République Algérienne Démocratique et Populaire

Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

Université Yahia FARES de Médéa



Faculté des Sciences et de la Technologie

Département Génie des Procédés et Environnement

Mémoire pour l'obtention du diplôme de Master en Génie des Procédés

Option : Génie des Procédés Pharmaceutique

Thème

Détermination et réduction du nombre de descripteurs structuraux des polymères dans la prédiction de leur perméabilité aux gaz

Réalisée par :

HASSEIN-BEY Amel Hind (M^{me} OUDJOUEDJ)

Examiné par :

-Président de Jury : Dr Anoun Mohamed

-Examinateur 1 : Pr Moulai-Mostafa Naji

-Examinateur 2 : M^{me} Kerrache Sonia

-Promoteur : Pr Krea Mohamed

-Copromoteur : M^{me} Hasnaoui Hanaa

Année universitaire

2013/2014

Sommaire

INTRODUCTION GENERALE 1

I. CHAPITRE1 4

I.1. DEFINITIONS 4

I.1.1. Les Polymères vitreux de type Polyimides 5

I.2. ARCHITECTURE MOLECULAIRE ET PROPRIETES DES POLYMERES 6

I.2.1. Structure microscopique 6

I.2.2. Structure des Polyimides 7

I.2.3. Structure macroscopique 8

I.2.4. Propriétés structurales 9

I.2.5. Propriétés structurales des Polyimides 10

I.3. SYNTHESE DES POLYMERES 12

I.3.1. Réaction de polymérisation 12

I.3.2. Techniques de polymérisation 14

I.4. MISE EN FORME DES POLYMERES 15

I.4.1. Polyimides thermoplastiques 16

I.5. PROCEDES DE SEPARATION MEMBRANAIRES 16

I.5.1. Généralités 16

I.5.2. Avantages et inconvénients de ces procédés 17

I.5.3. Principe de la perméation gazeuse 19

I.5.4. Méthode de mesure de la perméabilité dans les polymères 20

I.6. CONCLUSION 20

II. CHAPITRE 2 22

II.1. MODELISATION MOLECULAIRE DES STRUCTURES TRIMERES 23

II.1.1. Les méthodes quantiques 23

II.1.2. La méthode de Hückel 24

II.2. DESCRIPTEURS MOLECULAIRES 24

II.2.1. Types de descripteurs moléculaires 25

II.3. L'ANALYSE DE DONNEES OU « DATA MINING » 29

II.4. BUT DE L'ANALYSE DE DONNEES 30

II.5. ANALYSE EN COMPOSANTES PRINCIPALES 31

II.5.1. Présentation et description de la méthode factorielle : Analyse en composantes principales (ACP) 31

II.5.2. Interprétation des résultats d'une ACP 32

II.5.3. Limites 34

II.6. ANALYSE DES MOINDRES CARRES PARTIELS 34

II.6.1. Présentation 34

II.6.2. Intérêt de la Régression PLS 36

II.6.3. La méthode 36

II.6.4. Choix du nombre de composantes 36

II.7. LES APPROCHES GENETIQUES ET SELECTION 38

II.7.1. Algorithmes génétiques 38

II.8. APPLICATION DES RESEAUX DES NEURONES : 40

II.9. LOGICIELS DE CALCULS 40

II.10. CONCLUSION 41

III. CHAPITRE 3 43

III.1. DEMARCHE A SUIVRE 45
III.2. LES LOGICIELS ET MATERIELS UTILISES : 45
III.3. COLLECTE DES DONNEES 46
III.4. DESSIN DES MOLECULES SUR CHEMDRAW 46
III.5. MODELISATION MOLECULAIRE DES TRIMERES PAR HYPERCHEM 48
III.6. CALCUL DES DESCRIPTEURS PAR DRAGON 50
III.7. COLLECTION DES DONNEES DANS UN FICHIER EXCEL 51
III.8. REDUCTION DU NOMBRE DE VARIABLES 52
III.9. APPLICATION DE L'ANALYSE EN COMPOSANTES PRINCIPALES 55
III.10. APPLICATION DE LA METHODE DES MOINDRES CARREES PARTIELLES PLS 57
III.10.1. Selon la méthode de PLS avec validation croisée 58
III.10.2. Selon la méthode de PLS sans validation croisée 60
III.10.3. Selon la méthode de PLS d'élimination de la variable non informative (PLSUVE) 60
III.10.4. Selon la méthode de PLS analyse discriminante 61
III.11. APPLICATION DE LA METHODE DES ALGORITHME GENETIQUES 61
III.12. CONCLUSION 63

IV. APPLICATION DES DESCRIPTEURS RETENUS A LA PREDICTION DE LA PERMEABILITE 64

V. CONCLUSION GENERALE 70

VI. REFERENCES BIBLIOGRAPHIQUES 71

Liste des tableaux

Tableau I-1: classification des polyimides selon les réactifs employés pour leur synthèse.....9
Tableau I-2 : exemple commercial des polyimides et des polyetherimides 16
Tableau II-1:Nombre de descripteurs calculés par chaque famille des 22 blocs selon DRAGON (4) ..26
Tableau II-2 : Répartitions des descripteurs selon le type de structure par DRAGON (4).....27
Tableau IV-1 : résultats des ERAM et R du modèle neuronale pour chaque méthode 64

Liste des figures

Figure I-1 : groupement imide....6
Figure I-2: Les différentes structures microscopiques d'un polymère....7
Figure I-3 : structure des polymères polyimides....7
Figure I-4 : Les différentes structures macroscopiques d'un polymère....9
Figure I-5 : formation des polyimides au niveau macroscopique....11
Figure I-6 : représentation du transfert de charge formant les polyimides....11
Figure I-7 : Obtention d'un polyimide flexible....12
Figure I-8 : Exemple d'une réaction de polyaddition....12
Figure I-9**:** Exemple d'une réaction de polymérisation radicalaire....13
Figure I-10**:** Exemple d'une réaction de polymérisation cationique....13
Figure I-11**:**Exemple d'une réaction de polymérisation anionique....13
Figure I-12 : Exemple d'une réaction de polymérisation par coordination....13
Figure I-13 : Représentation des trois modes de séparation membranaire tangentielle....17
Figure I-14 : schéma du principe de la perméation membranaire....19
Figure II-1: Architecture générale d'un algorithme génétique....39
Figure II-2 : Comparaison schématique entre un neurone biologique et neurone artificiel....40
Figure III-1 : Organigramme des étapes suivies pour la préparation de la base de données....45
Figure III-2 : étapes de représentation des structures par ChemDraw....47
Figure III-3 : HyperChem version 8.0.7 pour la modélisation moléculaire des trimères....48
Figure III-4 : Etapes des pré-optimisations et d'optimisation des molécules respectivement....49
Figure III-5 : L'interface du logiciel DRAGON TALETE version 5.5....50
Figure III-6 : étapes de calcul des descripteurs par DRAGON TALETE....51
Figure III-7 : La base de données des descripteurs finale....52
Figure III-8 : représentation des valeurs propres en fonctions du nombre de composantes principales 55
Figure III-9 : Représentation des variables en fonction des premiers principaux facteurs....56
Figure III-10 : illustration des méthodes de PLS : Filtrantes, Enveloppantes et intégrés....57
Figure III-11 : Différentes variantes de PLS pour la sélection des variables....58
Figure III-12 : effet du nombre de composantes principales sur la variance exprimé des Y....58
Figure III-13 : choix des variables selon le coefficient de régression....59
Figure III-14 : choix des variables selon le poids des descripteurs....59
Figure III-15 : représentation du RMSEP en fonction du nombre de variables sélectionnées....62
Figure III-16 : représentation des fréquences de sélection après 100 itérations....62
Figure IV-1 : représentation de l'écart entre le modèle prédit et l'expérimental de la PLSUVE....65
Figure IV-2 : représentation de l'écart entre le modèle prédit et l'expérimental de la PLS sans CV....66
Figure IV-3 : représentation de l'écart entre le modèle prédit et l'expérimental de la PLSdisc....66
Figure IV-4 : représentation de l'écart entre le modèle prédit et l'expérimental de la PLSCV....67
Figure IV-5 : représentation de l'écart entre le modèle prédit et l'expérimental de la PLSAG....67
Figure IV-6 : représentation de l'écart entre le modèle prédit et l'expérimental de la PCA....68

Introduction générale

Introduction générale

Depuis le début des années 1950, la production des polymères est en continuelle ascension à cause d'un usage accru et très varié de ces produits dans le monde moderne; dont le secteur le plus important étant l'emballage dans le cas des aliments, et le conditionnement dans le cas des produits pharmaceutiques.

Ces deux domaines sensiblement importants pour l'Homme l'ont poussé à étudier scrupuleusement d'éventuelles interactions existantes entre le contenant (emballage ou conditionnement) et le contenu (aliments et médicaments) ; d'où la connaissance des phénomènes physique et chimique de l'évolution de ce dernier avec l'extérieur à travers cette membrane qui le protège, a abouti à la détermination de : la migration, de la sorption et de la perméation. Tous ces trois paramètres ont la même origine qui est la différence de concentration entre deux milieux séparés en un composé.

Pas mal de théories élaborées pour prédire ces propriétés dans le cas de la séparation par des membranes polymériques nouvelles n'ayant que leur structure ; -sans avoir recours à l'expérimentation souvent couteuse, dont les plus concluants sont les méthodes de contribution des groupes structuraux polymériques.

Cependant, l'accumulation de tant de données sur les polymères et leurs propriétés a nécessité leur numérisation pour une meilleure exploitation des bases de données incomplètes, et là aussi des recherches ont été faites comme par exemple l'étude de la perméabilité des polymères vitreux et élastomères aux gaz et vapeurs condensables selon la méthode de Yampolskii [1]

Dans la quête d'un modèle de prédiction universel pour tous les polymères existants et inexistants encore -à fin d'orienter leur synthèse pour une séparation ciblée- une approche qualifiée de fiable et puissante basée sur les réseaux de neurones est entreprise au niveau du laboratoire LME.

Introduction générale

Cette modélisation neuronale a permis de prédire la perméabilité de l'ensemble des polymères vitreux et élastomères aux gaz d'O_2, N_2, CO_2 et CH_4 avec des ERAM et coefficient R très intéressants. Cependant, la décomposition non systématique des polymères en groupe de Yampolski reste un inconvénient majeur de cette approche.

Il serait très intéressant de systématiser la décomposition des polymères et d'améliorer ce modèle neuronal par un choix plus universel des descripteurs en tenant compte de l'organisation spatiale de ceux-ci.

Ce travail s'inscrit dans ce contexte et constitue une suite logique des travaux antérieurs réalisés dans le laboratoire Matériaux et Environnement LME sur l'exploitation des données expérimentales pour la collection et la validation des modèles publiés de prédiction de perméabilité et de solubilité pour un large éventail de matériaux filmogènes utilisés soit dans les procédés de séparation ou comme emballage des produits pharmaceutiques ou agroalimentaires. Il a pour objectif de réduire le nombre de descripteurs structuraux des polymères, candidats à une séparation gazeuse par les membranes, par les trois techniques PCA, PLS et AG et comparer leurs résultats.

Le présent travail s'organise en trois chapitres selon le plan suivant :

Le premier chapitre donne des généralités sur les polymères étudiés et les procédés de séparations membranaires.

Le second chapitre introduit les trois méthodes linéaires proposées pour la réduction des descripteurs et qui sont : L'analyse en composantes principales (ACP) ; L'analyse des moindres carrés partiels (PLS) ; et les algorithmes génétiques(AG).

Le dernier chapitre est divisé en deux parties : la première détaille la préparation de la base de données (depuis la représentation des structures jusqu'aux calculs des descripteurs) et les logiciels employés ainsi que les résultats obtenus, analysés et expliqués dans la seconde partie.

La conclusion tirée à partir de ce travail et les perspectives qui peuvent être envisagées dans le futur sont présentées à la fin du mémoire.

Chapitre 1 :

Polymères & procédés membranaires

I. Chapitre1

Au début des années 1920, les chimistes doutaient de l'existence de molécules ayant des masses moléculaires supérieures à quelques milliers ; mais ce point de vue très limité fut vite contesté par le chimiste allemand Hermann Staudinger, qui proposa qu'elles fussent composées de macromolécules de 10000 atomes ou plus ; Flory énoncera plus tard les principes fondamentaux sur les polymères .

Le développement industriel consécutif de la science macromoléculaire a été accéléré ensuite par la Seconde Guerre mondiale. Les Etats-Unis ont été privés lors de leur entrée en guerre de leur approvisionnement en caoutchouc naturel en provenance d'Asie du Sud-est. Ils ont alors lancé un immense programme de recherche visant à trouver des substituts de synthèse. Dans ce chapitre nous présenterons notre famille des polyimides sur laquelle nous avons travaillé.

I.1. Définitions

Un *polymère* est un ensemble de macromolécules issues de l'assemblage covalent de motifs identiques appelés *monomères*. Il peut être inorganique ou organique, c'est à dire réalisé à partir de monomères composés d'atomes utilisés en chimie organique : le C, H, O, et N principalement ainsi que d'autres éléments comme les halogènes (F, Cl, Br, I) ou le souffre, le phosphore...Il existe d'autres polymères où l'atome de carbone est remplacé par Si (silicones).

Un *monomère* est la molécule de base qui réagit par des réactions dites de *polymérisation*. Il peut être un composé organique insaturé, ou un composé organique comportant une ou des fonctions réactives…

Un *macromère* (ou macro-monomère) est une espèce polymérisable de grande taille.

Un *ionomère* désigne des polymères porteurs de groupes ioniques ou ionisables.

Un *homopolymère* est un polymère préparé à partir d'un seul motif monomère.

Un *copolymère* est un polymère obtenu à partir de deux monomères ou plus. Cela donne une très grande variété de structures:

-*Le copolymère alterné*: alternance entre les séquences A et B (Figure I-2(d))

-*Le copolymère à blocs*: comporte des séquences plus ou moins longues de A et B(Figure I-2(b))

-*Le copolymère statistique*: les séquences A et B sont « au hasard » (aléatoires)Figure I-2(c))

- *Le copolymère greffé*: fixation latérale de blocs B sur une chaîne principale formée de chaînons A ou l'inverse (Figure I-2(e)).

