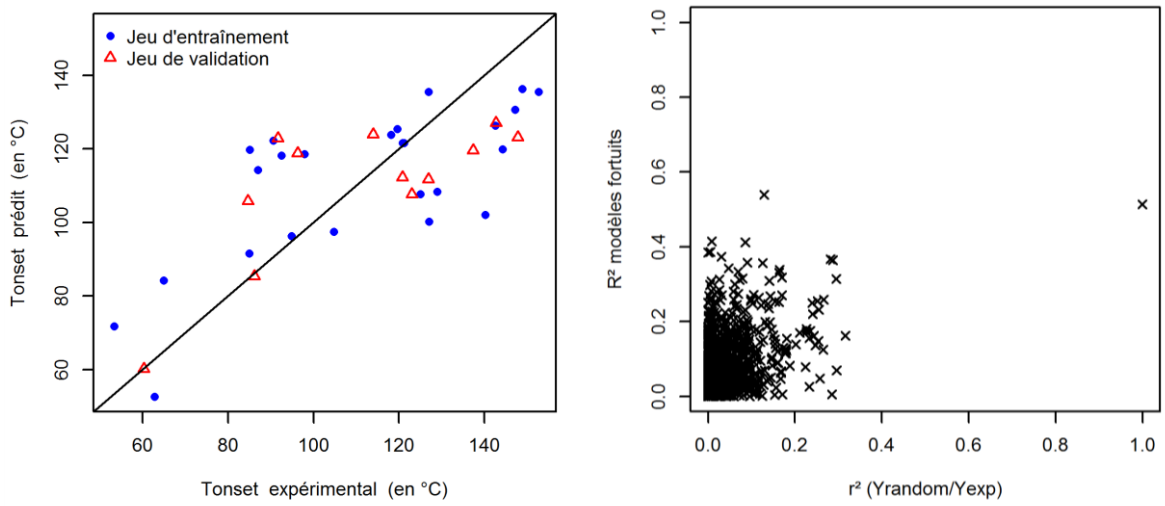


Figure 63 : Représentation des données expérimentales vs données prédites en utilisant l'équation (5. 15) et résultats de la procédure de Y-scrambling

Cette équation n'est pas un modèle car la validation interne ne montre aucune robustesse et indique que le modèle a été obtenu par chance (au moins une valeur de R^2 d'un modèle dit fortuit est supérieure à celui du modèle initial obtenu lors du développement du modèle).

Le modèle développé sur le jeu obtenu par la répartition manuelle présente également deux descripteurs :

(5. 16) $$T_{onset} = 28n_{OO} - 18n_O + 144$$

Avec n_{OO} le nombre de liaisons peroxy (t-test=2,97) et n_O le nombre d'atomes d'oxygène (t-test=-6,51). Ces descripteurs sont redondants et les performances sont faibles comparées à celle du modèle (5. 4) développé à partir de tous les descripteurs. Le Tableau 42 résume les performances du

modèle (5. 16) dont une seule molécule du jeu de validation est hors du domaine d'applicabilité : ethyl 3,3-di-(tert-amyl peroxy) butyrate.

Tableau 42: Performances du modèle (5. 16)

R^2	RMSE	MAE	MAE(%)	Q^2	$Q^2$5cv	$Q^2$10cv	$Q^2$7cv	R^2_{YS}	σ_{YS}
0,67	18	13	12,36%	0,60	0,55	0,58	0,58	0,08	0,08
R^2_{ext}	RMSEP	MAEP	MAEP(%)	Q^2F1	Q^2F2	Q^2F3	CCC		
0,25	24	18	16,46%	0,12	0,12	0,88	0,49		
R^2_{in}	$RMSEP_{in}$	$MAEP_{in}$	$MAEP_{in}$ (%)	Q^2F1_{in}	Q^2F2i_{in}	Q^2F3_{in}	CCC_{in}		
0,51	19	15	14,90%	0,41	0,20	0,94	0,68		

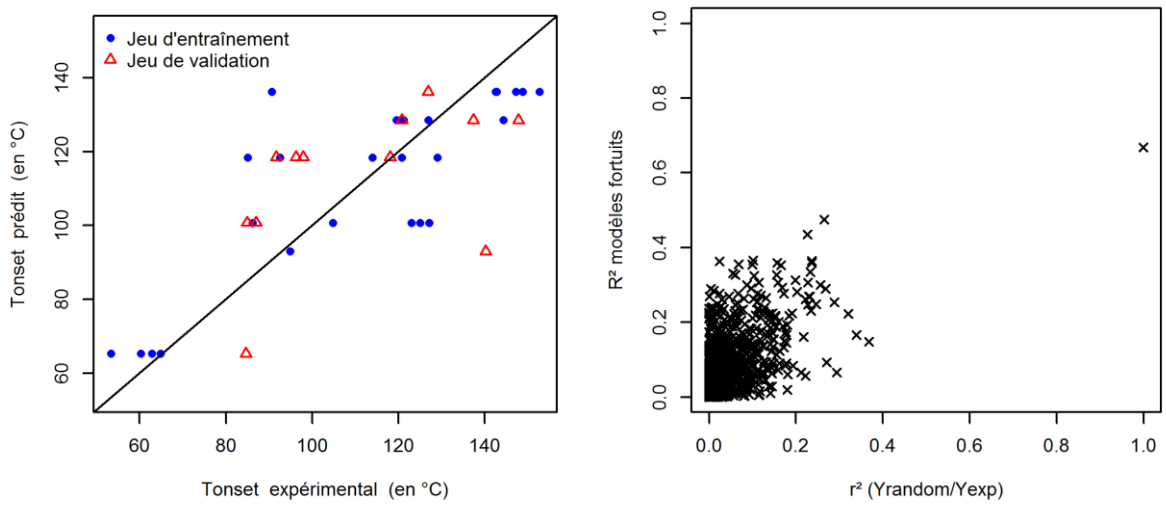


Figure 64 : Représentation des données expérimentales vs données prédites en utilisant l'équation (5. 16) et résultats de la procédure de Y-scrambling

Les performances de ce modèle pour la température onset sont mauvaises avec une erreur de 13°C. Les coefficients de mesure de l'ajustement R^2 et de la robustesse Q^2 ont des valeurs faibles (<0,70).

b) Modèles avec des descripteurs constitutionnels uniquement

Les deux modèles suivant obtenus à partir du partage 1/3-2/3 des jeux (équation (5. 17) et Tableau 43) et manuel (équation (5. 18) et Tableau 42) ont les mêmes descripteurs, qui sont aussi ceux du modèle (5. 16).

(5. 17) $$T_{onset} = 32n_{OO} - 16n_O + 130$$

(5. 18) $$T_{onset} = 28n_{OO} - 18n_O + 144$$

Avec n_{OO} le nombre de liaisons peroxy (t-test=3,86) et n_O le nombre d'atomes d'oxygène (t-test=-6,25).

Les performances du modèle (5. 17), qui présente une molécule du jeu de validation hors du domaine d'applicabilité (ethyl 3,3-di-(tert-amyl peroxy)butyrate), sont disponibles dans le Tableau 43.

Tableau 43 : Performances du modèle (5. 17)

R^2	RMSE	MAE	MAE(%)	Q^2	$Q^2$5cv	$Q^2$10cv	$Q^2$7cv	R^2_{YS}	σ_{YS}
0,67	17	12	12,74%	0,59	0,59	0,60	0,56	0,08	0,08
R^2_{ext}	RMSEP	MAEP	MAEP(%)	Q^2F1	Q^2F2	Q^2F3	CCC		
0,42	24	17	14,82%	0,37	0,37	0,88	0,59		
R^2_{in}	$RMSEP_{in}$	$MAEP_{in}$	$MAEP_{in}$ (%)	Q^2F1_{in}	Q^2F2i_{in}	Q^2F3_{in}	CCC_{in}		
0,56	20	15	13,41%	0,55	0,55	0,93	0,71		

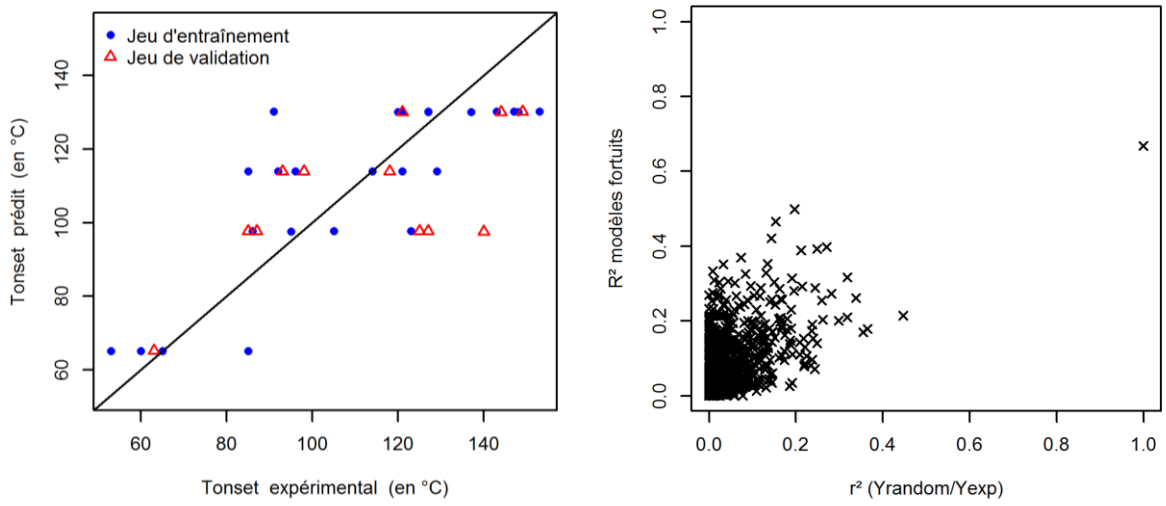


Figure 65 : Représentation des données expérimentales vs données prédites en utilisant l'équation (5. 17) et résultats de la procédure de Y-scrambling

Le modèle (5. 18) est le même que le modèle (5. 16), ses performances sont disponibles dans le Tableau 42 et Figure 64.

c) Comparaison des modèles

Les meilleurs modèles pour la température onset sont ceux développés avec tous les descripteurs (R^2_{in}=0,83 et 0,78 ; $MAEP_{in}$=9,33% et 8,66% voir dans le Tableau 44). Pour cette propriété, l'utilisation de descripteurs plus simples à calculer diminue beaucoup le pouvoir prédictif des modèles (passant de 9% à 14% en MAEP). Des descripteurs plus complexes à obtenir sont donc nécessaires. Notons également que le choix de la méthode de partage influe peu sur les modèles contrairement à ce qui avait été observé pour la prédiction de la chaleur de décomposition.

Tableau 44 : Modèles obtenus pour Tonset

Modèles - T_{onset}	Nombre de descripteurs	R^2	Q^2	MAE%	R^2_{ext}	MAEP%	R^2_{in}	$MAEP_{in}$ %
Tous les descripteurs	3	0,84	0,77	8,43%	0,80	9,97%	0,83	9,33%
Tous les descripteurs - Manuel	3	0,84	0,80	8,50%	0,78	8,66%	0,78	8,66%
Descripteurs 1/2D	2	0,51	0,36	16,31%	0,54	13,71%	0,54	13,71%
Descripteurs 1/2D - Manuel	2	0,67	0,60	12,36%	0,25	16,46%	0,51	14,90%
Descripteurs constitutionnels	2	0,67	0,59	12,74%	0,42	14,82%	0,56	13,41%
Descripteurs constitutionnels - Manuel	2	0,67	0,61	12,35%	0,25	16,50%	0,51	14,94%

3. Modèle pour la température maximale du pic

Quatre modèles ont été développés pour la température maximale du pic de décomposition en suivant toujours la même démarche.

a) Modèles avec des descripteurs constitutionnels et topologiques

Le jeu obtenu par le découpage 1/3-2/3 donne un modèle à deux descripteurs :

(5. 19) $$T_{\text{pic}} = -6,42\text{BO}_{100} - 3,14\phi + 40$$

Avec BO_{100} la balance en oxygène selon Kamlet[27] (t-test=-4,38) et ϕ l'indice de flexibilité de Kier (t-test=-4,83).

Ce sont les mêmes descripteurs que ceux de l'équation (5. 15) qui a été développée pour la température onset. Cela n'est pas si surprenant puisque nous avons déjà vu dans le paragraphe II.6 que ces propriétés sont corrélées et peuvent être prédites correctement avec le même modèle (aux coefficients près). Les performances de ce modèle (5. 19), dont une seule molécule du jeu de validation est hors du domaine d'applicabilité (dimyristyl peroxydicarbonate), sont disponibles dans le Tableau 45.