I.1.1. Les Polymères vitreux de type Polyimides

Les polyimides sont un groupe très intéressant de polymères incroyablement résistants mécaniquement, chimiquement et également thermiquement. Leur résistance est si grande que ces matériaux remplacent souvent le verre et des métaux comme l'acier dans des applications industrielles très exigeantes.

Ils sont utilisés partout dans notre quotidien pour les renforts et les châssis dans certaines voitures, ainsi que pour certaines parties situées sous le capot, car ils supportent la chaleur intense et les lubrifiants corrosifs, carburants, liquides de refroidissement que nécessite le bon fonctionnement d'une voiture ;en cuisine dans la construction de nombreux appareils électroménagers, ainsi que pour les plats de cuisson pour micro-onde et pour les emballages de nourriture à cause de leur stabilité thermique, leur résistance aux huiles, graisses, corps gras et leur transparence aux radiations micro-onde. Ils peuvent aussi être utilisés pour les plaques des circuits intégrés, l'isolation, les fibres de vêtement de protection, les composites et les colles.

Un polyimide est un polymère qui contient un groupe imide dont la structure générale est la suivante (Figure I-1) :

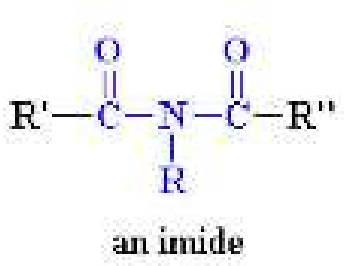


Figure I-1 : groupement imide

I.2. Architecture moléculaire et propriétés des polymères

I.2.1. Structure microscopique

On distingue trois types de chaines

I.2.1.1. Les polymères linéaires

Les motifs s'enchaînent de manière unidimensionnelle c'est-à-dire que la chaîne ne comporte pas d'autres ramifications que les substituants latéraux déjà présents sur le monomère. Ce sont en général des homopolymères constitués par la répétition dans la chaîne, du même motif (monomère) récurent (Figure I-2(a)).Ceci ne signifie pas que les groupes unitaires de la chaîne soient alignés. Ils peuvent être:

- en zig-zag

- en hélice simple ou double

- en pelote

I.2.1.2. Les polymères ramifiés

Ils possèdent des ramifications plus ou moins importantes constituées par la répétition. Les séquences de monomères identiques donnent soit des homopolymères branchés (Figure I-2(c) : les B sont des A), ou bien des homopolymères en étoile (Figure I-2(f)) et les séquences de monomères différents donnent des copolymères greffés (Figure I-2(e)).

I.2.1.3. Les polymères réticulés

Appelés aussi polymères tridimensionnels ; ils comportent des pontages en nombre variable entre les chaines moléculaires. Dans le cas d'un copolymère, les motifs s'arrangent au hasard (Figure I-2(g)).

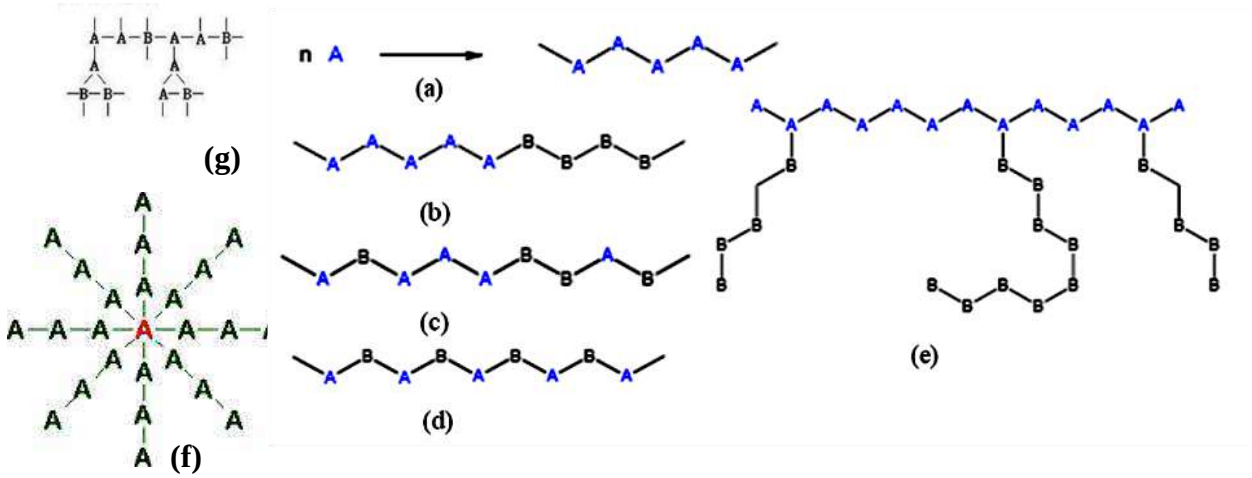


Figure I-2: Les différentes structures microscopiques d'un polymère

I.2.2. Structure des Polyimides

Les polyimides prennent en général une forme parmi deux possibles. La première de ces formes est une structure *linéaire* dans laquelle les atomes du groupe imide font partie de la chaîne linéaire. La seconde structure est une structure *hétérocyclique* dans laquelle le groupe imide est une partie d'une unité cyclique dans la chaîne du polymère (Figure I-3).

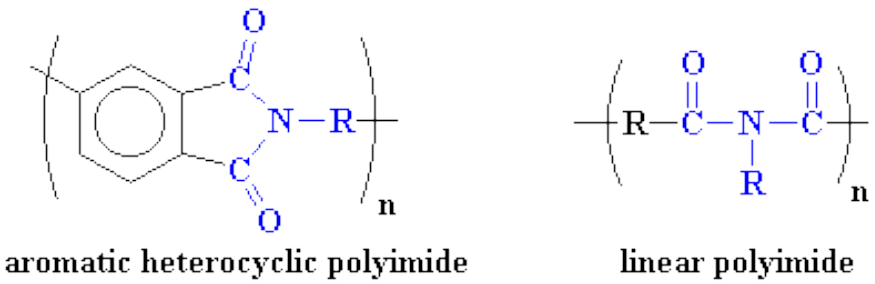


Figure I-3 : structure des polymères polyimides

Les polyimides aromatiques hétérocycliques, comme celui de gauche, représentent la plupart des polyimides commerciaux, comme l'Ultem de G.E. et le Kapton de Dupont, et sont par conséquent les polyimides avec lesquels nous sommes le plus en contact. Ces polymères ont des propriétés mécaniques et thermiques si incroyables qu'ils sont utilisés à la place du verre et des métaux dans beaucoup d'applications hautes performances en électronique, automobile

et même en aérospatiale. Ces propriétés sont dues aux forces intermoléculaires fortes entre les chaînes du polymère.

I.2.3. Structure macroscopique

En général, les polymères formés à partir de chaînes linéaires non réticulées et flexibles sont souples (à certaines températures) tandis que les polymères très réticulés, formant un réseau tridimensionnel sont plus rigides. Les premiers donnent lieu à des polymères thermoplastiques, les seconds à des polymères thermodurcissables.

I.2.3.1. Les thermoplastiques

Les polymères thermoplastiques sont essentiellement des polymères linéaires ou branchés donc peu réticulés et les liaisons intermoléculaires sont de type Van der Waals. Ils se ramollissent lorsqu'on les chauffe au-dessus d'une certaine température, mais ils redeviennent solides en dessous. Cette matière conserve de façon réversible sa thermoplasticité initiale ce qui permet leur mise en forme « moulage » (Figure I-4(a)), exemple : PE, PVC, PP, PS....

I.2.3.2. Les thermodurcissables

Ils deviennent durs à la chaleur de manière irréversible et l'élévation de température provoque plutôt leur dégradation que leur fusion. Les liaisons covalentes se développent dans les trois dimensions, ce qui les rend des polymères tridimensionnels (Figure I-4(d)), exemple: polyesters, résines glycérophtaliques, phénoplastes, etc….)

I.2.3.3. Les élastomères

Ce sont des matériaux amorphes[1], mais avec quelques pontages entre les chaînes macromoléculaires linéaires, ces liaisons sont assurées par des atomes C, S ou O. ces pontages assurent la mémoire de l'état initial. Selon le type de réticulation, les élastomères sont classés en deux familles :

-**Elastomères thermodurcissables** : faible réticulation avec des liaisons covalentes (Figure I-4(c));

[1]*Structure amorphe* : les chaines macromoléculaires sont disposées dans la matière d'une manière totalement anarchique et désordonnée

-Elastomères thermoplastiques (TPE) : faible réticulation avec des liaisons non covalentes comme les liaisons hydrogène (Figure I-4(b)).

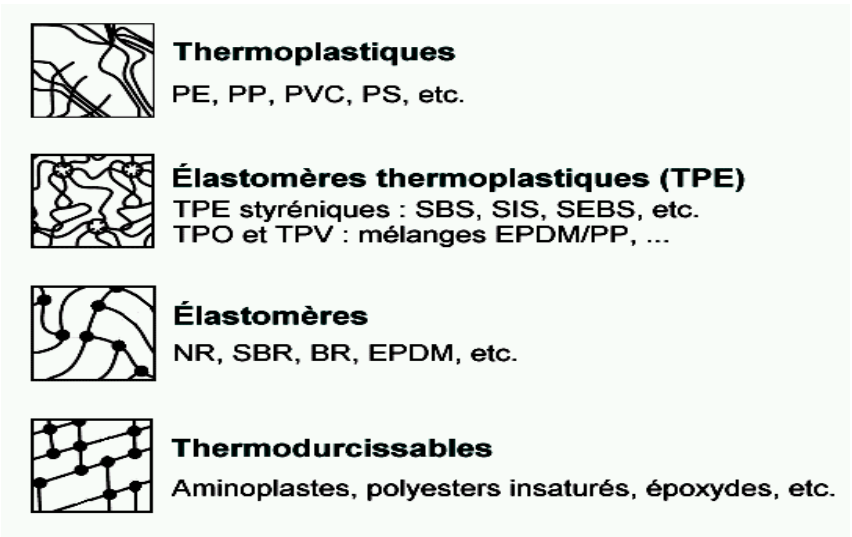


Figure I-4 : Les différentes structures macroscopiques d'un polymère

Selon les réactifs utilisés pour leur synthèse, les polyimides peuvent être (Tableau I-1):

Tableau I-1: classification des polyimides selon les réactifs employés pour leur synthèse

Famille de polyimide	Type de polyimide	Réactifs
Polyimides non réticulés	Polyimides linéaires (PI)	Deux monomères
	Polyimides thermoplastiques (TPI)	Un ou deux monomères et un pré-polymère
Polyimides réticulés (thermodurcissables)	Polybismaléimides (BMI)	Un monomère et un agent réticulant
	Polyimides PMR (Polymerization of Monomeric Reactants)	Deux monomères et un agent réticulant

I.2.4. Propriétés structurales

Lorsque le polymère est branché, les ramifications empêchent une trop grande approche des molécules et les liaisons intermoléculaires restent faibles. Cela se traduit par une structure totalement désordonnée, analogue à celle d'un verre. Cette structure est dite *amorphe*(ou *vitreuse*). Le polymère est alors dur et cassant.

Par contre, pour les polymères parfaitement linéaires, les molécules peuvent, s'aligner, se rapprocher et établir des liaisons assez fortes, appelées *liaisons pontales*. Cette cohésion entre les chaînes entraîne l'existence de zones du solide où l'organisation rappelle celle des cristaux.

Ces zones sont appelées *cristallites.* Leurs dimensions latérales sont de l'ordre du micromètre. Donc cette structure est dite *semi-cristalline*, car l'organisation ne s'étend pas à tout le solide ; il existe des zones amorphes où les chaînes sont disposées au hasard sans liaisons pontales entre elles.

Les propriétés physiques dépendent essentiellement du taux de cristallinité 2. Plus un polymère est ramifié, plus le taux de cristallinité est faible.

Les propriétés thermiques en dépendent aussi : les polymères amorphes passent de manière progressive de l'état solide à l'état fluide sans palier de changement d'état. La température au-delà de laquelle il est possible de réaliser les opérations de moulage est appelée *température de transition vitreuse* T_g. Cependant les polymères semi-cristallins présentent un palier de changement d'état à une température T_f appelée *température de fusion* qui correspond en fait à la désorganisation des cristallites.

La densité qui est une propriété mécanique augmente avec le taux de cristallinité, car elle augmente lorsque les chaînes sont courtes et rapprochées; par contre elle diminue lorsque des ramifications rendent les molécules encombrantes. Donc une forte densité se traduit par une bonne rigidité et une bonne résistance à la traction par exemple le Kevlar ; et une faible densité se traduit par une bonne étirabilité.

I.2.5. Propriétés structurales des Polyimides

Un polymère qui contient un transfert de charge complexe est constitué de deux différents types de monomères, un donneur et un accepteur. Le donneur est plein d'électrons à cause de ses groupes azote alors il prête quelques-uns de ses électrons à l'accepteur, et cela les lient fermement ensemble (Figure I-5).