Tableau 45: Performances du modèle (5. 19)

R^2	RMSE	MAE	MAE(%)	Q^2	$Q^2$5cv	$Q^2$10cv	$Q^2$7cv	R^2_{YS}	σ_{YS}
0,62	21	16	12,53%	0,54	0,52	0,54	0,56	0,08	0,07
R^2_{ext}	RMSEP	MAEP	MAEP(%)	Q^2F1	Q^2F2	Q^2F3	CCC		
0,46	27	20	16,06%	0,41	0,41	0,87	0,65		
R^2_{in}	$RMSEP_{in}$	$MAEP_{in}$	$MAEP_{in}$ (%)	Q^2F1_{in}	Q^2F2i_{in}	Q^2F3_{in}	CCC_{in}		
0,24	29	21	17,30%	0,18	0,17	0,88	0,44		

Ce modèle n'est pas prédictif avec une erreur de 17% en prédictivité dans le domaine d'applicabilité et une robustesse avec des coefficients d'une valeur inférieure à 0,60.

Le jeu, obtenu avec le partage manuel, donne un modèle à trois descripteurs :

(5. 20) $$T_{\text{pic}} = 28n_{OO} - 22n_O - 12\text{n}_{\text{double}} + 190$$

Avec n_{OO} le nombre de liaisons peroxy (t-test=3,70), n_O le nombre d'atomes d'oxygène (t-test=-9,72) et n_{double} le nombre de liaisons doubles dans la molécule (t-test=-2,65). Les deux premiers sont ceux de l'équation (5. 16) pour la température onset.

Pour ce modèle, dont les performances sont résumées dans le Tableau 46, deux molécules du jeu de validation sont hors du domaine d'applicabilité : ethyl 3,3-di-(tert-amyl peroxy) butyrate et 2,5-dimethyl-2,5-di-(2-ethylhexanoyl peroxy) hexane. Ce modèle présente une légère amélioration par rapport au modèle précédent développé sur le jeu obtenu par le partage 1/3-2/3.

Tableau 46 : Performances du modèle (5. 20)

R^2	RMSE	MAE	MAE(%)	Q^2	$Q^2$5cv	$Q^2$10cv	$Q^2$7cv	R^2_{YS}	σ_{YS}
0,84	14	11	7,95%	0,79	0,80	0,79	0,79	0,13	0,09
R^2_{ext}	RMSEP	MAEP	MAEP(%)	Q^2F1	Q^2F2	Q^2F3	CCC		
0,38	31	20	14,81%	0,20	0,20	0,87	0,61		
R^2_{in}	$RMSEP_{in}$	$MAEP_{in}$	$MAEP_{in}$ (%)	Q^2F1_{in}	Q^2F2i_{in}	Q^2F3_{in}	CCC_{in}		
0,57	26	16	12,13%	0,52	0,52	0,94	0,74		

Les performances de ces deux modèles sont très mauvaises, notamment en prédictivité ($RMSEP_{in}$=26°C et $MAEP_{in}$=12,13%) : ils ne sont pas utilisables pour la prédiction de la température maximale du pic. Les descripteurs sont les mêmes que ceux utilisés pour la prédiction de la température onset, ce qui confirme la possibilité d'utiliser le même modèle (aux coefficients près) pour la prédiction de ces deux propriétés.

b) Modèles avec des descripteurs constitutionnels uniquement

Le jeu obtenu avec le partage 1/3-2/3 donne le modèle suivant :

(5. 21) $$T_{\text{pic}} = 43n_{COOC} - 21n_O - 10\text{n}_{\text{double}} + 170$$

Avec n_{COOC} le nombre de liaisons peroxy hors liaisons hydroperoxy (t-test=5,57), n_O le nombre d'atomes d'oxygène (t-test=-7,20) et n_{double} le nombre de liaisons doubles dans la molécule (t-test=-2,30).

Par rapport à l'équation (5. 20), seul le descripteur n_{OO} change, remplacé par n_{COOC}. Cependant, ces descripteurs sont très similaires, l'un comptant le nombre de liaisons OO et l'autre les liaisons OO qui ne sont pas des liaisons O-OH. Ce modèle, dont les performances sont disponibles dans Tableau 47, ne présente aucune molécule du jeu de validation hors du domaine d'applicabilité. Celles-ci sont mauvaises ($RMSEP_{in}$=33 °C et MAE_{in}=15%) ce qui n'est pas surprenant puisque les performances du modèle (5. 20) développé à partir de plus de descripteurs étaient déjà faibles.

Tableau 47: Performances du modèle (5. 21)

R^2	RMSE	MAE	MAE(%)	Q^2	$Q^2$5cv	$Q^2$10cv	$Q^2$7cv	R^2_{YS}	σ_{YS}
0,76	17	12	9,65%	0,67	0,71	0,67	0,68	0,13	0,09
R^2_{in}	$RMSEP_{in}$	$MAEP_{in}$	$MAEP_{in}$ (%)	Q^2F1_{in}	Q^2F2i_{in}	Q^2F3_{in}	CCC_{in}		
0,39	33	21	14,91%	0,23	0,23	0,83	0,62		

Le jeu obtenu par le découpage manuel donne le modèle suivant :

(5. 22) $$T_{\text{pic}} = 28n_{OO} - 22n_O - 12\text{n}_{\text{double}} + 190$$

Avec n_{OO} le nombre de liaisons peroxy (t-test=3,70), n_O le nombre d'atomes d'oxygène (t-test=-9,72) et n_{double} le nombre de liaisons doubles dans la molécule (t-test=-2,65). Ce sont les mêmes descripteurs que les modèles (5. 20) et (5. 21) développés sur le jeu obtenu avec la partition 1/3-2/3. Il s'agit de la même équation (5. 20) que celle obtenue à partir du même jeu d'entraînement mais

avec les descripteurs topologiques et constitutionnels. Les performances sont donc les mêmes (Tableau 46).

c) Comparaison des modèles

Le Tableau 48 synthétise les performances des modèles développés pour la température maximale du pic de décomposition.

Tableau 48 : Modèles obtenus pour Tpic

Modèles - T_{pic}	Nombre de descripteurs	R^2	Q^2	MAE%	R^2_{test}	MAEP%	R^2_{in}	$MAEP_{in}$ %
Tous les descripteurs	3	0,86	0,81	7,52%	0,33	14,45%	0,87	9,11%
Tous les descripteurs - Manuel	4	0,94	0,91	4,83%	0,60	12,06%	0,80	8,30%
Descripteurs 1/2D	2	0,62	0,54	12,53%	0,46	16,06%	0,24	17,30%
Descripteurs 1/2D - Manuel	3	0,84	0 ,79	7,95%	0,38	14,81%	0,57	12,13%
Descripteurs constitutionnels	3	0,76	0,67	9,65%	0,39	14,91%	0,39	14,91%
Descripteurs constitutionnels - Manuel	3	0,84	0,79	7,95%	0,38	14,81%	0,57	12,13%

Comme pour la température onset, l'utilisation de descripteurs dits simples ne suffit pas pour obtenir des modèles prédictifs pour la température maximale du pic de décomposition et les meilleurs modèles sont ceux obtenus à partir de tous les descripteurs. Là encore, le modèle développé avec le jeu d'entrainement manuel (à partir de tous les descripteurs) permet l'obtention de meilleures performances. Il faut aussi noter la similitude des descripteurs présents dans les modèles pour les températures onset et maximale du pic. Cela appuie l'idée présentée dans le paragraphe II.6 qui consiste à utiliser le même modèle (aux coefficients près) pour ces dernières.

VI. AUTRES PROPRIÉTÉS

Dans cette partie, des modèles pour la prédiction de la densité et du point d'éclair des peroxydes organiques seront développés. La sensibilité à l'impact qui a été mesurée dans la cadre du projet mais qui est également une propriété qualitative, n'a pas été modélisée au cours de cette thèse. Étant donné le nombre insuffisant de données expérimentales pour ces deux propriétés, les modèles ne sont pas validés par un jeu de validation. Cependant, des méthodes de validation interne qui nécessitent seulement le jeu d'entraînement ont été employées.

Tout d'abord, pour chaque propriété, les modèles QSPR existants dans la littérature et applicables (pour les peroxydes ou pour les composés organiques en général) ont été utilisés pour prédire les propriétés des peroxydes de la base de données obtenue dans PREDIMOL. Puis, nous avons développés des modèles spécifiques aux peroxydes organiques à partir des données expérimentales de cette même base de données.

1. Densité

La densité (ρ en g/cm^3) d'une substance est une grandeur physique qui correspond au rapport entre sa masse m et son volume V.

(5. 23) $$\rho = m/V$$

a) Modèles existants

Une seule référence a été identifiée pour la prédiction de la densité des peroxydes organiques par modèles QSPR. Romanelli[29] a développé 8 modèles à partir de 14 molécules. Le meilleur modèle (équation (5. 24)), à 9 descripteurs, présente une corrélation R^2=0,9990 mais aucune validation n'a été effectuée.

(5. 24) $$\begin{aligned}\rho = & -3,4395 + 1,8337.10^{-1}SAG - 5,5427.10^{-4}(SAG)^2 + 5,4843.10^{-7}(SAG)^3 \\ & -1,1037.10^{-1}V + 2,3871.10^{-4}V^2 - 1,54144.10^{-7}V^3 + 9,0340.10^{-1}P - 6,1427.10^{-2}P^2 \\ & + 1,3307.10^{-3}P^3\end{aligned}$$

où SAG est la surface moléculaire accessible au solvant, V est le volume moléculaire et P est la polarisabilité calculée selon la formule de Miller[30].

Les revues de Katritzky[31] et Dearden[32] mettent en évidence plusieurs modèles QSPR applicables à la prédiction de la densité de différentes familles de molécules, les modèles de : Karelson[33] qui nécessite la densité relative, Cocchi[34] dont la valeur des coefficients pour chaque descripteur n'est pas disponible ainsi que les modèles de Gakh[35], Zhang[36], Toporov[37] et Cao[38] qui sont applicables aux alcanes uniquement.

Parmi eux, celui de Kuwata[39], développé pour les composés organiques ayant une densité entre 750 et 1900 kg/m^3, utilise uniquement des descripteurs constitutionnels.

(5. 25) $$\rho = 1000\frac{12 + \frac{n_H}{n_C} + 16\frac{n_O}{n_C}}{7,0 + 5,0\frac{n_H}{n_C} + 4,15\frac{n_O}{n_C}}$$

Les modèles de Romanelli et Kuwata seront utilisés pour prédire la densité des peroxydes organiques de la base de données pour lesquels la valeur expérimentale de la densité est disponible.

Tout d'abord, le modèle de Romanelli ayant été développé explicitement pour les hydroperoxydes d'alkyle a été appliqué. Le niveau de calcul n'étant pas précisé, la surface accessible au solvant disponible sur le site chemspider[40] (calculé par ChemAxon[41]) a été utilisée. De plus, le niveau DFT avec la fonctionnelle PBE0 est considéré comme fiable et a été utilisé pour optimiser les structures à partir desquels le volume moléculaire est calculé (avec le logiciel Codessa). Finalement, la

polarisabilité est calculée par la méthode de contribution de Miller, comme indiqué dans l'article. Les prédictions ont été faites pour les deux hydroperoxydes de la base de données pour lesquels la densité est renseignée (tert-amyl hydroperoxide et tert-butyl hydroperoxide). Les prédictions obtenues présentent une erreur moyenne très élevée (9 g/cm^3 soit 10%). Néanmoins, il faut noter que l'application de l'équation n'a pas été faite avec le logiciel ChemPlus avec lequel les descripteurs du jeu d'entraînement ont été calculés. Cependant, dans tous les cas, le domaine d'applicabilité de ce modèle (les hydroperoxydes d'alkyle) reste très limité et le modèle est sur-paramétré avec 9 descripteurs pour une base de données de 14 molécules.

L'application du modèle Kuwata aux peroxydes organiques de notre base de données donne un R^2 de 0,39 et une MAE de 138 kg/m^3 (soit 15%), ce qui est assez élevé pour cette propriété. Ce modèle a été développé pour les composés organiques, cependant il faut noter que la base de données utilisée ne contient pas de peroxydes organiques.

Les deux modèles identifiés pour la prédiction de la densité par modèle QSPR, applicables au peroxydes organiques, sont soit trop spécifiques (Romanelli) soit trop larges (Kuwata) et présentent par conséquent de mauvaises prédictions. C'est dans ce cadre que nous avons développé un modèle QSPR pour les peroxydes organiques incluant tous les types de familles (des peroxydes organiques).

b) Développement de modèle QSPR pour la densité

Parmi les peroxydes organiques de la base de données obtenue dans PREDIMOL, seulement 30 molécules ont une valeur pour cette propriété dont la plupart des valeurs dans la littérature sont des valeurs prédites et qui de plus ne sont pas toujours en accord avec celle mesurées dans le projet.

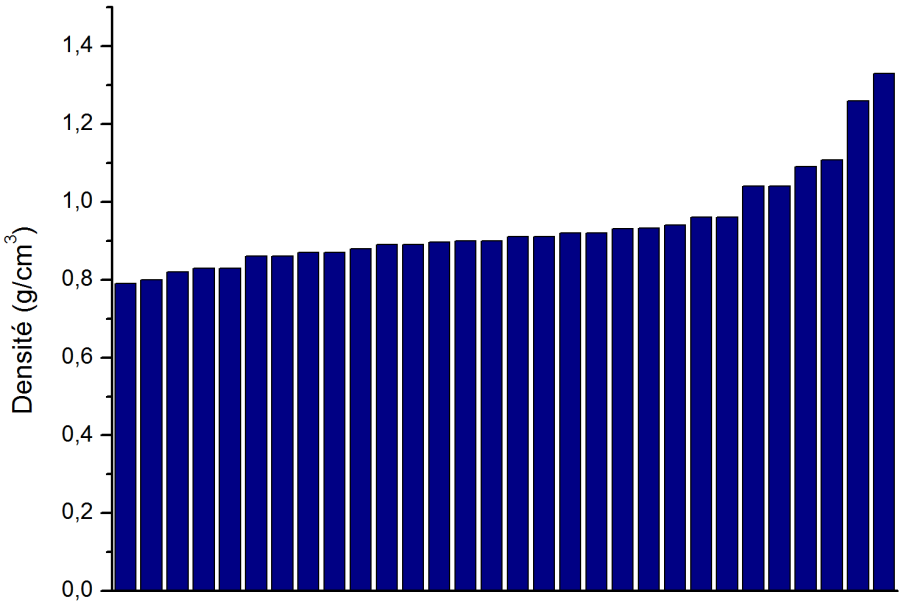


Figure 66: Diagramme des valeurs expérimentales pour la densité des 30 peroxydes organiques

Des modèles ont été développés avec ces 30 molécules mais il a été observé que le 2,5-di(tert-butylperoxy)-2,5-dimethyl-3-hexyne et le dibenzoyl peroxide influencent beaucoup la régression (voir Figure 67 avec la meilleure régression linéaire : ρ=-1,58 n_H/mw + 1,87). Elles ont donc été supprimées du jeu de données qui est donc réduit à 28 peroxydes organiques.

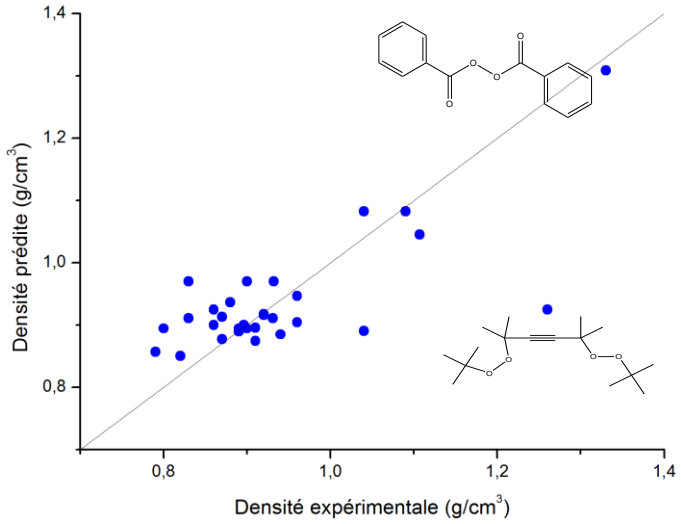


Figure 67 : Valeur de densité prédite par la meilleure équation linéaire vs. densité expérimentale

À partir de ces 28 peroxydes organiques et tous les descripteurs, un modèle à 2 descripteurs est obtenu (équation (5. 26), Figure 68 et Tableau 49) :

(5. 26) $$\rho = 0{,}813 n_{Ar,r} + 0{,}01222\ \mathrm{DPSA3} + 0{,}7232$$

Avec $n_{Ar,r}$ le nombre relatif de liaisons aromatiques (t-test=8,71) et DPSA3 la différence entre les surfaces partielles chargées positives et celles chargées négatives (t-test=7,61).

Aucun de ces descripteurs n'est lié à la liaison peroxy ou aux atomes d'oxygène, ce qui n'est pas étonnant puisque cette propriété est complètement différente de celles étudiées précédemment. En effet, la densité est une propriété physico-chimique qui n'est en aucun cas liée aux dangers et à la décomposition des peroxydes organiques.

Tableau 49: performances du modèle pour la densité

R^2	Q^2	RMSE	MAE	MAE(%)	$Q^2$5cv	$Q^2$10cv	$Q^2$7cv	R^2_{YS}	σ_{YS}
0,83	0,78	0,03	0,03	2,94%	0,74	0,76	0,76	0,07	0,07

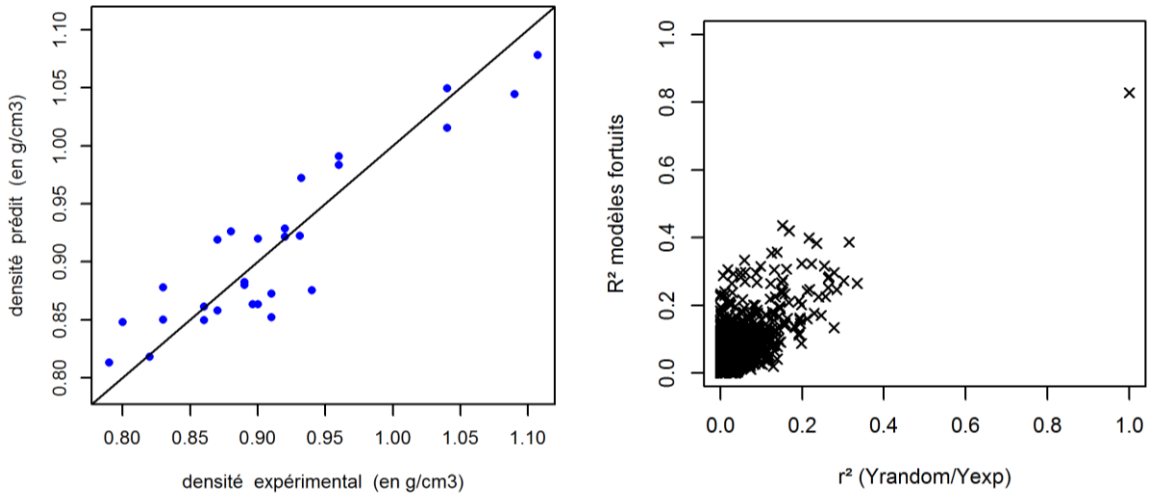


Figure 68 : Représentation des données expérimentales vs données prédites pour la densité; Représentation des résultats de la procédure d'Y-scrambling

Ce modèle, bien que non validé par un jeu de validation, présente de très bonnes performances (MAE≈3% contre une incertitude moyenne de 12% pour le modèle de Kuwata) avec une procédure de Y-scrambling représenté Figure 68 qui illustre bien le fait que le modèle n'a pas été obtenu par chance. Seulement deux descripteurs sont nécessaires pour une base de données de 28 molécules, ce qui écarte toute possibilité de sur-apprentissage malgré le petit nombre de données expérimentales. La mesure et le calcul de la densité pour les autres composés de la base de données PREDIMOL serait une bonne façon de valider. En effet les mesures seraient faites de manière homogène aux mesures utilisées pour le jeu d'entraînement.

2. Point d'éclair

Le point d'éclair est la température la plus basse à laquelle un corps combustible émet suffisamment de vapeurs pour former, à une pression standard à 101,3 kPa, un mélange gazeux qui s'enflamme sous l'effet d'une source d'énergie calorifique telle qu'une flamme ou une étincelle, mais pas suffisamment pour que la combustion s'entretienne d'elle-même.

Le point d'éclair est l'une des caractéristiques importantes pour les propriétés d'inflammabilité des liquides et des substances à faible point de fusion. Il fournit un indice simple et pratique, de l'inflammabilité et de la combustibilité des substances. Il est important car il donne des connaissances nécessaires pour la manutention et le transport du composé en grandes quantités ainsi qu'en termes de sécurité.

a) Modèles existants

De manière générale, de nombreux modèles ont été développés pour la prédiction du point d'éclair pour des produits purs et ont été recensés par Katritzky[31]. Certains d'entre eux sont basés sur d'autres propriétés telles que le point d'ébullition[42,43]. Aucun modèle QSPR n'a été identifié pour les

peroxydes uniquement. Cependant, des modèles ayant un domaine d'applicabilité plus large comme ceux de Katritzky[44,45], Stefanis[46], Stayanarayana[47], Gharaghetzi[48], Rowley[49], Bagheri[50] et Caroll[51] sont disponibles dans la littérature. Certains d'entre eux, ne nécessitant pas la valeur de la température d'ébullition expérimentale et dont les descripteurs peuvent être calculés avec les logiciels à notre disposition, ont été utilisés pour prédire le point d'éclair des peroxydes disponibles de la base de données obtenue dans le cadre du projet PREDIMOL.

En 2001, Katritzky[44] a développé plusieurs modèles à partir de 271 composés mais aucun peroxyde organique n'a été identifié parmi eux. Cependant, nous tenterons tout de même de prédire le point d'éclair des composés de notre base de données à partir de ces modèles qui présentent des performances intéressantes. L'erreur calculée est de 74 K pour le premier modèle et de 55 K pour le deuxième modèle applicable. En 2007, une mise à jour avec une base de données de 758 composés organiques (dont un peroxyde) a été publiée et l'application du modèle obtenu aux composés de notre base de données donne une erreur de 93 K. Ces modèles ne sont donc pas utilisables pour les peroxydes organiques.

En 2010, Rowley[49] a proposé un modèle, basé sur la méthode de contribution de groupe, développé sur plus de 1000 molécules dont la composition n'est pas connue. Cependant, l'absence du fragment –OO– laisse supposer que les peroxydes organiques ne sont pas en très grand nombre, voire pas du tout présents. L'application de ce modèle sur les peroxydes de notre base de données donne une erreur de 33,47 K.

En considérant qu'aucun modèle QSPR n'existe pour les peroxydes organiques et que les modèles existants sont la plupart du temps entraînés sur de grandes bases ne contenant pas de peroxydes organiques, nous avons développé un modèle QSPR pour la prédiction du point d'éclair des peroxydes organiques. Le nombre de données expérimentales à notre disposition ne permet pas une validation externe mais une validation interne sera effectuée.

b) Développement de modèle QSPR pour le point d'éclair

Un modèle préliminaire a été développé et présenté lors du congrès *Loss Prevention and safety Promotion in the Process Industries*[52] avec 23 molécules. Il s'agit d'un modèle à 5 descripteurs, développé à partir de plus de 300 decripteurs calculés par Codessa qui présente de très bonnes performances : R^2=0,92, Q^2 = 0,87, Q^2_{10CV}= 0,87 et Q^2_{5CV}=0,89. Cependant, ce modèle prend en compte deux molécules dont les valeurs extrêmes (Figure 70) influencent la régression (Figure 69).

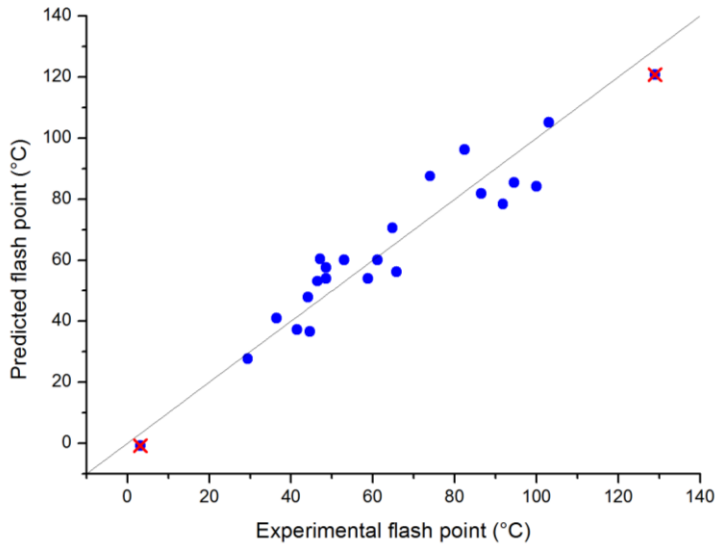


Figure 69 : Valeurs prédites en fonction des valeurs expérimentale pour le point d'éclair par le modèle préliminaire (les croix rouges représentent les deux valeurs extrêmes influençant la régression).