2 *Le taux de cristallinité* X_c en % d'un polymère : est le rapport de la masse de phase cristalline sur la masse totale du système

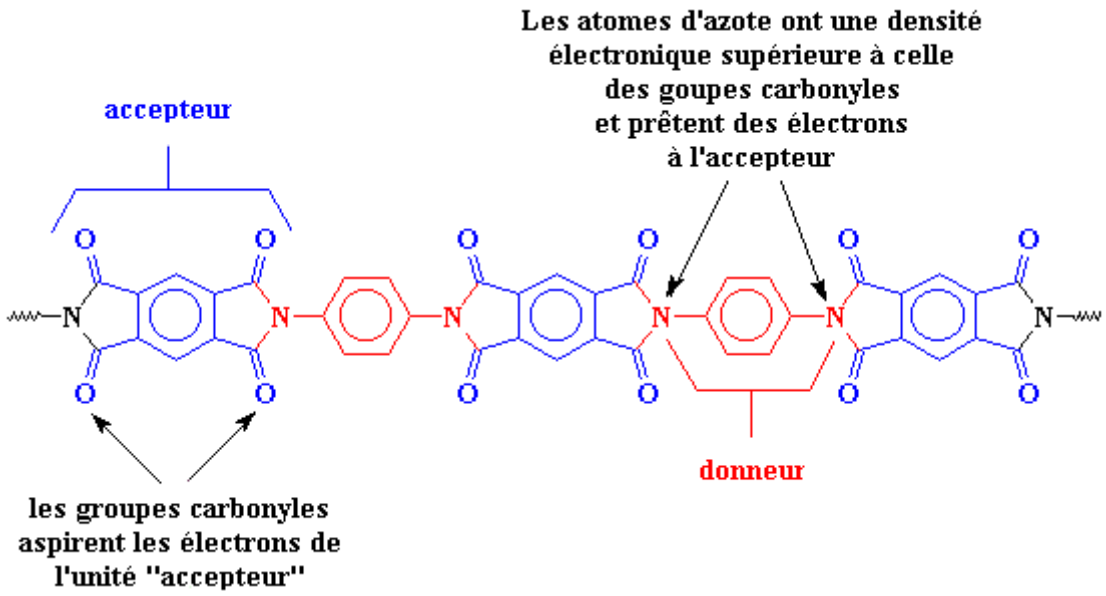


Figure I-5 : formation des polyimides au niveau macroscopique

Le transfert de charge complexe fonctionne non seulement entre les unités adjacentes sur une chaîne, mais aussi entre les chaînes. Les chaînes vont s'empiler les unes sur les autres comme des bandes de papier, les donneurs et les accepteurs étant appariés.

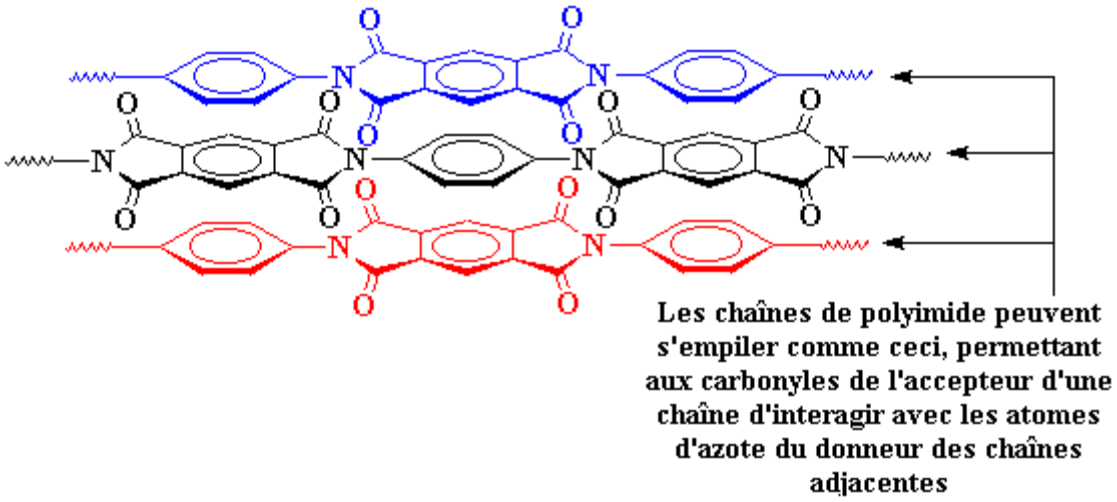


Figure I-6 : représentation du transfert de charge formant les polyimides

Ce transfert de charge complexe lie très fortement les chaînes ensemble, ne leur permettant pas de bouger beaucoup. Quand les choses ne bougent pas au niveau moléculaire, elles ne peuvent pas bouger non plus au niveau du matériau entier. C'est pourquoi les polyimides sont si résistants (Figure I-6).

Le transfert de charge est si fort, que quelques fois il devient nécessaire de faire des polymères plus mous pour pouvoir être mis en œuvre. Ceci est réalisé avec pont un de bisphénol-A qui est inséré dans la chaîne du polymère comme il est montré ci-dessous (Figure I-7).

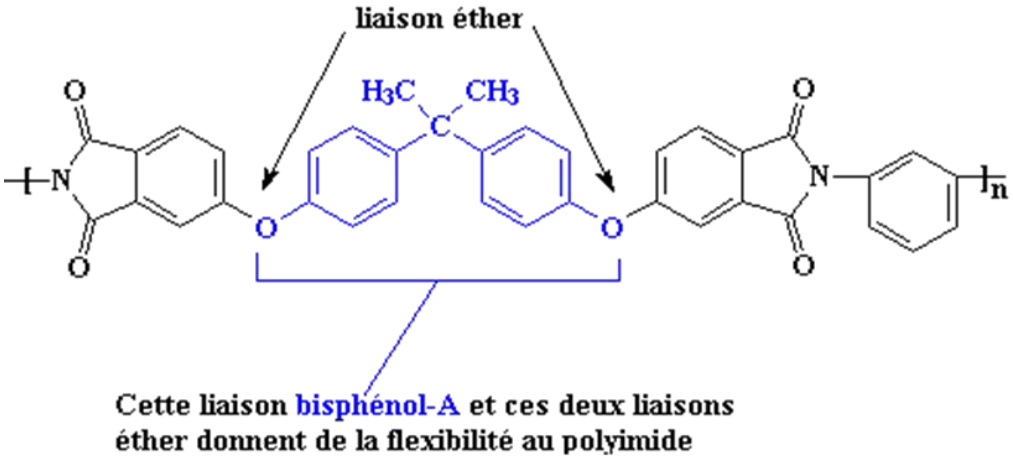


Figure I-7 : Obtention d'un polyimide flexible

I.3. Synthèse des polymères

I.3.1. Réaction de polymérisation

Il y a deux types de réactions permettant de produire des polymères:

- La polyaddition ou bien les réactions d'addition (mécanisme en chaîne)

- La polycondensation ou bien les réactions de condensation (mécanisme par étapes)

Les polymères d'addition résultent de réactions en chaîne provoquées par des centres actifs prenant naissance dans le milieu (Figure I-8). Il y a création d'un site actif localisé à l'extrémité du polymère en croissance. Une fois que la formation de cette entité réactive (ou site actif) a eu lieu, il n'y a plus de barrière à la création d'un très grand nombre de liaisons successives.

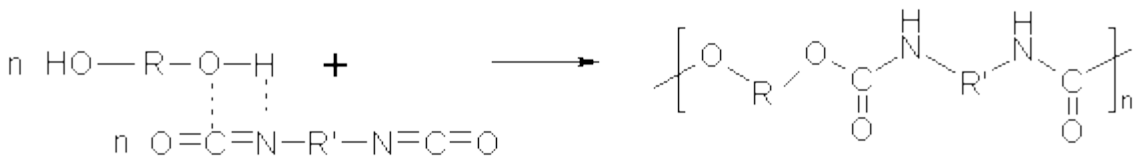


Figure I-8 : Exemple d'une réaction de polyaddition

Il existe trois sortes de site actif:

- Electron célibataire = polymérisation radicalaire (Figure I-9)

- Charge positive = polymérisation cationique (Figure I-10)

- Charge négative = polymérisation anionique (Figure I-11)

Et un quatrième mode, particulier, pour déclencher la polymérisation, c'est la polymérisation par coordination (Figure I-12).

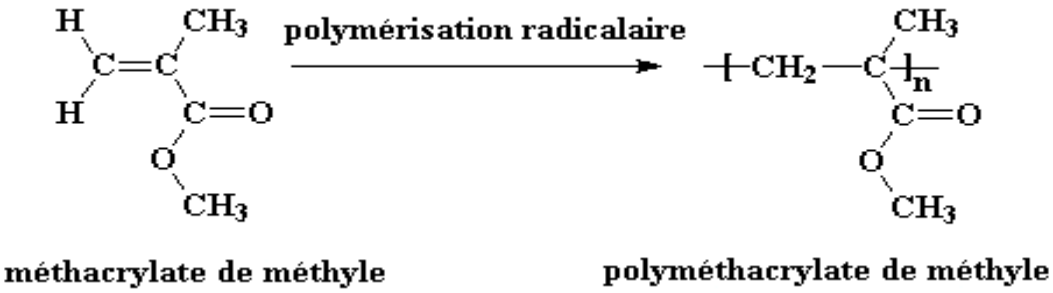


Figure I-9**:** Exemple d'une réaction de polymérisation radicalaire

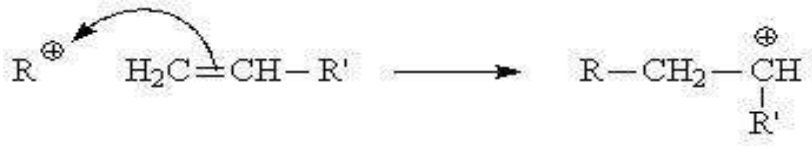


Figure I-10**:** Exemple d'une réaction de polymérisation cationique

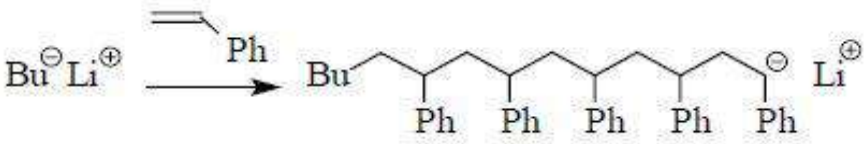


Figure I-11**:**Exemple d'une réaction de polymérisation anionique

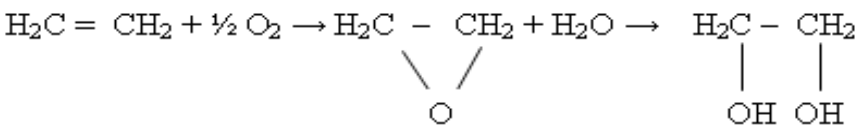


Figure I-12 : Exemple d'une réaction de polymérisation par coordination

Généralement, la polymérisation en chaîne comporte 3 étapes:

- *Amorçage* : une étape au cours de laquelle il y a création d'un nombre limité d'espèces actives.

- *Propagation* : des réactions successives de molécules de monomères sur une extrémité active ou activée conduisant à l'allongement de la chaîne macromoléculaire.

- *Terminaison* : désactivation de l'espèce ou de l'extrémité active et arrêt de la croissance de la chaîne.

Cependant les polymères de condensation sont formés généralement par condensation intermoléculaire par étapes de groupements réactifs. Au cours du processus, la formation du polymère s'effectue avec élimination d'une petite molécule.

I.3.2. Techniques de polymérisation

I.3.2.1. Polymérisation en masse

L'initiateur est ajouté directement dans le monomère ; et là on distingue deux cas :

- Si le polymère est soluble dans le monomère, la phase reste homogène, mais la viscosité augmente exemple: le polystyrène. Le polymère est recueilli puis refroidi et se fige en masse.

-Si le polymère est insoluble dans le monomère, il y a précipitation des macromolécules exemple: le polyméthacrylate de méthyle.

I.3.2.2. Polymérisation en solution

En présence d'un solvant du monomère, dont le rôle est de contrôler la température du milieu réactionnel et donc de maîtriser la réaction chimique. Il est intéressant quand le solvant du monomère est non solvant du polymère (le polymère précipite pendant la polymérisation). C'est un procédé industriel utilisé quand il n'y a pas besoin d'éliminer le solvant (utilisation directe du polymère) : c'est le cas des colles et peintures.

I.3.2.3. Polymérisation en suspension

Le monomère est dispersé sous forme de fines gouttelettes dans un liquide non miscible (ni avec le monomère ni avec le polymère) par exemple l'eau. L'initiateur est dissous dans le monomère par agitation mécanique. On peut considérer chaque gouttelette comme un milieu indépendant qui est le siège de polymérisation en masse. On obtient ainsi des perles de diamètre ~ 1 mm que l'on sépare par centrifugation.

I.3.2.4. Polymérisation en émulsion

Le monomère est dispersé en gouttelettes dans un liquide non miscible comme l'eau, additionné d'un émulsifiant (un savon) qui forme des micelles (agrégats sphériques). Ces micelles peuvent dissoudre un peu de monomères. Il se produit une polymérisation micellaire dont la vitesse de réaction est élevée, avec une bonne maitrise de la température, et un degré de polymérisation 3 élevé d'où une masse moléculaire élevée. Le seul inconvénient serait la séparation du polymère du tensioactif.

I.3.2.5. Polymérisation en phase gazeuse (ou en lit fluidisé)

Un flux gazeux de monomère s'établit de bas en haut d'un réacteur et l'on injecte le catalyseur sous forme de micro grains. Le polymère recouvre progressivement le grain de catalyseur et se trouve en suspension dans le flux gazeux puis est retiré régulièrement: c'est un procédé en continu par exemple pour le PEHD et le PP, ce qui a permis des réductions d'investissements intéressants.

I.4. Mise en forme des polymères

Les matières plastiques thermoplastiques se ramollissent au chauffage : elles sont utilisées sous forme de poudre ou de granulés que l'on fond. On ajoute souvent des agents de moulage, des stabilisants, des colorants, des plastifiants... La mise en œuvre pour obtenir les formes désirées est réalisée par différents procédés dont l'extrusion, moulage par injection, thermoformage, moulage avec soufflage, ou encore moulage avec compression. Le calandrage permet de faire des feuilles ou des plaques.

Pour les polymères thermodurcissables, le chauffage, la mise en forme et la polymérisation provoquant la synthèse et le durcissement du polymère ont lieu simultanément ; le moulage en compression est très souvent utilisé.

Les élastomères sont mis en œuvre de façon similaire aux thermodurcissables : on chauffe et on comprime dans un moule un mélange d'élastomère et d'agent de vulcanisation (réticulation).