Le nombre final de données expérimentales disponibles pour cette propriété est de 24 molécules (Tableau 21). D'autre part, comme on peut le voir sur la Figure 70 et Figure 69, deux molécules ont des valeurs extrêmement différentes : di-tert-butyl peroxyde et dicumyl peroxyde. Il ne reste alors plus que 22 molécules pour développer un modèle, sans validation externe. Ces deux molécules n'avaient pas été supprimées dans le modèle préliminaire, ci-dessus.

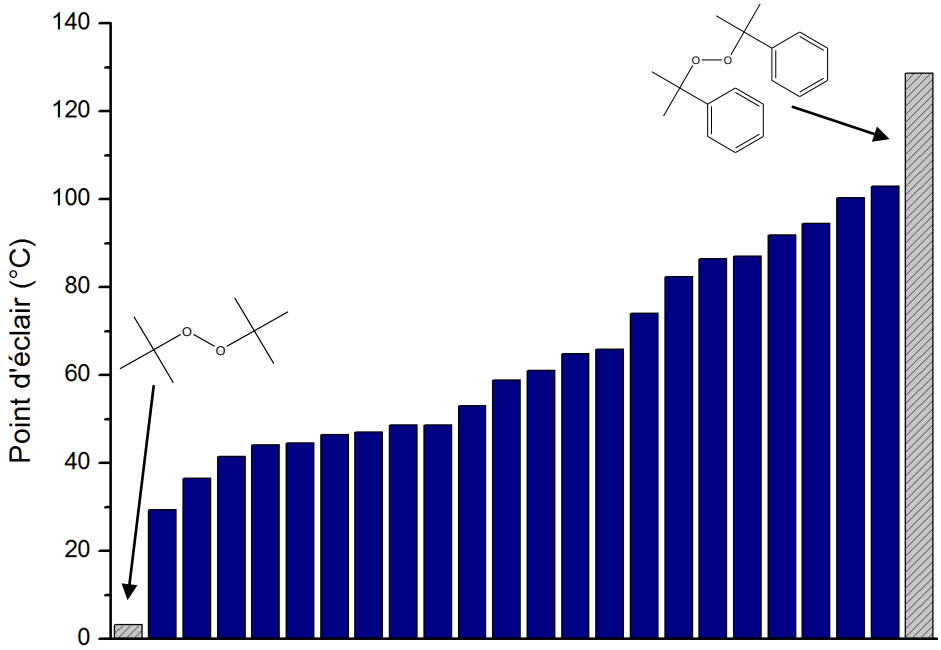


Figure 70 : Diagramme des valeurs expérimentales pour le point d'éclair des 24 peroxydes organiques

Au final, un modèle à deux descripteurs est obtenu (équation (5. 27) et Figure 71) à partir de tous les descripteurs :

(5. 27)
$$T_{eclair} = 379\mathrm{V}_{\mathrm{max,C}} - 30670\mathrm{N}_{\mathrm{min,C}} - 1423$$

Avec $V_{max,C}$ la valence maximale pour un atome de carbone (t-test=-3,33) et $N_{min,C}$ l'indice de réactivité maximale pour un atome de carbone.

Les performances du modèle (5. 27) sont nettement inférieures à celle du modèle présenté au congrès. En effet, l'ajutement passe de R^2=0,92 à 0,63 et la robutesse chute également (Q^2 = 0,87 à Q^2 = 0,49)

Tableau 50: performances du modèle pour le point d'éclair

R^2	Q^2	RMSE	MAE	MAE(%)	$Q^2$5cv	$Q^2$10cv	$Q^2$7cv	R^2_{YS}	σ_{YS}
0,63	0,49	15	10	17,37%	0,53	0,52	0,53	0,09	0,09

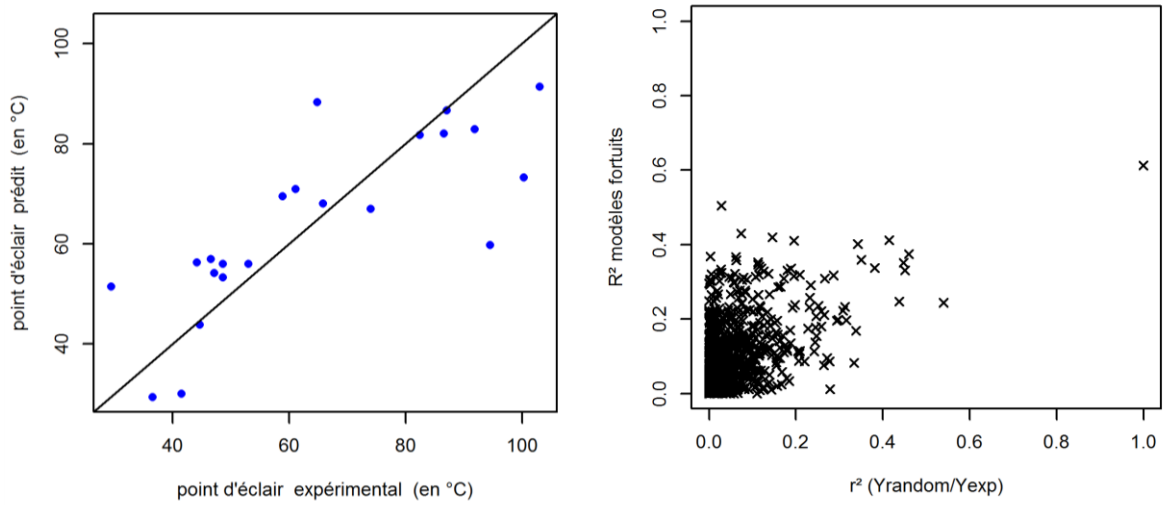


Figure 71 : Représentation des données expérimentales vs données prédites pour le point d'éclair; Représentation des résultats de la procédure d'Y-scrambling

Cette équation n'est pas un modèle : les performances obtenues sont médiocres (erreur de 19°C) et la validation interne montre qu'elle n'est pas robuste et qu'elle peut être le résultat d'une corrélation par chance (Figure 71).

VII. CONCLUSION

Dans le cadre du projet PREDIMOL, une base de données expérimentale de 38 peroxydes organiques à haute concentration a été construite. Elle a permis le développement des premiers modèles QSPR validés pour la prédiction des propriétés liées à la stabilité thermique (chaleur de décomposition divisée par la concentration, température onset et température maximale du pic de décomposition) des peroxydes organiques.

Les cinq principes de l'OCDE pour la validation des modèles QSAR/QSPR ont été suivis lors du développement des modèles QSPR. Ainsi, une validation interne a été effectuée avec notamment la procédure de Y-scrambling. Un jeu de validation a été utilisé pour la validation externe et le calcul de

la prédictivité des modèles. Le domaine d'applicabilité a aussi été défini. Trois modèles prédictifs ont donc été obtenus (à partir de tous les descripteurs) pour la chaleur de décomposition divisée par la concentration (ΔH/C avec les performances suivantes : MAE=5,66%, R^2=0,97, Q^2=0,94, et MAEP=$MAEP_{in}$=14,47%), la température onset (MAE=8,43%, R^2=0,84, Q^2=0,77, MAEP=9,97% et $MAEP_{in}$=9,33%) et la température maximale du pic de décomposition (MAE=7,52%, R^2=0,86, Q^2=0,81, MAEP=14,45% et $MAEP_{in}$=9,11%).

Dans le but d'obtenir des modèles plus simples à utiliser, notamment dans le cas de la prédiction de propriété pour l'enregistrement des substances par les industriels dans le cadre de REACH, des modèles ont été développés à partir de descripteurs qui ne nécessitent pas la détermination de la géométrie des molécules (c'est-à-dire sans calculs quantiques longs et complexes). Dans un premier temps, les descripteurs topologiques et constitutionnels ont servis pour le développement de modèles pour les trois propriétés précédentes. Un modèle performant a été obtenu pour ΔH/C seulement (MAE=5,89%, R^2=0,95, Q^2=0,91 et MAEP=$MAEP_{in}$=15,72%). Puis des modèles encore plus simples (descripteurs constitutionnels uniquement) ont été obtenus (avec des performances plus faibles). L'utilisation de descripteurs simples est intéressante mais pas toujours concluante et les performances dépendent de la propriété considérée.

Des modèles QSPR ont aussi été développés sur un jeu d'entraînement différent qui est obtenu en se basant sur la structure des molécules avant de considérer la valeur de la propriété. Ce second jeu d'entraînement (dit « manuel ») donne accès à d'autres modèles plus prédictifs ou équivalents aux précédents. En effet, cette méthode de partage implique une meilleure répartition des molécules des jeux dans l'espace des descripteurs. Comme, de plus, une partie de la répartition s'appuie aussi sur la distribution de la propriété, ce découpage donne des modèles avec de meilleures performances.

Au final, 18 modèles ont été obtenus (6 par propriété, pouvant être utilisés dans une approche consensus) parmi lesquels un descripteur se distingue : le nombre de liaisons peroxy (n_{OO}) qui est présent dans la plupart des modèles. La présence répétitive de ce descripteur confirme le caractère important de cette liaison pour la stabilité thermique des peroxydes organiques avec le fait que la rupture de la liaison peroxy est la première étape de décomposition de ces composés.

Les températures onset et maximale du pic de décomposition sont, quant à elles, liées en plus au descripteur fonction de Fukui localisée sur la liaison peroxy (F^-_{OO}). Issu de la DFT conceptuelle, ce descripteur représente la réactivité de cette liaison et souligne encore une fois le rôle important de cette liaison dans la décomposition des peroxydes organiques.

La densité et le point d'éclair ont été modélisés bien que le faible nombre de données expérimentales ne permette pas une validation externe. Ces modèles présentent des performances encourageantes, aussi l'acquisition de données expérimentales pour ces propriétés est recommandée pour obtenir un modèle robuste et prédictif.

VIII. RÉFÉRENCES

(1) Chang, R. H.; Tseng, J. M.; Jehng, J. M.; Shu, C. M.; Hou, H. Y. Thermokinetic model simulations for methyl ethyl ketone peroxide contaminated with H2SO4 OR NaOH by DSC and VSP2. *J Therm Anal Calorim* **2006**, *83*, 57–62.

(2) Tseng, J.-M.; Chang, Y.-Y.; Su, T.-S.; Shu, C.-M. Study of thermal decomposition of methyl ethyl ketone peroxide using DSC and simulation. *Journal of Hazardous Materials* **2007**, *142*, 765–770.

(3) Hou, H.-Y.; Shu, C.-M.; Tsai, T.-L. Reactions of cumene hydroperoxide mixed with sodium hydroxide. *Journal of Hazardous Materials* **2008**, *152*, 1214–1219.

(4) Levin, M. E.; Gonzales, N. O.; Zimmerman, L. W.; Yang, J. Kinetics of acid-catalyzed cleavage of cumene hydroperoxide. *Journal of Hazardous Materials* **2006**, *130*, 88–106.

(5) Lu, Y.; Ng, D.; Mannan, M. S. Prediction of the Reactivity Hazards for Organic Peroxides Using the QSPR Approach. *Industrial & Engineering Chemistry Research* **2011**, *50*, 1515–1522.

(6) Peduzzi, P.; Concato, J.; Feinstein, A. R.; Holford, T. R. Importance of events per independent variable in proportional hazards regression analysis II. Accuracy and precision of regression estimates. *Journal of Clinical Epidemiology* **1995**, *48*, 1503–1510.

(7) Golbraikh, A.; Tropsha, A. Predictive QSAR modeling based on diversity sampling of experimental datasets for the training and test set selection. *Journal of Computer-Aided Molecular Design* **2002**, *16*, 357–369.

(8) Puzyn, T.; Mostrag-Szlichtyng, A.; Gajewicz, A.; Skrzyński, M.; Worth, A. P. Investigating the influence of data splitting on the predictive ability of QSAR/QSPR models. *Structural Chemistry* **2011**, *22*, 795–804.

(9) Ando, T.; Fujimoto, Y.; Morisaki, S. Analysis of differential scanning calorimetric data for reactive chemicals. *Journal of Hazardous Materials* **1991**, *28*, 251–280.

(10) Fayet, G.; Rotureau, P.; Joubert, L.; Adamo, C. On the prediction of thermal stability of nitroaromatic compounds using quantum chemical calculations. *Journal of Hazardous Materials* **2009**, *171*, 845–850.

(11) Fayet, G.; Rotureau, P.; Joubert, L.; Adamo, C. Development of a QSPR model for predicting thermal stabilities of nitroaromatic compounds taking into account their decomposition mechanisms. *Journal of Molecular Modeling* **2010**, *17*, 2443–2453.

(12) Sang, P.; Zou, J.-W.; Xu, L.; Liu, Y.-H. QSPR of Thermal Stability of Nitroaromatic Explosives using Theoretical Descriptors Derived from Electrostatic Potentials on the Molecular Surface. *chinese journal of structural chemistry* **2011**, *30*, 533–537.

(13) Sang, P.; Zou, J.; Xu, L.; Zhou, P. Linear and Nonlinear QSPR Models for Predicting Thermal Stabilities of Nitroaromatic Compounds. *Chemical research in chinese university* **2011**, *27*, 891–895.

(14) Atalar, T.; Zeman, S. A New View of Relationships of the N-N Bond Dissociation Energies of Cyclic Nitramines. Part I. Relationships with Heats of Fusion. *Journal of Energetic Materials* **2009**, *27*, 186–199.

(15) Keshavarz, M. H. Predicting heats of fusion of nitramines. *Indian journal of engineering & materials sciences* **2007**, *14*, 386–390.

(16) Zeman, S. Some predictions in the field of the physical thermal stability of nitramines. *Thermochimica Acta* **1997**, *302*, 11–16.

(17) Gharagheizi, F.; Sattari, M.; Ilani-Kashkouli, P.; Mohammadi, A. H.; Ramjugernath, D.; Richon, D. Quantitative structure-property relationship for thermal decomposition temperature of ionic liquids. *Chemical Engineering Science* **2012**, *84*, 557–563.

(18) Yu, X.; Xie, Z.; Yi, B.; Wang, X.; Liu, F. Prediction of the thermal decomposition property of polymers using quantum chemical descriptors. *European Polymer Journal* **2007**, *43*, 818–823.

(19) Tenenhaus, M.; Gauchi, J.-P.; Ménardo, C. Régression PLS et applications. *Revue de Statistique Appliquée* **1995**, *43*, 7–63.

(20) Karelson, M. *Molecular Descriptors in QSAR/QSPR*; Wiley: New York, 2000.

(21) Kier, L. B. A Shape Index from Molecular Graphs. *Quantitative Structure-Activity Relationships* **1985**, *4*, 109–116.

(22) *Scigress*; FUJITSU, 2008.

(23) Muehlbacher, M.; Kerdawy, A. E.; Kramer, C.; Hudson, B.; Clark, T. Conformation-Dependent QSPR Models: logPOW. *J. Chem. Inf. Model.* **2011**, *51*, 2408–2416.

(24) Golbraikh, A.; Shen, M.; Xiao, Z. Y.; Xiao, Y. D.; Lee, K. H.; Tropsha, A. Rational selection of training and test sets for the development of validated QSAR models. *Journal of Computer-Aided Molecular Design* **2003**, *17*, 241–253.

(25) Prana, V.; Fayet, G.; Rotureau, P.; Adamo, C. Development of validated QSPR models for impact sensitivity of nitroaliphatic compounds. *J. Hazard. Mater.* **2012**, *235-236*, 169–177.

(26) Fayet, G.; Joubert, L.; Rotureau, P.; Adamo, C. On the use of descriptors arising from the conceptual density functional theory for the prediction of chemicals explosibility. *Chemical Physics Letters* **2009**, *467*, 407–411.

(27) Kamlet, M. J. The relationship of impact sensitivity with structure of organic high explosives. I Polynitroaliphatic explosives. In; Coronado, California, 1976; p. 312.

(28) Arrêté du 29 mai 2009 relatif au transport de marchandises dangereuses par voies terrestres (dit "arrêté TMD") NOR: DEVP1241087A Version consolidée au 01 janvier 2013 **2013**.

(29) Romanelli, G. P.; Cafferata, L. R. F.; Castro, E. A. Ameliorate QSPR Study of Alkyl Hydroperoxides. *Russian Journal of General Chemistry* **2001**, *71*, 257–260.

(30) Miller, K. J. Additivity methods in molecular polarizability. *J. Am. Chem. Soc.* **1990**, *112*, 8533–8542.

(31) Katritzky, A. R.; Kuanar, M.; Slavov, S.; Hall, C. D.; Karelson, M.; Kahn, I.; Dobchev, D. A. Quantitative correlation of physical and chemical properties with chemical structure: utility for prediction. *Chem. Rev.* **2010**, *110*, 5714–5789.

(32) Dearden, J. C.; Worth, A. In Silico Prediction of Physicochemical Properties. *European Commission, Joint Research Centre* **2007**.

(33) Karelson, M.; Perkson, A. QSPR prediction of densities of organic liquids. *Computers & Chemistry* **1999**, *23*, 49–59.

(34) Cocchi, M.; De Benedetti, P. G.; Seeber, R.; Tassi, L.; Ulrici, A. Development of Quantitative Structure−Property Relationships Using Calculated Descriptors for the Prediction of the Physicochemical Properties (nD, ρ, bp, ε, η) of a Series of Organic Solvents. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 1190–1203.

(35) Gakh, A. A.; Gakh, E. G.; Sumpter, B. G.; Noid, D. W. Neural Network-Graph Theory Approach to the Prediction of the Physical Properties of Organic Compounds. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 832–839.

(36) Zhang, R.; Liu, S.; Liu, M.; Hu, Z. Neural network-molecular descriptors approach to the prediction of properties of alkenes. *Computers & Chemistry* **1997**, *21*, 335–341.

(37) Toropov, A. A.; Toropova, A. P. QSPR modeling of alkanes properties based on graph of atomic orbitals. *Journal of Molecular Structure: THEOCHEM* **2003**, *637*, 1–10.

(38) Cao, C.; Shuo Gao Density Models for Alkanes and Monoderivatives of Hydrocarbons. *Internet Electron. J. Mol. Des.* **2005**, 671–697.

(39) Kuwata, M.; Zorn, S. R.; Martin, S. T. Using Elemental Ratios to Predict the Density of Organic Material Composed of Carbon, Hydrogen, and Oxygen. *Environ. Sci. Technol.* **2012**, *46*, 787–794.

(40) Royal Society of Chemistry ChemSpider - The free chemical database http://www.chemspider.com/ (accessed Jun 12, 2013).

(41) ChemAxon chemicalize.org http://www.chemicalize.org/ (accessed Jun 12, 2013).

(42) Hshieh, F.-Y. Correlation of closed-cup flash points with normal boiling points for silicone and general organic compounds. *Fire and Materials* **1997**, *21*, 277–282.

(43) Vidal, M.; Rogers, W. J.; Holste, J. C.; Mannan, M. S. A review of estimation methods for flash points and flammability limits. *Process Safety Progress* **2004**, *23*, 47–55.

(44) Katritzky, A. R.; Petrukhin, R.; Jain, R.; Karelson, M. QSPR Analysis of Flash Points. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1521–1530.

(45) Katritzky, A. R.; Stoyanova-Slavova, I. B.; Dobchev, D. A.; Karelson, M. QSPR modeling of flash points: An update. *Journal of Molecular Graphics and Modelling* **2007**, *26*, 529–536.

(46) Stefanis, E.; Constantinou, L.; Tsivintzelis, I.; Panayiotou, C. New Group-Contribution Method for Predicting Temperature-Dependent Properties of Pure Organic Compounds. *Int J Thermophys* **2005**, *26*, 1369–1388.

(47) Satyanarayana, K.; Rao, P. G. Improved equation to estimate flash points of organic compounds. *Journal of Hazardous Materials* **1992**, *32*, 81–85.

(48) Gharagheizi, F.; Eslamimanesh, A.; Mohammadi, A. H.; Richon, D. Empirical Method for Representing the Flash-Point Temperature of Pure Compounds. *Ind. Eng. Chem. Res.* **2011**, *50*, 5877–5880.

(49) Rowley, J. r.; Rowley, R. l.; Wilding, W. v. Estimation of the flash point of pure organic chemicals from structural contributions. *Process Safety Progress* **2010**, *29*, 353–358.

(50) Bagheri, M.; Bagheri, M.; Heidari, F.; Fazeli, A. Nonlinear molecular based modeling of the flash point for application in inherently safer design. *Journal of Loss Prevention in the Process Industries* **2012**, *25*, 40–51.

(51) Carroll, F. A.; Lin, C.-Y.; Quina, F. H. Simple Method to Evaluate and to Predict Flash Points of Organic Compounds. *Ind. Eng. Chem. Res.* **2011**, *50*, 4796–4800.

(52) Fayet, G.; Rotureau, P.; Prana, V.; Adamo, C. Prediction of physico-chemical properties for REACH based on QSPR models. *Chemical Engineering Transactions* **2013**, *31*, 925–930.

CONCLUSION

La réglementation européenne REACH demande l'enregistrement, d'ici 2018, de toutes les substances chimiques produites ou importées en Europe à plus d'une tonne par an. Les dossiers d'enregistrement de ces substances requièrent l'évaluation de leurs propriétés physico-chimiques, toxicologiques et écotoxicologiques pour autoriser leur utilisation et mise sur le marché. Étant donné le grand nombre de propriétés et de substances, le développement de méthodes alternatives à la caractérisation expérimentale est encouragé. C'est dans ce contexte que le projet ANR PREDIMOL (PREDIction des propriétés physico-chimiques des produits par modélisation MOLéculaire), où la modélisation moléculaire représente une de ces voies, a été mis en place en 2010.

L'objectif de cette thèse, intégrée dans ce projet de recherche, était le développement de modèles QSPR (*Quantitative Structure-Property Relationship*) prédictifs pour la caractérisation des propriétés physico-chimiques des peroxydes organiques pour lesquels aucun modèle prédictif et validé n'avait été développé. La méthode *Best Multi-Linear Regression* (BMLR) implémentée dans Codessa a été utilisée pour le développement de ces modèles. Cette méthode intervient au niveau de la sélection des descripteurs et permet l'obtention de modèles multilinéaires. Auparavant, la base de données est partagée en deux jeux de molécules : le jeu d'entraînement sur lequel le modèle est développé, qui sert aussi à la détermination du domaine d'applicabilité, et le jeu de validation avec lequel le pouvoir prédictif du modèle est calculé. Plus de 300 descripteurs ont été calculés par Codessa à partir de la structure des peroxydes optimisée au niveau DFT (PBE0//6-31+G(d,p))

Pour commencer, des modèles ont été développés à partir de la base de données Datatop du TNO qui recense les résultats des essais expérimentaux pour la classification des peroxydes organiques selon la réglementation relative au transport de marchandises dangereuses (TMD) pour plus d'une centaine de molécules (les valeurs de toutes les propriétés ne sont pas nécessairement connues pour chacune des molécules). Cette base de données est une compilation de résultats obtenus par différents organismes dans des conditions expérimentales différentes. En particulier, les concentrations des peroxydes organiques peuvent être différentes et parfois la valeur n'est pas connue exactement. Les résultats de cette étude ne sont pas concluants car aucun modèle performant n'a pu être obtenu. Des modèles meilleurs en ajustement ont été obtenus pour les peroxyesters (famille de peroxydes majoritaire dans cette base de données avec 34 composés) mais la validation interne montre un manque de robustesse et une forte probabilité de corrélation par

chance. Ces modèles ne sont donc pas utilisables pour la prédiction des propriétés de la Datatop. Le modèle le plus prometteur est celui développé pour la température de décomposition auto-accélérée (TDAA) à partir de 22 peroxyesters, avec une équation à 4 paramètres, dont l'erreur en ajustement est de 5°C. L'amélioration des performances de ce modèle nécessiterait la validation sur un jeu de molécules externes. Cependant, cette propriété dépend de plusieurs paramètres expérimentaux comme le type d'emballage contenant les peroxydes organiques. Cette indication figure rarement dans les fiches de données de sécurité (FDS) des peroxydes organiques et n'existe pas non plus dans la Datatop.