3 *Le degré de polymérisation moyen en nombre*: est le degré d'avancement d'une réaction de polymérisation. Il est lié à la viscosité du milieu.

On peut produire des fibres de polymères en forçant le polymère fondu ou en solution à traverser des filières (petits trous). Les fibres formées sont torsadées pour faire un fil puis tissés pour faire une étoffe. Ainsi que la fabrication des mousses en générant des bulles (CO_2) au sein du polymère fondu ou de la résine en train de polymériser lors de sa fabrication.

On a aussi pu produire des membranes polymériques fines, résistantes, sélectives et perméables à certaines espèces chimiques comme les gaz permanents et les vapeurs condensables étudiées dans ce mémoire. Leur fonctionnement repose sur la diffusion Knudsen qui nécessite l'étudie des procédés de séparation membranaires.

I.4.1. Polyimides thermoplastiques

Pour contourner la difficulté de leur mise en œuvre, de nouveaux polyimides ont été fabriqués dès le début des années 1980. Ces polyimides thermoplastiques portent des groupements chimiques qui permettent de les rendre fusibles et solubles et ceci sans altérer de manière significative leurs propriétés. Le tableau suivant donne deux exemples de ces polyimides thermoplastiques qui font l'objet de notre étude (Tableau I-2).

Tableau I-2 : exemple commercial des polyimides et des polyetherimides

Polyimides thermoplastiques	**Exemple**	**Température maximale d'utilisation en continu**
Polyetherimide (PEI)	Ultem (Sabic suite à son achat en 2007 de la division Plastiques de General Electric)	170°C
Polyamideimide (PAI)	Torlon (Solvay Advanced Polymers)	200°C

I.5. Procédés de séparation membranaires

I.5.1. Généralités

Généralement avant chaque procédé de fabrication industriel, des opérations d'extraction, de purification ou de séparation conditionnent leurs performances et déterminent la qualité du produit final. Parmi les différentes classes de cette dernière technique, nous avons la séparation membranaire qui repose sur les propriétés de rétention sélective des membranes vis-à-vis des molécules d'un mélange de fluide (liquide ou gazeux).

Les mélanges à séparer selon leur taille peuvent être :

-Homogènes : liquides miscibles ou gaz

-Hétérogènes : liquides non miscibles, liquide-solide, liquide-gaz.

La partie du mélange retenue par la membrane est appelée *retentât* alors que celle qui traverse cette dernière est appelée *perméat.*

Selon le mode de passage du fluide à travers la membrane, les procédés de séparation membranaires sont classés en :

-Frontales: le flux du produit alimenté est perpendiculaire au filtre ;

-Tangentielles : le flux du produit alimenté est parallèle au filtre, trois possibilités se présentent (Figure I-13):

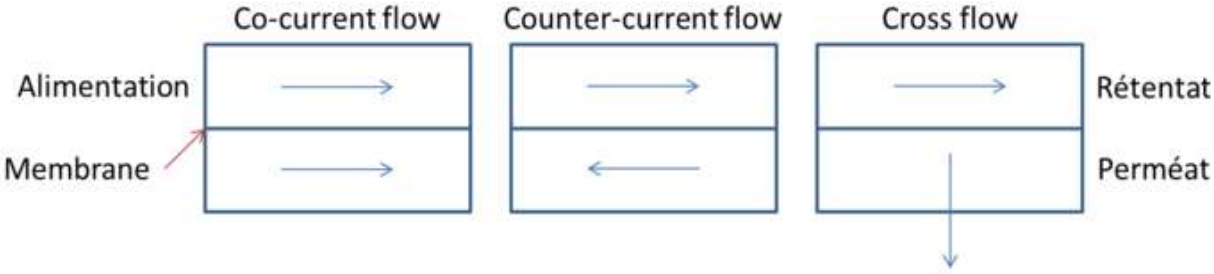


Figure I-13 : Représentation des trois modes de séparation membranaire tangentielle

I.5.2. Avantages et inconvénients de ces procédés

Les méthodes de séparation membranaire permettent de travailler dans les conditions favorables suivantes :

-Une température modérée favorable à la séparation de composés thermosensibles.

-Un caractère compact et modulaire des installations et donc une construction sur mesure par ajout d'éléments standards et une maintenance relativement souple.

-Une extraction de produits en continu de milieux entièrement isolés de l'extérieur par le film séparateur et donc l'élimination de toute contamination croisée entre l'amont et l'aval.

-Une sélectivité élevée.

-Un coût d'exploitation modéré.

-Un temps de séparation relativement court.

-Un grand respect de l'environnement :

-L'absence d'ajout de produits chimiques.

-Un très faible rejet de polluant.

-Un besoin énergétique moindre.

Cependant, ces méthodes souffrent encore de certaines limitations voire inconvénients :

-Un coût d'investissement relativement élevé.

-Un traitement spécifique à certains types d'effluents.

-Le colmatage, par des particules ou des microorganismes et donc la diminution des flux transmembranaires et la nécessité de mettre en œuvre des opérations de lavage.

-La durée de vie limitée des membranes.

-Le coût énergétique résultant des fortes pressions à appliquer.

Toutes les possibilités des procédés à membranes n'ont pas encore été explorées. En particulier, leur couplage avec d'autres procédés de séparations (distillation par exemple) ou procédés hybrides, qui sont des domaines encore relativement peu étudiés.

Lors de notre étude, on s'est intéressé aux phénomènes de la perméation gazeuse du N_2 qui fractionnent un mélange gazeux en mettant à profit l'inégale diffusivité de ses constituants à travers une membrane.

En effet, sous l'action du gradient d'activité établi de part et d'autre du film polymère, les espèces moléculaires pénètrent dans le matériau dense et diffusent selon des profils de concentration liés aux interactions existant avec le polymère. La nature et les propriétés du polymère, élastomère, plastomère, cristallin, polaire, apolaire sont les paramètres clés du transport.

On peut distinguer trois types de montage qui sont représentés dans le schéma de la Figure I-14.

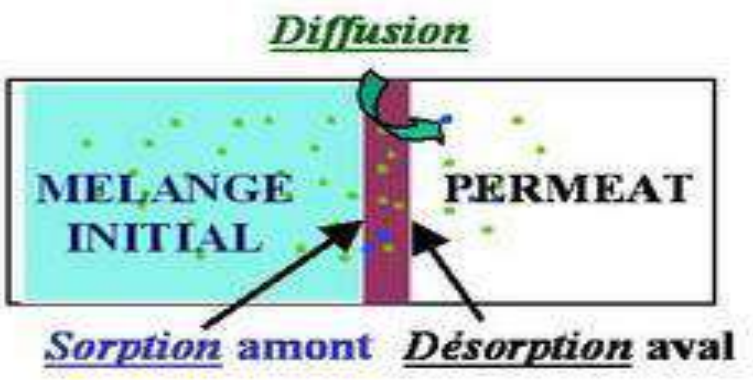
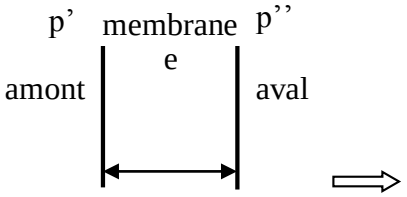


Figure I-14 : schéma du principe de la perméation membranaire

I.5.3. Principe de la perméation gazeuse

Dans le cas d'un gaz pur ; on considère une membrane plane, homogène et isotrope, d'épaisseur e qui est soumise à des pressions constantes amont p', et aval p'', d'un gaz i. En régime stationnaire, il s'établit un profil de concentration stable qui permet de calculer le flux de perméation en intégrant la première loi de Fick.



$$\mathbf{J_i = \frac{1}{e}\int_{c''_i}^{c'_i} D_i(c_i)\,dc_i} \qquad \text{I-1}$$

En appliquant une loi de type Henry aux interfaces, les concentrations limites s'expriment selon :

$$c_i = S_i(p_i) \times p_i \qquad \text{I-2}$$

(dans cette expression, $S_i(p_i)$ est le coefficient de solubilité de Henry, dépendant de la pression partielle p_i) ; il vient alors :

$$J_i = \frac{1}{e}\int_{c''_i}^{c'_i} D_i(S_i(p_i) \times p_i) \times S_i(p_i) \times dp_i \qquad \text{I-3}$$

$$= \frac{p_i'' - p_i'}{e} \frac{c_i'' - c_i'}{p_i'' p_i'} \int_{c_i''}^{c_i'} \frac{D_i(S_i(p_i) \times p_i)}{c_i'' - c_i'} \times dp_i \quad \text{I-4}$$

$$J_i = \frac{\Delta p_i}{e} \overline{S_1} \overline{D_1} = \overline{P_1} \frac{\Delta p_i}{e} \quad \text{I-5}$$

Membrane plane de surface unitaire,

Le produit $\overline{S_1}\overline{D_1}$ (solubilité moyenne*diffusivité moyenne) est appelé *Perméabilité* ou coefficient moyen de perméabilité ($\overline{P_1}$) et a pour dimension le produit d'une longueur par un flux divisé par une pression. Cette grandeur devrait normalement être exprimée en *mole* $m^{-1}s^{-1}Pa^{-1}$ ou en $m^3m^{-1}s^{-1}Pa^{-1}$ dans le système international, mais l'usage fait qu'elle est habituellement exprimée en *Barrer* : $1Barrer=10^{-10}Ncm^3cm\ cm^{-2}s^{-1}(cmHg)^{-1}$. Les conditions normales de mesure des volumes gazeux sont : T=0°C et P=1,013 10^5 Pa.

I.5.4. Méthode de mesure de la perméabilité dans les polymères

La perméabilité peut être obtenue par trois méthodes [2] :

-La méthode de perméation intégrale : elle consiste à la mesure de l'accumulation du gaz en aval du film,

-La méthode de perméation différentielle : elle consiste à mesurer directement le flux de gaz traversant la membrane,

-La méthode de sorption.

Les deux techniques de séparation par perméation présentent de nombreux avantages. Elles ne sont pas destructives et permettent donc le recyclage des gaz ou COV récupérés ce qui peut conduire dans certains cas, à des gains en temps, fonds et énergie.

I.6. Conclusion

Dans ce chapitre, nous avons introduit des notions générales sur les polymères en les définissant et en les classant pour passer après à leur synthèse et leur mise en forme afin de mieux comprendre leur emploi comme membrane dans les procédés de séparation membranaire. Nous avons aussi expliqué brièvement quelques méthodes de mesure de la perméabilité dans les polymères ainsi nous pourrions décrire les différents modèles de sa prédiction dans le prochain chapitre.

Chapitre 2:

Méthodes de réduction

II. Chapitre 2

Aujourd'hui, le but de la recherche en chimie thérapeutique, agrochimique…, est de concevoir de nouveaux composés répondant à différents objectifs: une ou plusieurs activités dont l'une au moins étant supérieurs à l'activité rencontrée pour d'autres molécules ou la combinaison des activités présentant un intérêt biologique particulier; haute tolérance et non toxicité; coût de fabrication faible.

La recherche chimique prédit les propriétés de nouvelles molécules en s'appuyant sur les résultats obtenus par analyse de l'activité de molécules précédemment testées. La comparaison des résultats permet de classer des familles de composés en fonction de leur(s) propriété(s) et de rassembler ces données dans un tableau. L'analyse statistique des données obtenues permet d'obtenir une relation de corrélation générale se présentant sous la forme d'une équation linéaire: appelée relation quantitative structure-activité ou QSAR ou comme dans notre étude relation quantitative structure-propriété ou QSPR. La découverte d'une telle relation permet de guider la synthèse et le test d'autres molécules. Le processus cyclique de QSAR (formulation, prédiction, synthèse et teste) se poursuit jusqu'à ce qu'une molécule possédant l'activité biologique souhaitée soit identifiée. Ceci engendre un nombre relativement important de caractéristiques ou de descripteurs moléculaires qui peut cependant manquer de pertinence sur l'interprétation structurale alors on fait appel à l'Analyse des données.

Pour obtenir un modèle QSAR/QSPR valide, il est nécessaire de respecter certaines conditions:

1. *Utiliser une méthode statistique* qui détecte une relation de type QSAR/QSPR (méthode des moindres carrées -PLS-, régression multilinéaire -MR-, régression linéaire –LR).

2. *Sélectionner les paramètres* à faire intervenir pour les différents composés (paramètres associés aux atomes, associés aux substituants, associés à la topologie 1D ,2D ou à la topologie 3D).

3. *Obtenir des résultats reproductibles de l'activité biologique ou propriété physicochimique* en utilisant des paramètres obtenus dans les mêmes conditions et pour la même conformation moléculaire.

4. *Valider le modèle QSAR/QSPR* obtenu par une méthode.

La validité d'un modèle QSAR dépendra donc du choix que l'on aura fait sur les paramètres. Il est ainsi judicieux d'identifier et d'évaluer quantitativement les paramètres les *plus pertinents* à utiliser en fonction de l'activité et des molécules sélectionnées.

Durant ce chapitre nous expliquerons les descripteurs moléculaires calculés pour construire notre base de données puis les différentes méthodes pour leurs réductions choisies convenablement à notre cas qui est p descripteurs et une perméabilité pour n polymères de la même famille (PolyImides PI), où n est très petit par rapport à p.

II.1. Modélisation moléculaire des structures trimères

La modélisation et la simulation sont devenues des pratiques courantes dans de nombreux domaines scientifiques et techniques et en particulier en Chimie. Elles s'imposent souvent lorsque l'expérience réelle est : trop difficile, trop dangereuse, trop coûteuse, trop longue ou trop rapide, éthiquement inacceptable, ou même impossible à réaliser.

La modélisation moléculaire a pour but de prévoir la structure et la réactivité des molécules ou des systèmes de molécules.

Les méthodes de la modélisation moléculaire comprennent : les méthodes quantiques, la mécanique moléculaire et la dynamique moléculaire.