Suite à cette première démarche, une seconde base de données a été construite pour des peroxydes organiques pouvant être transportés et disponibles sur le marché européen. Les mesures des propriétés liées à la stabilité thermique (chaleur de début de décomposition, température onset, température maximale du pic) de ces molécules ont été réalisées dans les mêmes conditions expérimentales dans le cadre du projet PREDIMOL par les partenaires Arkema et INERIS. Cette base de données de 38 peroxydes organiques (dont la concentration est connue exactement) a servi pour le développement de nouveaux modèles validés et prédictifs.

Ainsi, en considérant tous les types de descripteurs et une répartition des molécules de la base de données entre jeux d'entraînement et de validation effectuée sur la valeur de la propriété (⅓-⅔), trois modèles multilinéaires prédictifs ont été obtenus. Le premier, pour la chaleur de décomposition divisée par la concentration, est un modèle à 4 descripteurs qui présente une erreur de 14% en prédictivité. Il présente une meilleure robustesse (Q^2=0,94 contre -0,81) que le modèle de Lu pour la chaleur de décomposition (seule référence dans la littérature). Le second modèle, à 3 descripteurs, est obtenu pour la prédiction de la température onset avec une $MAEP_{in}$=9,33% (erreur en prédictivité dans le domaine d'applicabilité). Ce modèle présente de meilleures performances que celui de Lu à 4 descripteurs, avec une robustesse de 0,77 contre 0,11 et une erreur de 11 °C (MAEP) en prédictivité contre 62 °C. Le troisième modèle, à 3 descripteurs, qui a été développé pour la température maximale du pic de décomposition, présente une $MAEP_{in}$=9,11%. Rappelons que la température onset et la température maximale du pic de décomposition sont deux propriétés liées, qui peuvent être décrites par les mêmes descripteurs. Ainsi, une seule équation (aux coefficients près) pourrait suffire pour la prédiction de ces deux propriétés.

L'influence de différents paramètres (les jeux de molécules, les descripteurs) sur les performances des modèles a été étudiée. Des jeux d'entraînement et de validation ont été réalisés avec deux méthodes différentes pour le développement des modèles. La première méthode de partage repose

sur la valeur de la propriété uniquement, la seconde méthode dite « manuelle » est avant tout basée sur la structure des molécules en utilisant l'approche d'analyse en composantes principales (ACP ou PCA). Il a été remarqué que les modèles développés sur les jeux dits « manuels » ont généralement des performances meilleures ou équivalentes aux jeux obtenus en considérant la valeur de la propriété uniquement.

De même, différents types de descripteurs (tous, constitutionnels et topologiques, constitutionnels uniquement) ont été utilisés lors de l'entraînement des modèles. Dans le premier cas, la prise en compte de tous les descripteurs dont les descripteurs quantiques, en particulier ceux issus de la DFT conceptuelle, donne accès à des modèles avec des descripteurs chimiquement interprétables. En effet, pour les trois premiers modèles, on peut observer plusieurs descripteurs quantiques (F^-_{OO}, E_{HOMO}, $R_{avg,O}$, $gap_{HOMO-LUMO}$, d_{OO} et Q_{OO}). On peut donc constater que ces descripteurs sont utiles pour obtenir de bonnes prédictions mais aussi une confirmation du mécanisme de décomposition. Dans un second temps, l'utilisation de descripteurs « simples » uniquement a été considérée dans le but d'obtenir des modèles faciles à utiliser par un industriel ou une instance réglementaire dans le cadre des dossiers d'enregistrement de REACH.

Finalement, six modèles ont été obtenus pour chacune de ces trois propriétés liées à la stabilité thermique des peroxydes organiques en considérant les différents types de descripteurs et la répartition des données entre le jeu d'entraînement et de validation. Ces modèles ont tous été développés en accord avec les principes de l'OCDE pour la validation des modèles QSPR/QSAR, à savoir :

1) Une propriété ciblée définie (avec un protocole expérimental identifié) ;
2) Un algorithme sans équivoque (BMLR) ;
3) Un domaine d'applicabilité défini ;
4) Des mesures appropriées de la qualité d'ajustement (R^2, MAE, RMSE), de robustesse (Q^2, Y-scrambling) et du pouvoir prédictif (MAEP, RMSEP, Q^2_{F1}, Q^2_{F2}, Q^2_{F3}, CCC) ;
5) Si possible, une interprétation des mécanismes sous-jacents.

Seulement quelques uns sont réellement prédictifs et validés selon ces principes (9/18 modèles). Ainsi, ceux qui présentent de bonnes performances parmi les plus simples pourraient être soumis pour leur intégration dans la Toolbox de l'OCDE et de l'ECHA pour une meilleure visibilité et une utilisation par les industriels.

Dans le cas de la stabilité thermique, le cinquième principe de l'OCDE correspond à une interprétation des mécanismes de décomposition. Pour les peroxydes organiques, la rupture de la

liaison peroxy est considérée comme la première étape de ces mécanismes. Cela semble être confirmé par l'apparition redondante du descripteur « nombre de liaisons peroxy » dans les différents modèles sur la stabilité thermique développés dans cette thèse. Néanmoins, aucune relation directe entre l'énergie de dissociation de cette liaison (calculée pour 109 peroxydes organiques) et les descripteurs liés à la chimie des peroxydes (en particulier la liaison peroxy) ou les propriétés expérimentales liées à la décomposition n'a été identifiée. Une étude centrée sur les mécanismes de décomposition des peroxydes organiques en distinguant les différentes familles serait intéressante à réaliser.

D'autres modèles ont été obtenus pour la densité et le point d'éclair. Cependant, seule une validation interne a été effectuée car le nombre de données disponibles pour ces deux propriétés ne permet pas pour l'instant une validation externe avec un jeu de validation. La mesure de ces propriétés manquantes dans la base de données et une augmentation du nombre de peroxydes organiques de la base de données PREDIMOL serait une bonne solution pour valider ces modèles ou en développer de nouveaux.

Enfin, bien que les modèles multilinéaires développés soient performants, d'autres types de méthodes (régression des moindres carrés partiels (PLS), réseaux de neurones, algorithmes génétiques) peuvent également être utilisés selon les propriétés considérées. En particulier, le développement de modèles pour les propriétés qualitatives, comme la sensibilité à l'impact des peroxydes organiques, qui nécessitent des méthodes différentes (telles que les arbres de décisions ou séparateurs à vaste marge (SVM)) serait une poursuite envisageable de ces travaux de thèse.

Au niveau technique, les perspectives pour le développement des modèles seraient de poursuivre les efforts en direction de la validation des modèles par l'utilisation de la méthode de bootstrap (validation interne par ré-échantillonnage) dont les résultats sont aussi demandés dans les fichiers QMRF (pour *QSAR Model Reporting Format*), structurés selon les principes de validation de l'OCDE qui récapitulent les informations clés sur les modèles et sont nécessaires pour la reconnaissance réglementaire des modèles QSPR. De même, la mise en place d'une sélection automatique des variables pour le développement de modèles pourrait être faite dans le but d'effectuer la procédure de Y-scrambling (mélange des valeurs expérimentales de la propriété) en incluant la sélection des descripteurs. Toujours dans un esprit de simplification des modèles pour leur acceptation réglementaire et leur utilisation, la mise en place d'un domaine d'applicabilité plus simple à utiliser est à creuser.

En parallèle, un programme (Callisto pour *Conformational Analysis In Silico*) a été développé afin d'effectuer un post-traitement de conformations générées automatiquement auparavant. Il s'agit de sélectionner rapidement, parmi toutes ces conformations, les plus représentatives afin de réduire le nombre de conformations à analyser. La méthode utilisée est le clustering hiérarchique qui permet un regroupement progressif des conformations ayant des géométries semblables. Au final, un petit nombre de conformations est sélectionné par Callisto. Des essais ont été réalisés dans le but d'observer l'influence des conformations sur l'application des modèles, en particulier ceux développés précédemment pour la stabilité thermique contenant des descripteurs géométriques ou quantiques (partage sur la valeur de la propriété) .

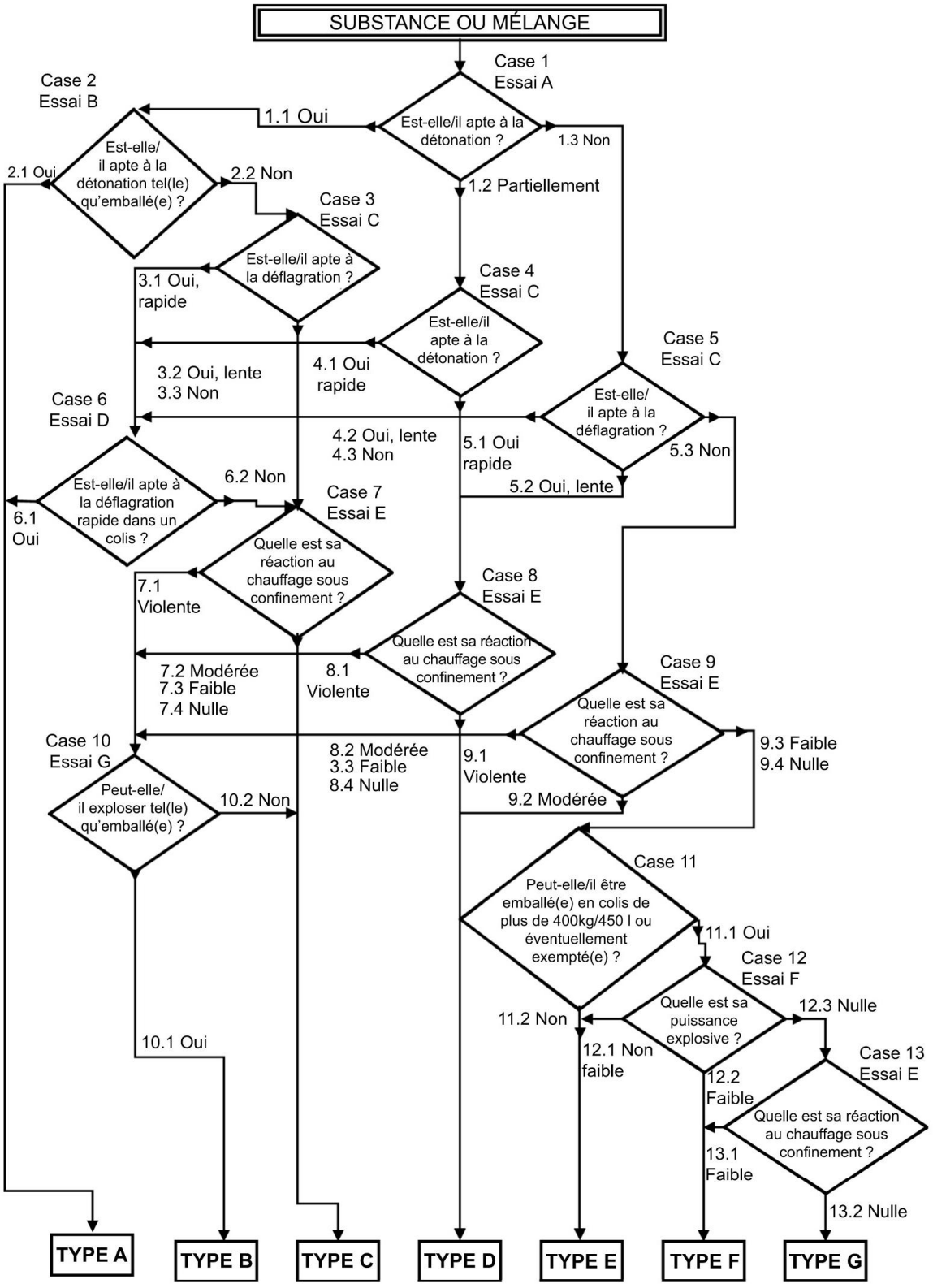
La suite logique de cette étude conformationnelle est l'analyse de l'influence des conformations au niveau du développement des modèles afin de savoir si cette analyse est nécessaire par rapport au gain de performance. Ainsi, des modèles sont à développer soit sur les structures minimales proposées pour chaque molécule par le logiciel générant automatiquement l'ensemble des conformations, soit sur les structures minimales pour chaque molécule obtenue après une optimisation au niveau DFT des conformations proposées par Callisto, soit sur les structures de toutes les conformations sélectionnées pour chaque molécule par Callisto sans et avec pondération. Cette dernière analyse demande beaucoup de temps puisqu'il faut considérer une moyenne de 10 conformations par molécule.

Ces travaux de thèse ouvrent donc de nombreuses perspectives en termes de méthodes de développement (méthode de classement comme les arbres de décisions) et de validation des modèles (Y-scrambling avec sélection des descripteurs et bootstrap), de propriétés (qualitatives comme la sensibilité à l'impact) et de molécules (amines prévues dans le cadre du projet PREDIMOL). D'autre part, l'utilisation des modèles développés dans le cadre de l'enregistrement des substances pour la règlementation REACH (après leur reconnaissance par l'OCDE) sera développée.

ANNEXES

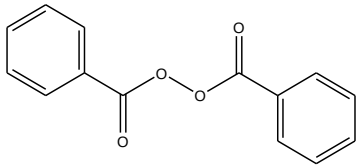
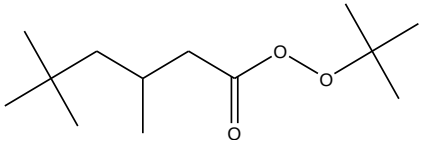
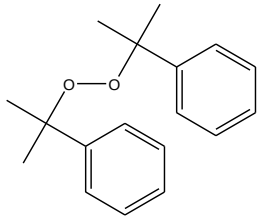
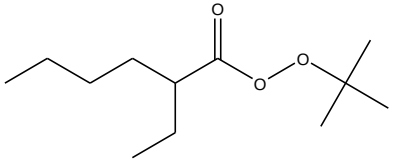
Annexe I

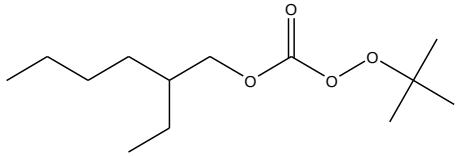
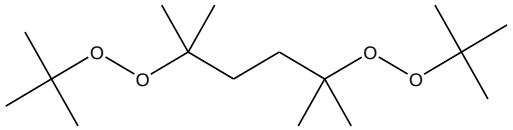
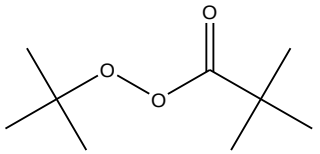
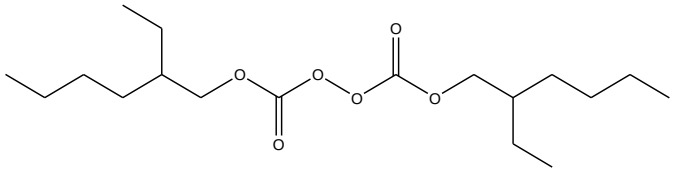
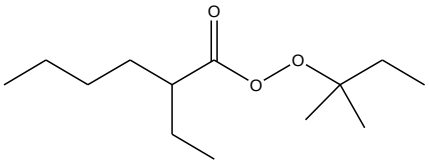
I. LE DIAGRAMME DE DÉCISION POUR LE CLASSEMENT DES MATIÈRES AUTORÉACTIVES ET DES PEROXYDES ORGANIQUES SELON LE MANUEL D'ÉPREUVES ET DE CRITÈRES.

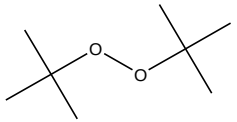
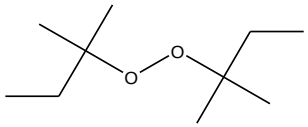
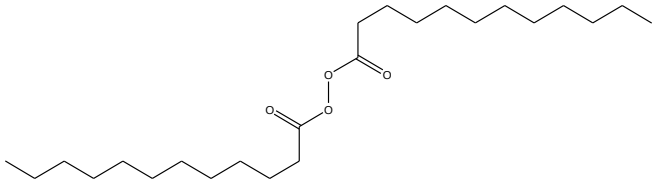
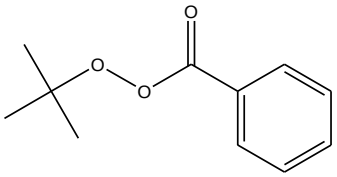
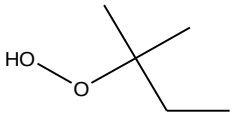


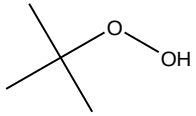
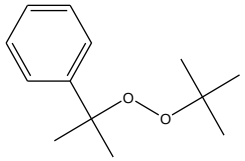
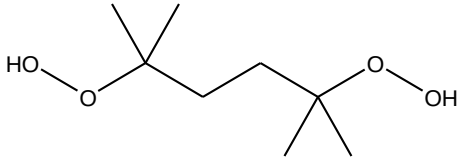
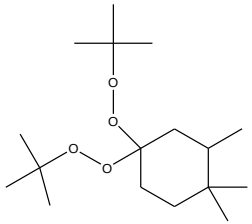
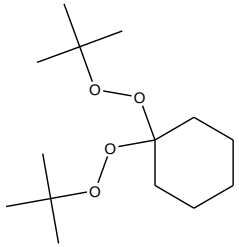
Annexe II

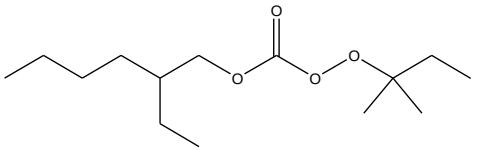
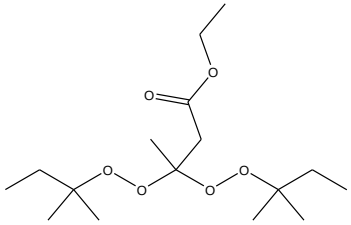
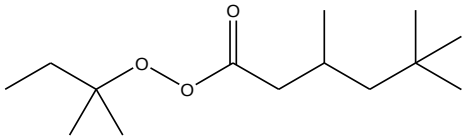
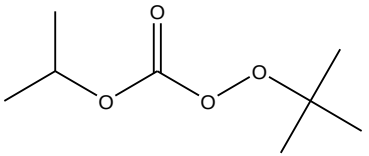
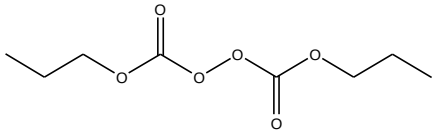
II. BASE DE DONNEES DES PEROXYDES ORGANIQUES DANS LE CADRE DE PREDIMOL

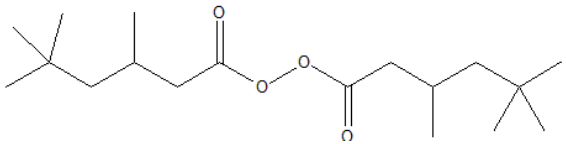
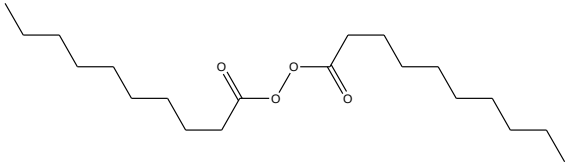
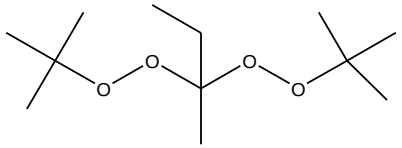
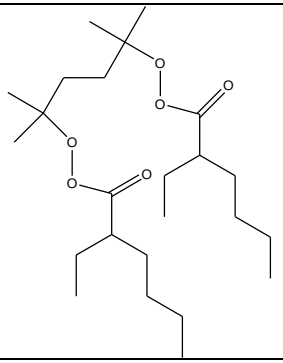
Nom	N° CAS	Structure
dibenzoyl peroxide	94-36-0	
tert-butyl peroxy-3,5,5-trimethylhexanoate	13122-18-4	
dicumyl peroxide	80-43-3	
tert-butyl peroxy-2-ethylhexanoate	3006-82-4	

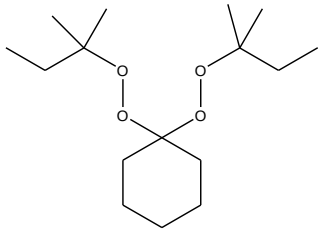
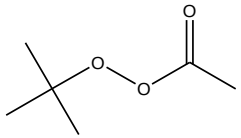
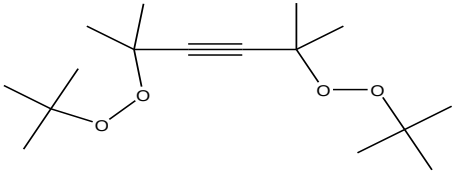
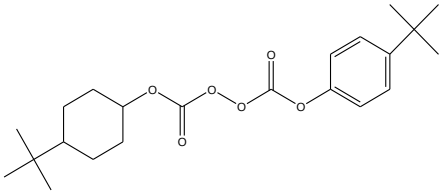
Nom	N° CAS	Structure
tert-butyl peroxy-2-ethylhexylcarbonate	34443-12-4	
2,5-dimethyl-2,5-di(tert-butylperoxy)hexane	78-63-7	
tert-butyl peroxypivalate	927-07-1	
di-(2-ethylhexyl) peroxydicarbonate	16111-62-9	
tert-amyl peroxy-2-ethylhexanoate	686-31-7	

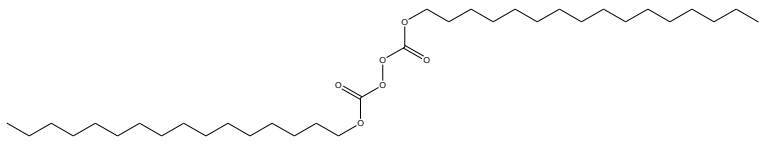
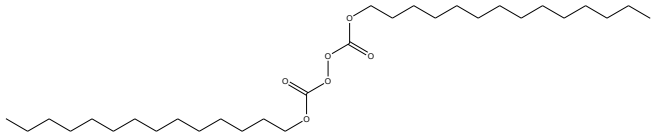
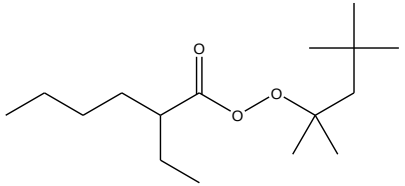
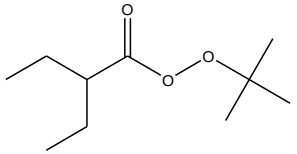
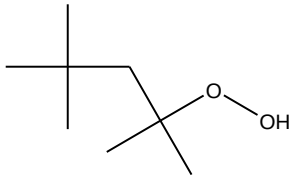
Nom	N° CAS	Structure
di-tert-butyl peroxide	110-05-4	
di-tert-amyl peroxide	10508-09-5	
dilauroyl peroxide	105-74-8	
tert-butyl peroxybenzoate	614-45-9	
tert-amyl hydroperoxide	3425-61-4	

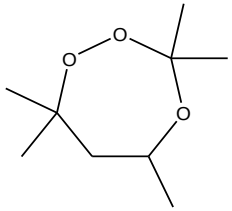
Nom	N° CAS	Structure
tert-butyl hydroperoxide	75-91-2	
tert-butyl cumyl peroxide	3457-61-2	
2,5-dimethyl-2,5-dihydroperoxyhexane	3025-88-5	
1,1-di-(tert-butylperoxy)-3,3,5-trimethylcyclohexane	6731-36-8	
1,1-di-(tert-butyl peroxy) cyclohexane	3006-86-8	

Nom	N° CAS	Structure
tert-amyl peroxy-2-ethylhexyl carbonate	70833-40-8	
ethyl 3,3-di-(tert-amyl peroxy) butyrate	67567-23-1	
tert-amyl peroxy-3,5,5-trimethylhexanoate	68860-54-8	
tert-butyl peroxyisopropylcarbonate	2372-21-6	
di-n-propyl peroxydicarbonate	16066-38-9	

Nom	N° CAS	Structure
di-(3,5,5-trimethylhexanoyl) peroxide	3851-87-4	
didecanoyl peroxide	762-12-9	
2,2-di-(tert-butyl peroxy) butane	2167-23-9	
2,5-dimethyl-2,5-di-(2-ethylhexanoyl peroxy) hexane	13052-09-0	

Nom	N° CAS	Structure
1,1-di-(tert-amyl peroxy)cyclohexane	15667-10-4	
tertio-butyl peracétate	107-71-1	
2,5-di(tert-butylperoxy)-2,5-dimethyl-3-hexyne	1068-27-5	
di-(4-tert-butylcyclohexyl) peroxydicarbonate	15520-11-3	

Nom	N° CAS	Structure
dicetyl peroxydicarbonate	26322-14-5	
dimyristyl peroxydicarbonate	53220-22-7	
1,1,3,3-tetramethylbutyl peroxy-2-ethylhexanoate	22288-43-3	
tert-butyl peroxydiethylacetate	2550-33-6	
1,1,3,3-tetramethylbutyl hydroperoxide	5809-08-5	

Nom	N° CAS	Structure
3,3,5,7,7 pentamethyl 1,2,4 trioxepane	215877-64-8	

Annexe III et IV

III. CALLISTO : CONFORMATIONAL ANALYSIS IN SILICO

1. Installation

Pour installer le programme, suivre les étapes suivantes :

1) Installation des librairies : python et numpy
2) Décompresser l'archive (tar –xvjf callisto.tar.bz2).
Allez dans le dossier "Callisto".
3) Taper les lignes de commandes suivantes :
python setup.py build
python setup.py install

2. Fichier d'entrée

CAllISto (Clustering Analysis In Silico) ne génère pas les différentes conformations possibles pour une molécule. Les données d'entrée nécessaires sont donc la géométrie (les coordonnées cartésiennes des différents atomes) des conformères possibles et leur énergie respective. Ces informations sont entrées sous la forme d'un fichier *structure data file* (.sdf voir annexe IV) contenant toutes ces structures.