II.1.1. Les méthodes quantiques

Ces méthodes sont basées sur le calcul des orbitales moléculaires (OM). Leur complexité augmente rapidement avec le nombre d'électrons. Les principales variantes sont :

II.1.2. La méthode de Hückel

C'est la plus simple de toutes. Elle ne prend en compte que les électrons π et utilise des approximations assez draconiennes. Malgré cela, elle permet d'interpréter une grande part de la réactivité chimique.

II.1.2.1. Les méthodes de champ auto-cohérent (Self Consistent Field, SCF)

Ces méthodes prennent en compte les électrons σ et reposent sur des calculs plus élaborés que la méthode de Hückel. On distingue deux variantes, suivant la manière dont les termes d'énergie électroniques sont calculés :

- **Les méthodes ab initio**

Tous les termes énergétiques (bi-électroniques) sont calculés, les temps de calcul étant alors très longs.

- **Les méthodes semi-empiriques**

Les termes énergétiques les plus difficiles à calculer sont estimés à partir de données expérimentales. Les temps de calcul sont considérablement raccourcis, mais la méthode est tributaire des composés qui ont servi à l'étalonner. Selon la nature des approximations utilisées, on distingue plusieurs variantes (MNDO, AM1,...). Cette dernière méthode de calcul est utilisée au cours de ce travail (Austin Model 1) développée par Dewar. Ce modèle AM 1 est basé sur une approche Hartee-Fock qui réalise un certain nombre d'approximations de manière à réduire le nombre d'intégrales bi électronique et gagner du temps [3].

II.2. Descripteurs moléculaires

De nombreuses recherches ont été menées, au cours des dernières décennies, pour trouver la meilleure façon de représenter l'information contenue dans la structure des molécules, en un ensemble de nombres réels appelés descripteurs.

Une fois que ces nombres sont disponibles, il est possible d'établir une relation entre ceux-ci et une propriété ou activité moléculaire, à l'aide d'outils de modélisation. Ces descripteurs numériques réalisent de ce fait un codage de l'information chimique en un vecteur de réels.

On en dénombre aujourd'hui plus de 3200 types de descripteurs moléculaires, qui quantifient des caractéristiques physico-chimiques ou structurelles de molécules. Ils peuvent être obtenus de manière empirique ou non-empirique, mais les descripteurs calculés, et non mesurés, sont à privilégier : ils permettent en effet d'effectuer des prédictions sans avoir à synthétiser les molécules, ce qui est un des objectifs de la modélisation.

Avant toute modélisation, il est nécessaire de calculer ou de mesurer un grand nombre de descripteurs différents, car les mécanismes qui déterminent l'activité d'une molécule ou une de ses propriétés sont fréquemment mal connus. Il faut ensuite sélectionner parmi ces variables celles qui sont les plus pertinentes pour la modélisation.

II.2.1. Types de descripteurs moléculaires

Le descripteur moléculaire est le résultat final d'une procédure logique et mathématique qui transforme l'information chimique chiffrée dans une représentation symbolique d'une molécule en un nombre utile qui est le résultat de quelque expérience standard.

Nous allons présenter les descripteurs moléculaires les plus courants, en commençant par les descripteurs les plus simples, qui nécessitent peu de connaissances sur la structure moléculaire, mais véhiculent peu d'informations. Nous verrons ensuite comment les progrès de la modélisation moléculaire ont permis d'accéder à la structure 3D de la molécule, et de calculer des descripteurs à partir de cette structure.

II.2.1.1. Groupes ou Blocs des descripteurs moléculaires

Les descripteurs moléculaires sont les traits communs les plus considérables de structure moléculaire qui peut être utilisée pour développer la « Relation Structure – property». En effet, les descripteurs moléculaires sont basés sur plusieurs théories différentes, tel que : quantum chimie, théorie de l'information, chimie organique, théorie du graphique, et ainsi de suite.

Actuellement, il est possible de calculer plus de 3200 descripteurs par le logiciel DRAGON [4], répartis sur 22 blocs dont les deux derniers représentent la nouveauté de cette version(4) et selon la structure calculée (Tableau II-2).

- **Descripteurs 1D**

Sont accessibles à partir de la formule brute de la molécule (par exemple C_6H_6O pour le phénol), et décrivent des propriétés globales du composé. Il s'agit par exemple de sa composition, c'est-à-dire les atomes qui le constituent, ou de sa masse molaire. On peut remarquer que ces descripteurs ne permettent pas de distinguer les isomères de constitution.

Tableau II-1:Nombre de descripteurs calculés par chaque famille des 22 blocs selon DRAGON (4)

Blocs descripteurs	Nombre de descripteurs calculés	Blocs descripteurs	Nombre de descripteurs calculés
Descripteurs constitutionnels	48	Descripteurs topologiques	119
Comptes de marche et chemin	47	Indices de connectivités	33
Indices d'information	47	Autocorrélations 2D	96
Indices de contiguïté affilez	107	Charges des valeurs propres	64
Indices de charges topologiques	21	Indices basés sur les valeurs propres	44
Profils moléculaires Randic	41	Descripteurs géométriques	74
Descripteurs RDF	150	Descripteurs 3D morse	160
Descripteurs WHIM	99	Descripteurs escapade	197
Compte groupe fonctionnel	154	Fragments d'atome centré	120

Descripteurs de charge	14	Propriétés moléculaires	29
Empreintes digitales 2D binaires	780	Empreintes digitales 2D de fréquences	780

Tableau II-2 : Répartitions des descripteurs selon le type de structure par DRAGON (4)

Couleur de descripteur					
Descripteur de structure :	0D	1D	2D	3D	autres

- **Descripteurs 2D**

Sont calculés à partir de la formule développée de la molécule. Ils peuvent être de plusieurs types :

— **Les indices constitutionnels :** caractérisent les différents composants de la molécule. Il s'agit par exemple du nombre de liaisons simples ou multiples, du nombre de cycles…

— **Les indices topologiques :** peuvent être obtenus à partir de la structure 2D de la molécule, et donnent des informations sur sa taille, sa forme globale et ses ramifications. Les plus fréquemment utilisés sont l'indice de Wiener, l'indice de Randic, l'indice de connectivité de valence de Kier-Hall et l'indice de Balaban. L'indice de Wiener permet de caractériser le volume moléculaire et la ramification d'une molécule : si l'on appelle distance topologique entre deux atomes le plus petit nombre de liaisons séparant ces deux atomes, l'indice de Wiener est égal à la somme de toutes les distances topologiques entre les différentes paires d'atomes de la molécule. L'indice de Randic est un des descripteurs les plus utilisés ; il peut être interprété comme une mesure de l'aire de la molécule accessible au solvant.

Ces descripteurs 2D reflètent bien les propriétés physiques dans la plupart des cas, mais sont insuffisants pour expliquer de façon satisfaisante certaines propriétés ou activités. Des descripteurs, accessibles à partir de la structure 3D des molécules, ont pu être calculés grâce au développement des techniques instrumentales et de nouvelles méthodes théoriques.

- **Descripteurs 3D**

D’une molécule sont évalués à partir des positions relatives de ses atomes dans l'espace, et décrivent des caractéristiques plus complexes; leurs calculs nécessitent donc de connaître, le plus souvent par modélisation moléculaire empirique, la géométrie 3D de la molécule. Ces descripteurs s'avèrent donc relativement coûteux en temps de calcul, mais apportent davantage d'informations, et sont nécessaires à la modélisation de propriétés ou d'activités qui dépendent de la structure 3D. On distingue plusieurs familles importantes de descripteurs 3D :

— **Descripteurs géométriques** les plus importants sont le volume moléculaire, la surface accessible au solvant, le moment principal d'inertie.

— **Descripteurs électroniques** permettent de quantifier différents types d'interactions inter- et intramoléculaires, de grande influence sur l'activité biologique de molécules. Le calcul de la plupart de ces descripteurs nécessite la recherche de la géométrie pour laquelle l’énergie stérique est minimale, et fait souvent appel à la chimie quantique. Par exemple, les énergies de la plus haute orbitale moléculaire occupée et de la plus basse vacante sont des descripteurs fréquemment sélectionnés. Le moment dipolaire, le potentiel d'ionisation, et différentes énergies relatives à la molécule sont d'autres paramètres importants.

— **Descripteurs spectroscopiques :** les molécules peuvent être caractérisées par des mesures spectroscopiques, par exemple par leurs fonctions d'onde vibrationnelles. En effet, les vibrations d'une molécule dépendent de la masse des atomes et des forces d'interaction entre ceux-ci ; ces vibrations fournissent donc des informations sur la structure de la molécule et sur sa conformation. Les spectres infrarouges peuvent être obtenus soit de manière expérimentale, soit par calcul théorique, après recherche de la géométrie optimale de la molécule. Ces spectres sont alors codés en vecteurs de descripteurs de taille fixe. Le descripteur EVA est ainsi obtenu à partir des fréquences de vibration de chaque molécule. Les descripteurs de type MoRSE *(*Molecule Representation of Structures based on Electron diffraction*)* sont calculés à partir d'une simulation du spectre infrarouge ; ils font appel au calcul des intensités théoriques de diffraction d'électrons.

Afin d’établir des modèles de prédiction de la perméabilité des polymères, des outils de chémoinformatique permettent une meilleure manipulation de l’ensemble des données bien collectés.

Après l'injection de données obtenues, on passe à leur analyse statistique et réduction par ce qu'on appelle « Analyse des données ».

II.3. L'analyse de données ou « Data mining »

L'*analyse des données* est une famille de méthodes statistiques dont les principales caractéristiques sont d'être multidimensionnelles et descriptives. Certaines méthodes, pour la plupart géométriques, aident à faire ressortir les relations pouvant exister entre les différentes données et à en tirer une information statistique qui permet de décrire de façon plus brève les principales informations contenues dans ces données. D'autres techniques permettent de regrouper les données de façon à faire apparaître clairement ce qui les rend homogènes, et ainsi mieux les connaître.

Dans l'acception française, la terminologie *analyse des données* sert à désigner un sous-ensemble de ce qui est nommé d'une façon plus générale la *statistique multivariée*. Elle comprend essentiellement :

- L'*analyse en composantes principales* (ACP), utilisée pour des données quantitatives, comme pour notre étude.

- L'*analyse factorielle discriminante* (AFD) ou *analyse discriminante* qui permet d'identifier des groupes homogènes au sein de la population du point de vue des variables étudiées,

- L'*analyse factorielle des correspondances* (AFC), utilisée pour des données qualitatives nominales (tableau d'association).

- La *classification automatique*.

- L'*analyse en composantes indépendantes* (ACI).

Ces méthodes dont l'esprit ne diffère pas, permettent surtout de manipuler et de synthétiser l'information provenant de tableaux de données de grande taille. Pour cela, il est particulièrement important de bien estimer les corrélations entre les variables qu'on étudie. On a alors fréquemment recours à la matrice des corrélations (ou la matrice de variance-covariance) entre les variables ; par des logiciels permettant d'effectuer de l'analyse des données directe ou indirecte.

II.4. But de l'analyse de données

Le « Data mining » est un processus désigné pour explorer des données de grandes tailles ('Big data '), à fin d'obtenir des schémas cohérents et/ou des relations systématiques entre les variables puis de valider les résultats en appliquant les modèles détectés à de nouveaux sous-ensembles de données. Le but final d'une analyse de données est la prédiction selon les étapes suivantes : 1-l'explorationinitiale ; 2-la construction de modèles ou l'identification de motif avec validation/vérification, 3-le déploiement (i.e, l'application du modèle aux nouvelles données afin de générer des prédictions).

L'analyse des données repose sur l'utilisation de techniques statistiques, dans une recherche quantitative. A ce niveau, un premier choix se présente à nous entre le recours à des statistiques descriptives ou à des statistiques inférentielles ou encore l'utilisation combinée de ces deux techniques.

Les « statistiques descriptives » servent à décrire et à présenter les données par des indices de tendance centrale comme la moyenne, le mode, la médiane ou encore d'indices qui révèlent l'allure de la distribution comme l'étendue, l'écart-type, la symétrie ou l'aplatissement (kurtosis). Pour prolonger ces techniques descriptives, on trouve des techniques dites *exploratoires* dont le but est de faciliter la visualisation des données par l'utilisation de représentations graphiques adaptées. Ce type de statistique ne permet pas de faire d'inférence ou de prédiction à partir des données recueillies, mais rapporte simplement les résultats sous une forme synthétique qui facilite leur interprétation.

Les « statistiques inférentielles » ont pour ambition de réaliser des inférences et des prédictions à partir des données rassemblées. Il existe de nombreuses techniques inférentielles qui permettent de tester des hypothèses en comparant des moyennes ou des variances (t de Student, analyse de variance) ou encore de vérifier le lien entre des variables (corrélation et régression). Le choix de la technique statistique qui sera utilisée pour traiter les données dépend aussi du type de données : nominaux, ordinaux, à intervalles ou métriques.

Il existe de nombreuses techniques qui relèvent de l'analyse exploratoire. Tout d'abord des techniques de base qui n'exigent que des calculs simples comme les diagrammes, les histogrammes, les boites à moustache (*boxplot*), les diagrammes de série chronologique. A côté de ces techniques, on trouve des méthodes d'analyses exploratoires qui exigent des calculs plus

complexes comme l'analyse factorielle en composantes principales, l'analyse des correspondances ou encore l'analyse hiérarchique par cluster.

II.5. Analyse en composantes principales

L'analyse en composantes principales (ACP), développée en France dans les années 1960 par JP.Benzécri, est une méthode de statistique exploratoire permettant de décrire un grand tableau de données de type individus / variables. Lorsque les individus sont décrits par un nombre important de variables, aucune représentation graphique simple ne permet de visualiser le nuage de points formé par les données. L'ACP propose une représentation dans un espace de dimension réduite, permettant ainsi de mettre en évidence d'éventuelles structures au sein des données. Pour cela, nous recherchons les sous-espaces dans lesquels la projection du nuage déforme le moins possible le nuage initial [5].