3. Utilisation et liste des options

Allez dans le dossier contenant le fichier d'entrée (sdf)

Tapez la commande : *callisto file_name.sdf*

La liste des options est disponible avec l'option --h :

--debug	affiche les informations de débogage
--divalg	exécute l'algorithme de clustering divisif
--aggalg=method	exécute l'algorithme de clustering agglomératif (par défaut)
with method=average	utilisation du average linkage (par défaut)
single	utilisation du single linkage
complete	utilisation du complete linkage
ward	utilisation du ward linkage
weighted	utilisation du weighted linkage
flexible	utilisation du flexible linkage
--tclust=value	Choix du seuil RMSD
with value=float	(par défaut float=1.5)
--nclust=value	Choix du nombre de cluster finaux
with value=int	
--tboltz=value	Choix du pourcentage seuil pour l'analyse de population de Boltzmann
with value=float	(par défaut =0.1)
--help or -h	Affiche l'aide

4. Sorties

Le programme développé renvoie plusieurs informations, à la fois dans le terminal mais aussi par la création de fichiers. Certains fichiers sont utilisables pour ré-exécuter le programme pour une même molécule avec des options différentes pour un temps de calcul plus faible. D'autres fichiers sont créés afin de poursuivre l'analyse conformationnelle avec d'autres logiciels après l'obtention d'une sélection réduite de conformations ou encore pour une illustration graphique des résultats.

a) Sortie standard

- Liste des clusters générés et des conformations regroupées dans chacun d'eux.

```
List of clusters after clustering selection (H=1.500):
 0: 11 18 24 31 38 40 41 47 50 52 56 57 76 83 85 86 98 101 106 108 109 116
122 125 131 149 152 158 167 175 183
 1: 1 5 8 13 25 28 35 37 42 44 54 61 67 74 81 82 88 105 107 132 139 142
148 151 159 166 168 169 171 173 174 180 190
 2: 0 4 6 7 9 12 21 23 46 60 64 65 70 71 77 79 80 84 91 94 99 100 111 112
114 115 120 121 134 136 144 145 155 160 164 181 186 189
 3: 3 14 15 17 19 22 32 34 39 43 69 72 75 78 90 92 95 97 119 123 124 126
129 140 141 147 153 157 161 162 163 182 184 185 187 191 204
 4: 2 10 16 20 26 27 29 30 33 36 45 48 49 51 53 55 58 59 62 63 66 68 73 89
93 96 102 103 110 127 128 130 133 135 137 138 143 146 150 154 156 176 177
178 179 188
 5: 193 195 202 206 209 210
 6: 87 104 113 117 118 165 170 172
 7: 192 194 196 205 208 213 215 216
 8: 197 198 199 200 201 203 207 211 212 214
```

- Le profil silhouette qui mesure la précision du clustering est calculé : Le profil silhouette S moyenne est les différents s(i) de chaque cluster avec le nombre d'objets qu'il contient.

```
Silhouette profile: 0.15
 0: 31 | 0.16
 1: 33 | 0.14
 2: 38 | 0.15
 3: 37 | 0.13
 4: 46 | 0.13
 5: 6 | 0.11
 6: 8 | 0.20
 7: 8 | 0.21
 8: 10 | 0.22
```

- La liste des conformations représentatives de chaque cluster

```
0 1 2 3 11 87 192 193 197
```

- La liste des conformations représentatives de chaque cluster après la suppression des conformations trop hautes en énergie selon l'analyse de population de Boltzmann

```
0 1 2 3 11 87
```

Le calcul de la matrice des RMSD est l'étape la plus longue et peut être réutilisée dans le cas de tests (types de clustering, valeurs seuils). Cette matrice est donc conservée dans un fichier csv qui est créé dans le répertoire dans lequel se trouve le fichier d'entrée.

Tous les autres fichiers de sortie sont disponibles dans le répertoire nouvellement créé.

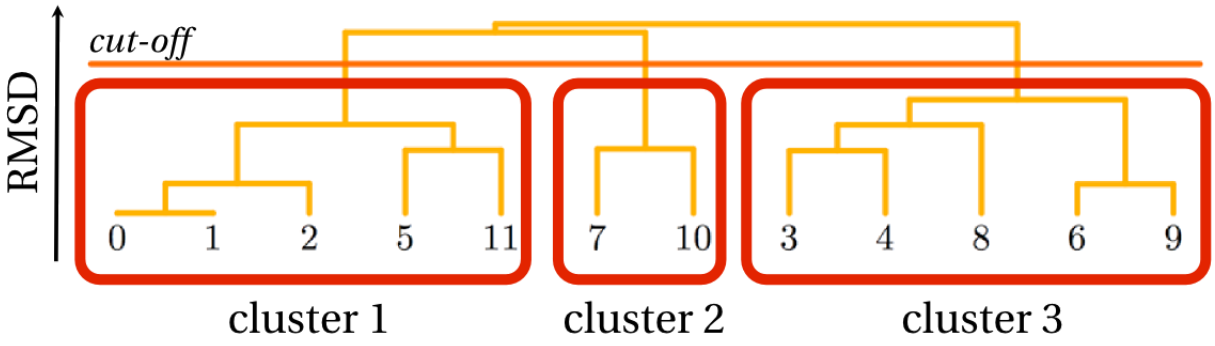
b) Fichier de sorties

Fichier d'entrée pour Gaussian

Les géométries des conformations sélectionnées après le clustering sont sauvées dans des fichiers d'entrée Gaussian (.gjf).

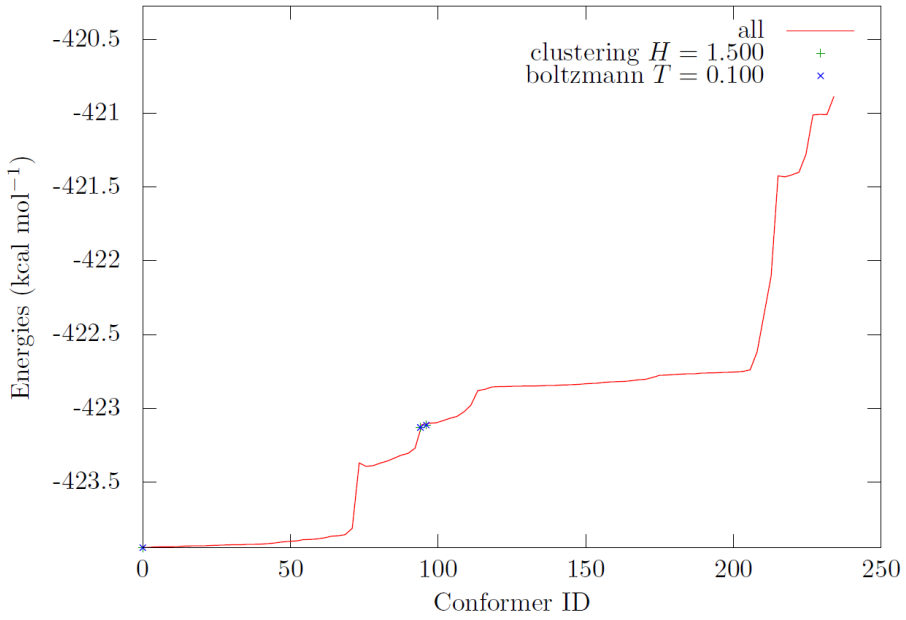
Dendrogramme

Un fichier latex (.tex) représentant le dendrogramme est créé. Pour le transformer en image pdf, latex doit être installé.



Courbe énergétique

Un fichier gnuplot (.gnu) contenant les coordonnées énergétiques des différentes conformations est aussi créé. Afin d'obtenir l'image, le programme gnuplot doit être installé. Les énergies de toutes les conformations de départ sont représentées sous forme de courbe, l'énergie des conformations sélectionnées après clustering, avant puis après une analyse de population de Boltzmann est représentée sous forme de points sur cette courbe.



IV. FICHIER STRUCTURE DATA FILE (.SDF)

Un fichier SDF est une série de fichiers de type MOL qui sont séparés par “$$$$”.

Fichier MOL

Description d’un fichier MOL :

Exemple :

```
" Molecule 1
MDL .mol file

 16 15  0     0               999 V2000
    1.5921   -0.6522   -0.1351 C   0  0
    0.3876    0.1118   -0.1085 O   0  0
   -0.7292   -0.8182   -0.1030 O   0  0
    2.7498    0.3626   -0.1422 C   0  0
    1.6880   -1.5307    1.1254 C   0  0
    1.6493   -1.5044   -1.4160 C   0  0
   -1.4499   -0.2066   -0.1797 H   0  0
    2.7035    1.0321   -1.0308 H   0  0
    2.7323    1.0120    0.7621 H   0  0
    3.7393   -0.1466   -0.1637 H   0  0
    1.6067   -0.9237    2.0553 H   0  0
    0.8891   -2.3048    1.1625 H   0  0
    2.6615   -2.0687    1.1671 H   0  0
    0.8521   -2.2806   -1.4426 H   0  0
    1.5358   -0.8785   -2.3299 H   0  0
    2.6226   -2.0377   -1.5004 H   0  0
  2  1  1
  2  3  1
  1  4  1
  1  5  1
  6  1  1
  3  7  1
  8  4  1
  4  9  1
  4 10  1
  5 11  1
  5 12  1
  5 13  1
  6 14  1
 15  6  1
  6 16  1
M  END "
```

Nombre d’atomes
Nombre de liaisons

Description des atomes:
coordonnées (x y z)
type (O, C, H, N …)
charge

Description des liaisons:
Liste des atomes liés
Type de liaison

1 : liaison simple
2 : liaison double

Fichier SDF

Description d'un fichier SDF :

```
" Molecule_1

 16 15  0     0               999 V2000
    1.5921   -0.6522   -0.1351 C   0  0
    0.3876    0.1118   -0.1085 O   0  0
   -0.7292   -0.8182   -0.1030 O   0  0
    2.7498    0.3626   -0.1422 C   0  0
    1.6880   -1.5307    1.1254 C   0  0
    1.6493   -1.5044   -1.4160 C   0  0
   -1.4499   -0.2066   -0.1797 H   0  0
    2.7035    1.0321   -1.0308 H   0  0
    2.7323    1.0120    0.7621 H   0  0
    3.7393   -0.1466   -0.1637 H   0  0
    1.6067   -0.9237    2.0553 H   0  0
    0.8891   -2.3048    1.1625 H   0  0
    2.6615   -2.0687    1.1671 H   0  0
    0.8521   -2.2806   -1.4426 H   0  0
    1.5358   -0.8785   -2.3299 H   0  0
    2.6226   -2.0377   -1.5004 H   0  0
  2  1  1
  3  2  1
  4  1  1
  1  5  1
  6  1  1
  7  3  1
  8  4  1
  9  4  1
  4 10  1
 11  5  1
  5 12  1
  5 13  1
  6 14  1
 15  6  1
  6 16  1
M  END
> <Potential Energy [kcal/mol]>
12.070737838745117

> <CONFLEX_Flag>
3.0
```

1ère conformation

Fichier MOL de la conformation n°1

```
$$$$
Molecule_2
 […] "
```

Fichier MOL de la conformation n°2

TABLES

TABLE DES FIGURES

TABLE DES TABLEAUX