II.5.1. Présentation et description de la méthode factorielle : Analyse en composantes principales (ACP)

II.5.1.1. Les données

Les données sont sous la forme d'un tableau à n lignes et p colonnes, que l'on stocke sous la forme d'une matrice X de taille n × p :

– chaque ligne $(x^1_i, \ldots, x^p_i)$ de X représente les valeurs prises par l'**individu** i sur les p variables,

– de même chaque colonne $(x_j^1, \ldots, x_j^n)^t$ de X représente les valeurs de la **variable** j pour les n individus.

Par simplicité de langage, l'individu i sera identifié au vecteur $x_i = (x^1_i, \ldots, x^p_i)^t$ de R^p tandis que la variable j sera identifiée au vecteur $x_j = (x_j^1, \ldots, x_j^n)^t$ de R^n. L'ensemble des données peut être schématisé par une matrice X à n lignes et p colonnes.

$$X^j = \begin{bmatrix} X_{1j} \\ X_{2j} \\ \vdots \\ X_{nj} \end{bmatrix} \qquad \mathbf{X} = \text{Individus}$$

Variables quantitatives

	1		j		p	
1	x_{11}	x_{12}	x_{1j}		x_{1p}	1
i	x_{i1}	x_{i2}	x_{ij}		x_{ip}	i
n	x_{n1}		x_{nj}		x_{np}	n

Nous travaillons désormais en ACP normée, avec le tableau de données centrées $Z = (X - 1g^t)D_{1/s}$ et la métrique identité. Les individus et les variables sont désormais les z_i et z^j. L'ACP consiste à chercher un sous-espace F_k de dimension k inférieure à celle de l'espace de départ, tel que le nuage, une fois projeté dans ce sous-espace, soit au minimum déformé. Comme la projection diminue nécessairement les distances, on cherche le sous-espace F_k qui maximise la moyenne des carrés des distances entre individus :

$$\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} d^2(i,j) = 2I_g \qquad \text{(II-1)}$$

Il faut donc que l'inertie du nuage projeté soit maximale.

Il est donc possible de rechercher le sous espace F_k séquentiellement axe par axe, en cherchant tout d'abord l'axe portant l'inertie maximale, puis celui orthogonal à ce dernier portant l'inertie maximale, et ainsi de suite.

II.5.2. Interprétation des résultats d'une ACP

II.5.2.1. Qualité de représentation sur les axes factoriels

Un des points les plus délicats de l'ACP est d'apprécier la perte d'information engendrée par la réduction de la dimension. On a vu précédemment que la qualité de représentation du nuage dans le sous-espace factoriel est exprimée par le pourcentage d'inertie expliquée $\frac{\lambda_1+\cdots+\lambda_K}{p}$.

Cette mesure globale doit être complétée par d'autres considérations comme la représentation des individus et des variables dans ce sous-espace factoriel. Il peut en effet arriver que les axes retenus, bien que représentant une part importante de l'inertie globale du nuage, ne suffisent pas à expliquer correctement certains individus ou variables : deux individus peuvent être proches en projection sur les axes factoriels retenus tout en étant très éloignés en réalité, si leurs représentations sur ces axes ne sont pas de bonne qualité.

La **qualité de représentation d'un individu** i sur un axe factoriel est mesurée par le **cosinus carré** de l'angle entre l'axe factoriel et le vecteur z_i. Plus le cosinus est grand, plus z_i sera proche de l'axe factoriel et donc sera bien représenté sur cet axe.

La **qualité de représentation d'une variable** j sur le k-ème axe factoriel est exprimée par le coefficient de corrélation linéaire $r(c^k, z^j)$ entre la variable initiale z^j et la nouvelle variable, composante principale, c^k. La valeur de cette corrélation sera également très importante pour interpréter les nouveaux axes factoriels en fonction des variables initiales. Le calcul de cette corrélation, non développé dans ce document, montre que

$$\mathbf{r(c^k, z^j) = \sqrt{\lambda_k} v_k^j} \textit{(II-1)} \qquad \text{II-2}$$

Où $\boldsymbol{v_k^j}$ est la j-ème coordonnée du vecteur v_k.

Ainsi, bien qu'ayant retenu un sous-espace factoriel expliquant une part importante de l'inertie totale, il est possible que certaines variables ou individus d'intérêt soient mal représentés dans ce sous-espace. Il sera alors intéressant de compléter le sous-espace factoriel en ajoutant des axes factoriels supplémentaires de sorte que ces variables ou individus d'intérêt soient bien représentés.

II.5.2.2. Sélection du nombre k d'axes factoriels à retenir

La sélection du nombre d'axes à retenir, abordée précédemment, est une étape importante d'une ACP. Les critères utilisés sont en général empiriques : la méthode du coude qui consiste à détecter un coude sur le diagramme des valeurs propres, ou le critère de Kaiser. Ce dernier consiste à ne retenir que les valeurs propres supérieures à la moyenne : puisqu'on est en ACP normée la somme des valeurs propres est égale au nombre p de ces valeurs propres, et la moyenne des valeurs propres est donc égale à 1. On ne retiendra donc par le critère de Kaiser que les valeurs propres supérieures à 1.

II.5.2.3. Interprétation

La méthode la plus naturelle pour interpréter les composantes principales (ou les axes factoriels) est de les relier aux variables initiales, en utilisant les corrélations précédemment introduites. Généralement, pour un couple de composantes principales c^1 et c^2, on représente les corrélations dans une figure appelée cercle de corrélation, où chaque variable z^j est représentée par un point de coordonnées $(r(c^1, z^j), r(c^2, z^j))$.

Une fois les composantes principales interprétées, on représente les individus dans les plans factoriels formés en croisant deux à deux les axes factoriels retenus. Chaque représentation est interprétée en essayant de repérer des groupes d'individus et en donnant à ces groupes une signification en fonction de leur place dans les plans factoriels.

Rappelons que notre étude ne s'intéresse pas aux interprétations et représentations graphiques, mais à la réduction de notre base de données réalisée par cette méthode à fin de la comparée à celle obtenue par la méthode de régression des moindres carrées partiels PLSR dont on introduit la description.

II.5.3. Limites

Comme pour toute autre méthode exploratoire nous pouvons citer deux limites principales de l'ACP :

• la première est la visualisation globale des données. Dans certains cas, cette visualisation est suffisante ; dans d'autres, par contre, elle permet seulement de situer, dans l'ensemble des données, une recherche plus poussée, qui peut être soit définie a priori, soit établie à la lumière des résultats de l'ACP.

• la deuxième limite est technique. La mise en œuvre de l'ACP demande le calcul préalable de la matrice carrée de covariance des données, qui est de taille p^2 pour des vecteurs de dimension p. Cette matrice est déjà couteuse à calculer, et sa taille et son traitement deviennent prohibitifs en haute dimension. Ainsi des données de dimension 1000 donneront lieu à une matrice d'un million d'´éléments. D'autres méthodes peuvent être élaborées pour remédier à ce problème.

II.6. Analyse des moindres carrés partiels

De nombreux problèmes industriels peuvent être décrits sous la forme d'un modèle de régression, ou l'on possède des variables X sur lesquelles on peut plus ou moins agir et des variables Y que l'on ne peut qu'observer.

L'objectif est alors de décrire les relations entre Y et X, en l'absence de modèle théorique. Le problème est que le nombre de variables X est souvent très important par rapport au nombre d'observations. La régression PLS (Partial Least Squares Regression) est une méthode d'analyse des données spécifiquement construite pour l'étude de ce type de problème.

II.6.1. Présentation

La régression PLS est donc une extension du modèle de régression linéaire multiple. Dans sa forme la plus simple, un modèle linéaire spécifie la relation (linéaire) entre une (ou des) variables dépendantes (réponses) Y et un ensemble de variables prédictives X telles que :

$$Y = b_0 + b_1 X_1 + b_2 X_2 + \ldots + b_p X_p \qquad \text{II-3}$$

où les *bi* sont les coefficients de régression.

Dans beaucoup de problèmes d'analyses de données, l'estimation de la relation linéaire entre deux variables est adéquate pour décrire les données observées et pour faire de bonnes prévisions pour de nouvelles observations.

Le modèle de régression multiple a été étendu de plusieurs façons afin de s'adapter aux problèmes d'analyse de données plus sophistiqués. Il sert donc de base pour de nombreuses méthodes multivariées comme l'analyse discriminante, la régression sur composantes principales (RCP) et la corrélation canonique.

La régression PLS est une technique récente qui généralise et combine les caractéristiques de l'analyse sur composantes principales et de la régression multiple.

Elle est particulièrement utile quand on a besoin de prédire un ensemble de variables dépendantes partir d'un ensemble très grand de variables explicatives (prédicteurs) qui peuvent être très fortement corrélées entre elles.

Quand les prédicteurs sont peu nombreux, non significativement colinéaires et ont une relation connue avec les réponses, alors la régression linéaire multiple est la meilleure méthode pour utiliser les données. Cependant si l'une de ces trois conditions n'est pas vérifiée, la régression linéaire multiple peut être inefficace et inappropriée.

PLS est donc une méthode pour construire des modèles de prédiction quand les facteurs sont nombreux et très colinéaires.

Notons que cette méthode met l'accent sur la prédiction de la réponse et pas nécessairement sur la mise en évidence d'une relation entre les variables. Ce qui signifie que PLS n'est pas appropriée pour désigner les variables ayant un effet négligeable sur la réponse, mais quand le but est la prédiction et qu'il n'y a pas besoin de limiter le nombre de variables mesurées, PLS est un outil très utile.

II.6.2. Intérêt de la Régression PLS

La régression PLS pour le modèle linéaire s'applique au champ d'application où l'on possède peu d'observations sur des variables très corrélées et en très grand nombre, c'est le cas pathologique de la régression linéaire, là où les méthodes habituelles ne fonctionnent pas (n étant très petit on ne peut pas tendre vers l'infini).

II.6.3. La méthode

La méthode PLS est une méthode de régression linéaire de *c* variables réponses sur *p* variables explicatives toutes mesurées sur les mêmes *n* individus. Les tableaux des observations, notes respectivement *Y* et *X*, de dimensions *n*c* et *n*p*, sont supposés centrés et éventuellement réduits par rapport aux poids ($p_1; \ldots; p_n$). On note $D = diag(p_1; \ldots ; p_n)$ la matrice diagonale des poids.

L'intérêt de la méthode comparée à la régression sur composantes principales (RCP), réside dans le fait que les composantes PLS sur les *X*, noté es*t*, sont calculées "dans le même temps" que des régressions partielles sont exécutées. Cette simultanéité leur confère un meilleur pouvoir prédictif que celles de la RCP. La question est donc d'examiner comment cette simultanéité est mise en œuvre.

Notons $E_0 = X$ et $F_0 = Y$ les tableaux centrés et réduits au sens de *D* qui en général est égal à $1/nI_n$. La méthode procède par étapes successives permettant le calcul des composantes principales. On notera *A* le nombre total d'étapes, c'est-à-dire de composantes indicées par $k = 1; \ldots ; A$.

II.6.4. Choix du nombre de composantes

La borne supérieure de *A* sera évidemment le rang de la matrice de départ, c'est à dire $A \cdot rang(X)$. On aura des critères permettant de déterminer *A* de façon raisonnable.

Critères du choix de *A*

CR1 Ajustement ("FIT") sur les données

CR2 Prédiction interne ("Validation Croisée")

CR3 Prédiction externe sur un jeu de données tests ("Validation Externe")

On ne veut pas avoir un sur ajustement donc le FIT ne devra pas être trop grand (*A* trop grand), car cela dégradera la prédiction. Donc il faut une balance entre le critère "ajustement" et le critère "prédiction" (interne ou externe).

Critère 1 : Le "FIT" ou "Critère d'Ajustement"

Le nombre total *A* d'axes ou de composantes (la dimension du modèle) est le super paramètre de la méthode. Lors de l'exécution de l'algorithme, il est utile de voir évoluer, à chaque nouvel axe construit, la reconstruction de la variance des prédicteurs et des réponses. La stratégie consiste à s'arrêter lorsque le pourcentage de la variance de *X* est suffisamment grand, pour un gain faible dans le pourcentage de variance de *Y*. Ce critère est donc basé sur l'ajustement des données.

En termes d'ajustement il est aussi intéressant de regarder l'évolution des coefficients du modèle de régression linéaire $\hat{Y}_A = X\hat{\beta}_A$ en fonction de la valeur de A.

Pour une réponse *j* choisie, l'étendue des valeurs du vecteur $\hat{\beta}_A^j$ reste stable jusqu'à une certaine valeur de *A*. Au-delà de cette valeur *A*, pour certains prédicteurs fortement corrélés, les valeurs des coefficients $\hat{\beta}_A^j$ semblent exploser. On choisit alors la plus petite valeur *A* à partir de laquelle ce phénomène apparait.

Critère 2 : Le "PRESS" ou "Critère de Validation Croisée"

Après avoir enlevé un individu *i* (une ligne) aux matrices *X* et *Y* (ou10% des individus si leur nombre est élevé), on calcule $\hat{\beta}_A^j$, qui représente la matrice des coefficients du modèle construit avec les individus restants. Puis, on calcule l'erreur de prédiction faite sur l'individu i, $E^{(i)}{}_A = Y_i - X_i \hat{\beta}_A^j$. L'erreur de prédiction est définie par le PRESS, Prédiction Sum of Squares :

$$PRESS(A) = \frac{1}{n}\sum_{i=1}^{n} \|E_A^{(i)}\|^2. \qquad \text{II-4}$$

On choisit donc *A* tel que *PRESS*(*A*) soit le plus petit possible.

Critère 3 : "Critère de Validation Externe" et Prédiction

Pour utiliser ce critère de choix de *A*, on doit posséder outre l'échantillon d'apprentissage (*X; Y)* sur les *n* individus, un échantillon test (X_{test}; Y_{test}) sur *N* observations mesurées sur les mêmes variables. Différents modèles PLS basés sur (*X; Y)* sont calculés pour différentes dimensions. La valeur optimale de *A* correspond à la plus petite des erreurs de prédiction calculée sur (X_{test}; Y_{test}).

II.7. Les approches génétiques et sélection

Les algorithmes génétiques (AG) ont montré avec succès leur grande capacité à résoudre des problèmes d'optimisation. Ils ont aussi été utilisés dans le domaine de la sélection de caractéristiques. De nombreuses études rapportées dans la littérature ont montré que les méthodes qui utilisent les AGs comme technique de recherche ont donné des meilleurs résultats en comparaison avec les autres méthodes de sélection. Dans ce chapitre, nous commençons d'abord par décrire les concepts de base des algorithmes génétiques ainsi que les techniques d'optimisation multi-objectifs et ensuite nous aborderons la présentation des applications des algorithmes génétiques dans les domaines de la sélection de caractéristiques

II.7.1. Algorithmes génétiques

Les algorithmes génétiques (AG) sont des méthodes d'optimisation stochastiques qui font partie des algorithmes évolutionnaires. Les algorithmes évolutionnaires sont des métas heuristiques qui s'inspirent des mécanismes d'évolution darwinienne des populations biologiques.

Les AGs ont été utilisés pour la première fois en 1950 par des biologistes dans le but de simuler l'évolution des organismes. La première adaptation aux problèmes d'optimisation combinatoire de ces algorithmes a été réalisée dans les années 70 par John en développant une analogie entre un individu dans une population et une solution d'un problème dans un ensemble de solutions potentielles. Des travaux de recherche ont été développés ensuite par Goldberg et Michalewicz. Ces travaux ont montré la possibilité d'appliquer les AGs à des problèmes concrets.

II.7.1.1. Principe

Les AGs sont des algorithmes itératifs basés sur la reproduction et l'évolution naturelle des individus en utilisant les principes de la survie des individus considérés comme les plus forts ou les mieux adaptées à l'environnement. Il s'agit alors de combiner les points forts de chaque individu pour en créer de nouveaux de manière à ce que leur efficacité soit meilleure. Avec les AGs on cherche à optimiser une fonction (objectif) donnée dans un espace de recherche, celui des individus. Pour l'optimiser, on définit une fonction d'évaluation (fitness) reliée à cette fonction objective et appliquée sur chaque individu ou chromosome.

En général, le fonctionnement d'un AG est basé sur les phases suivantes (Figure II-1) :

1. Initialisation : Générer aléatoirement une population initiale de taille N chromosomes.

2. Evaluation : Evaluer chaque individu de la population par la fonction d'évaluation appropriée au problème (fonction de fitness).

3. Reproduction : Créer une nouvelle population de N chromosomes par l'utilisation d'une méthode de sélection appropriée et l'application d'opérateurs génétiques (croisement et mutation) sur certains chromosomes au sein de la population courante.

4. Retour à la phase 2 tant que la condition d'arrêt du problème n'est pas satisfaite.

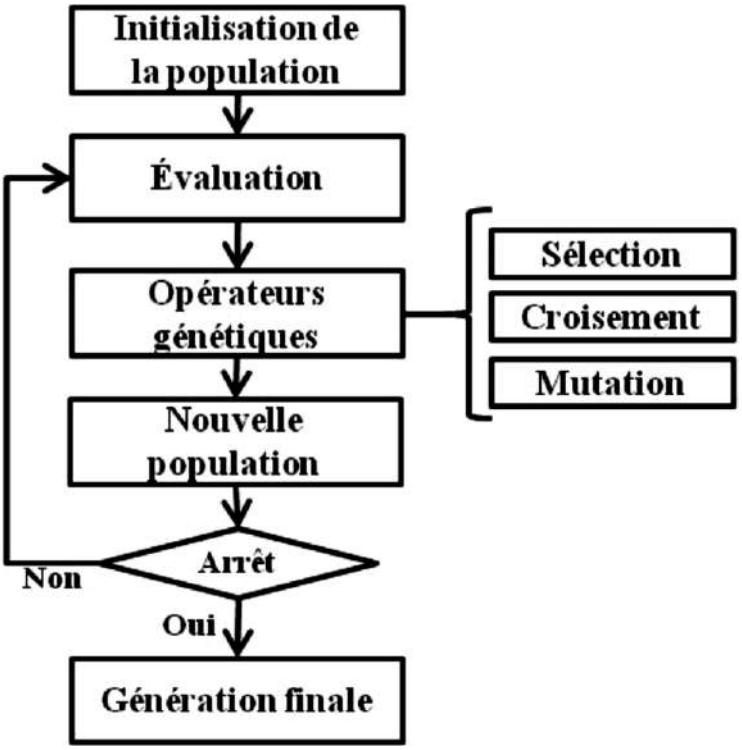


Figure II-1: Architecture générale d'un algorithme génétique

II.7.1.2. Opérateurs génétiques

La reproduction joue un rôle fondamental pour le passage d'une génération à une autre. Elle représente un cycle de l'algorithme génétique. Elle se fait par l'application d'opérations génétiques : la sélection, le croisement et la mutation [6].

II.8. Application des réseaux des neurones :

Un réseau de neurones artificiel (RNA) est un processeur massivement distribué en parallèle, composé d'unités de traitement simples, qui a une propension naturelle pour stocker la connaissance empirique et de la rendre disponible à l'usage. Il ressemble au cerveau humain sur deux aspects:

- La connaissance est acquise par le réseau de son environnement à travers un processus d'apprentissage.
- Les forces de liaison inter-neuronale, connues sous le nom de poids synaptiques, sont employées pour stocker la connaissance acquise.

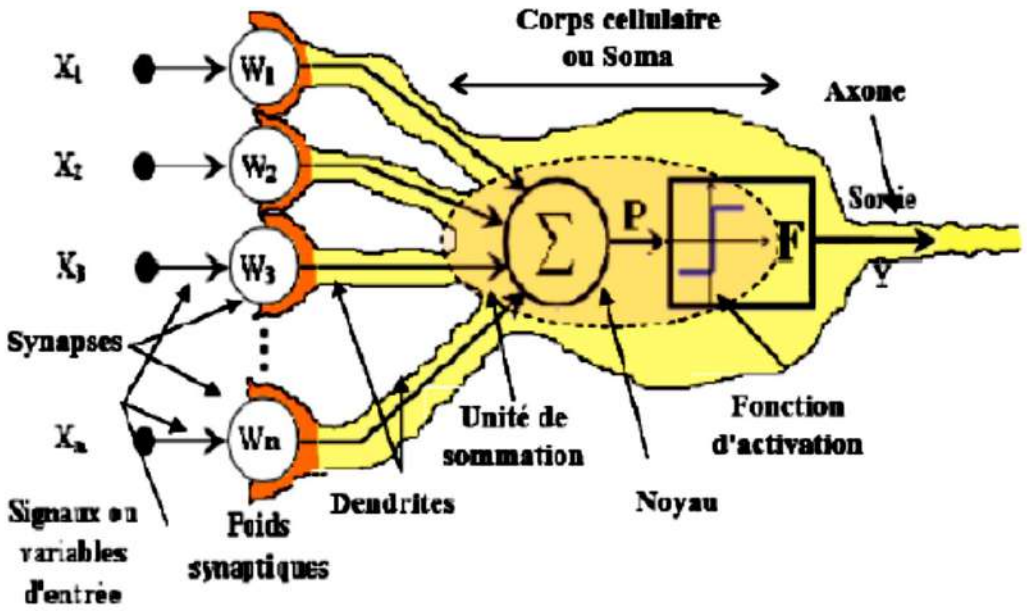


Figure II-2 : Comparaison schématique entre un neurone biologique et neurone artificiel

La Figure II-2 résume la correspondance entre le neurone artificiel et le neurone biologique. On remarque, -dans cette description schématique, que chaque composant ou constituant du neurone artificiel trouve son inspiration biologique. Il existe plusieurs types de réseaux de neurones, mais celui de type feed forward (existe par défaut dans les interfaces graphiques de réseaux de neurones, ***nntool sur MATLAB)*** le plus utilisé pour l'approximation des fonctions des données statiques a été employé comme application dans notre étude.

II.9. Logiciels de calculs

On trouve sur le marché informatique une panoplie de logiciels dédiés aux calculs et de résolution des problèmes des grandes tailles qui emploient ces méthodes par exemple : le XLSTAT ; Tanagra, Matlab ;IBM SPSS, MaterialStudio ; Statistica ...

II.10. Conclusion

Dans ce chapitre nous avons présenté la modélisation moléculaire et les méthodes de réduction des bases de données appropriées à notre étude.

Nous avons expliqué la méthode d'analyse en composante principale ; comparativement avec la méthode des moindres carrées partiels avec quatre de ses variantes ainsi que les algorithmes génétiques en tant que méthode, passons par un rappel des réseaux de neurones qui sera employé en tant qu'application du modèle de prédiction de la perméabilité des polyimides aux gaz permanents et nous prenons le N_2 comme exemple.

Chapitre 3:

Réduction & analyse

III. Chapitre 3

Ce chapitre comporte deux parties ; la première expose la méthodologie suivie pour construire notre base de données, en commençant par la récolte et la représentation des structures jusqu'au calcul des descripteurs par les logiciels appropriés ; puis le traitement et la réduction des descripteurs –qui constitue l'objectif de notre travail- par différentes méthodes de régression linéaire et non-linéaire.

Enfin, la deuxième partie constitue les résultats obtenus et la comparaison entre les méthodes employées à fin de traiter l'effet de la méthode de réduction sur le modèle de prédiction, et ceci sous forme d'une analyse graphique approfondie.

Parmi ces techniques nous avons choisis d'utiliser l'analyse en composantes principales et l'analyse des moindres carrés partiels, ainsi qu'une troisième méthode nouvelle et immergent dans le domaine de la recherche de type modélisation et prédiction, et qui est les algorithmes génétiques ; pour réduire le nombre des descripteurs de notre base de données ; et comparer leur effet sur la qualité de prédiction de la perméabilité des polyimides aux gaz permanent N_2.

1ERE PARTIE : ELABORATION DE LA BASE DE DONNEES

III.1.Démarche à suivre

Comme cette modélisation repose sur l'utilisation des descripteurs moléculaires calculés par le logiciel Dragon Talete, nous avons suivi la démarche illustrée dans l'organigramme représenté dans la Figure III-1. Des traitements préliminaires pour la structure de chaque polymère sont réalisés par deux autres logiciels afin d'accélérer le travail et convertir l'extension du fichier compatible avec chaque logiciel.

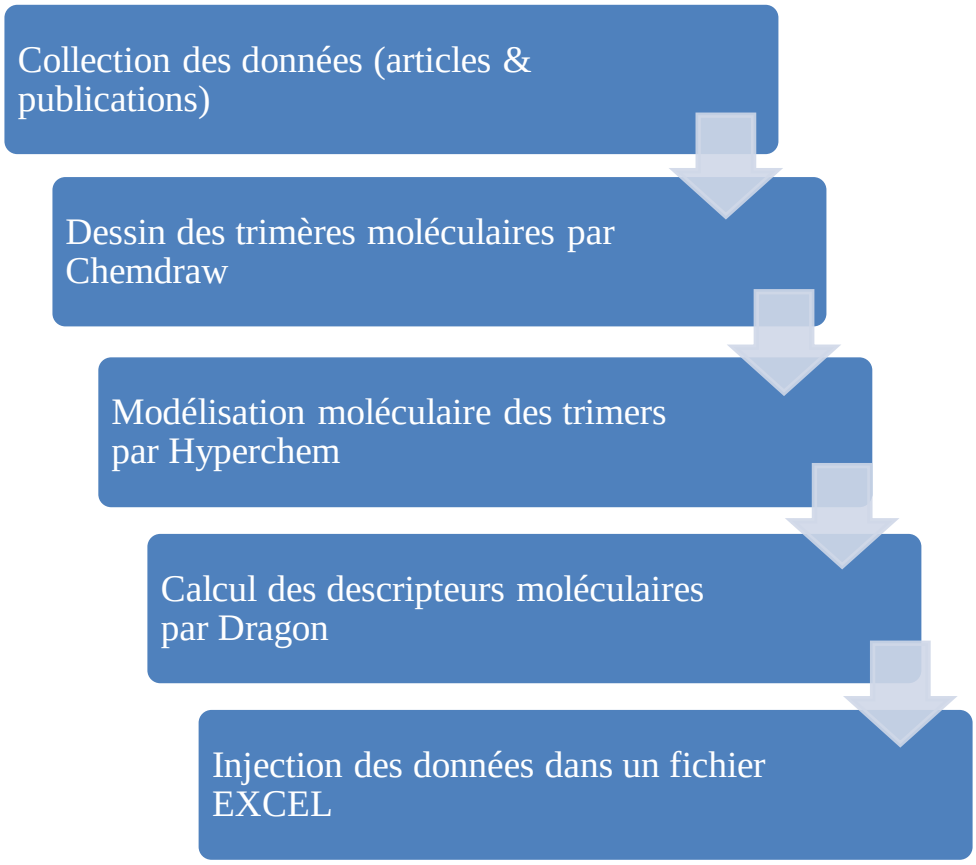


Figure III-1 **:** Organigramme des étapes suivies pour la préparation de la base de données

III.2.Les logiciels et Matériels utilisés :

Nous avons démarré les calculs avec une station de travail Samsung intel (R) core ™ i7 cpu ; 2.93 GHz de 8 Go RAM avec système d'exploitation 64 bits, mais il nous a fallu un deuxième PC plus puissant pour lancer plusieurs calculs simultanément et converger plus vite. Ainsi on s'est doté d'un nouveau ordinateur plus puissant hp probook ayant un microprocesseur Intel quatrième génération 450g1/i7-2.20ghz et de 16 Go de ram avec système d'exploitation Windows 8.0.1 original compatible avec la majorité des logiciels utilisés. Ces derniers sont présentés au fur et à mesure de leur emploi.

III.3.Collecte des données

Dans cette étape nous avons sélectionné 60 molécules de polymères vitreux de la famille des polyimides PI dont une seule molécule appartient aux polyetherimides PEI qui est (PDCDA_6F) ; à partir des différents articles et publications disponibles sur le Net [7] ; leur formules monomériques se trouvant sur le Tableau VII.1 de l'annexe. Les structures des polymères ont été collectées et restructurées selon [8] ainsi que leur perméabilités expérimentales respectives aux gaz permanent N_2 à Température correspondante [9].

III.4.Dessin des molécules sur ChemDraw

Pour réaliser cette tâche, nous avons choisi le logiciel Chem & Bio Draw Version 12.0 qui est un outil puissant et pas compliqué conçu par Cambridge Soft pour représenter une immense variété de molécules biologiques et/ou structures chimiques. Ce choix est dû à la facilité, ergonomie et convivialité de son interface.

Nous avons dessiné selon le type fils de fer (Wire) i.e que les liaisons entre les atomes qui sont des traits (Figure III-2) sous forme de trimères [10] pour les considérations suivantes :

-Un trimer est une représentation très simple d'un polymère et est valable pour son optimisation géométrique à fin d'évaluer un maximum d'interactions intramoléculaires.

-L'avantage de travailler avec un trimère réside dans la rapidité de l'optimisation de la structure et la facilité du calcul des descripteurs.

-Le segment de trimère qui représente le mieux la structure originale de polymère est celui du milieu (unité répétitive), tel qu'il est influencé par les caractéristiques physico-chimiques, électroniques et stériques des unités adjacentes et aussi préserve les caractéristiques structurales du polymère...

-Cette optimisation moléculaire n'a pas l'intention d'imiter la conformation 3D de la molécule de polymère, mais de considérer les interactions moléculaires entre les atomes des motifs répétitifs voisins à petite échelle.

-En outre, les valeurs des descripteurs proposés sont affectées par ces interactions [10].

Une foi la molécule dessinée elle est sauvegardée sous forme d'un fichier MDL Molfile (*.mol) pour l'étape suivante qui est la pré-optimisation par HyperChem

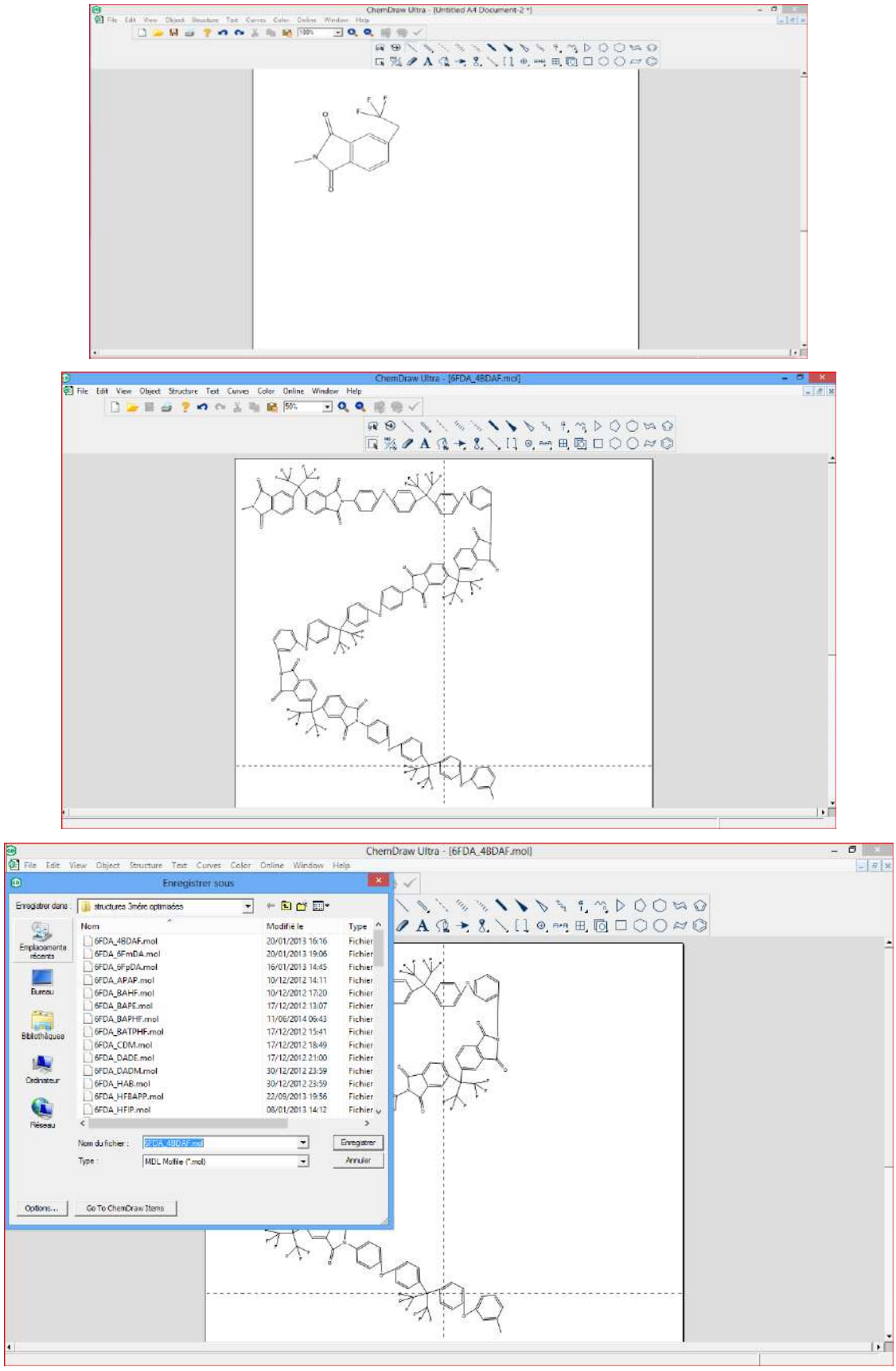


Figure III-2 : étapes de représentation des structures par ChemDraw

III.5.Modélisation moléculaire des trimères par HyperChem

L'HyperChem (version 8.0.7 molecular modeling system) (Figure III-3) est un logiciel très apprécié par les scientifiques à cause de la diversité des modules qu'il présente ainsi que sa puissance présentée à travers les diverses méthodes de modélisation et d'optimisation géométriques qu'il procure [11] ; en suivant les étapes :

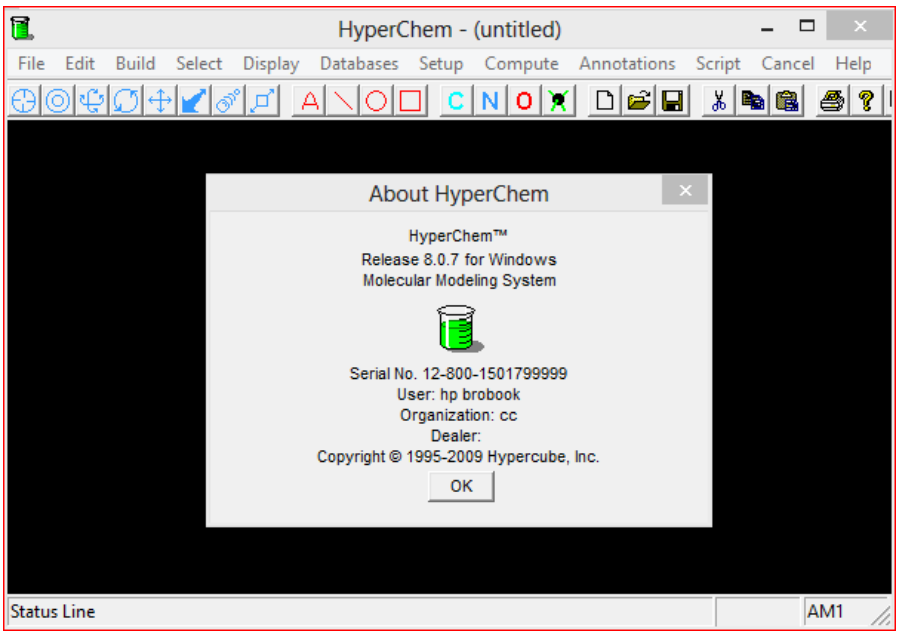


Figure III-3 : HyperChem version 8.0.7 pour la modélisation moléculaire des trimères

En ouvrant l'interface d'Hyperchem on extrait le dossier qui contient les fichiers d'extension (*.mol) déjà calculé par ChemDraw puis on réalise pré-optimisation des structures (Setup/Molecular Mechanics/MM+/ Compute/Geometry Optimization/Gradient conjugate polak Ribière).La convergence des calculs puis on sauvegarde la structure 3D optimisée. L'optimisation de notre structure par la procédure de la mécanique moléculaire de champs de force (MM+ : Force Field Molecular). Puis on refait pour la même structure une deuxième optimisation avec la méthode des orbitales moléculaires semi empirique AM1 (Austin Model1) avec l'algorithme de Polak-Ribière d'une valeur limite du gradient= 0.01kcal^{-1}mol^{-1} pour la raffiner [10]. Cette étape peut être très longue, elle prend des temps considérables parfois plus de 12 heures, mais indispensable au calcul des descripteurs.

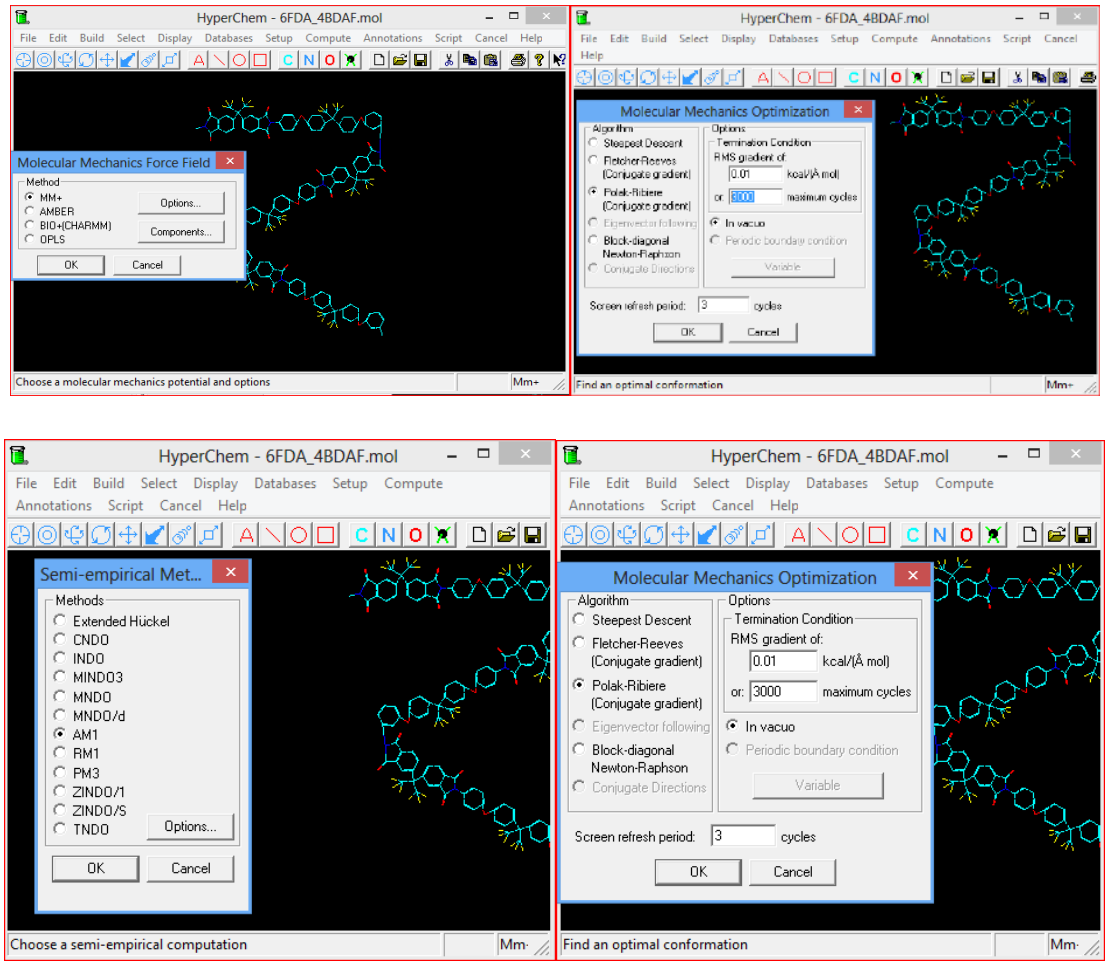


Figure III-4 : Etapes des pré-optimisations et d'optimisation des molécules respectivement

Dans notre travail ; la modélisation moléculaire a été employée à fin d'obtenir la conformation énergétique la plus stable -celle dont l'énergie est la plus petite possible [12]- et qui selon [10] représente la géométrie adoptée par le polymère selon les forces intramoléculaires.

Une fois la convergence atteinte, on enregistre le résultat sous forme d'un fichier d'extension (*.HIN) à fin de le rendre compatible au calcul des descripteurs par le logiciel Dragon.

III.6. Calcul des descripteurs par DRAGON

Une fois que nos structures ont été modélisées, on fait calculer les descripteurs une par une dans le logiciel de calcul des descripteurs moléculaires le plus employé dans la majorité des recherches de type modélisation structurale qui est DRAGON Talete version 5.5 année 2007 (Figure III-5)



Figure III-5 : L'interface du logiciel DRAGON TALETE version 5.5

Pour calculer les descripteurs, on ouvre les fichiers *.HIN, car il ne lie pas les fichiers d'extension *.Mole et on choisit tous les 22 blocs de descripteurs disponibles dont les deux derniers font la nouveauté de cette version pour toutes les structures (0D ,1D ,2D ,3D) puis une foi le calcul terminé enregistrez les fichiers obtenus sous forme de fichiers texte (.*txt) prêts pour la dernière étape (Figure III-6).

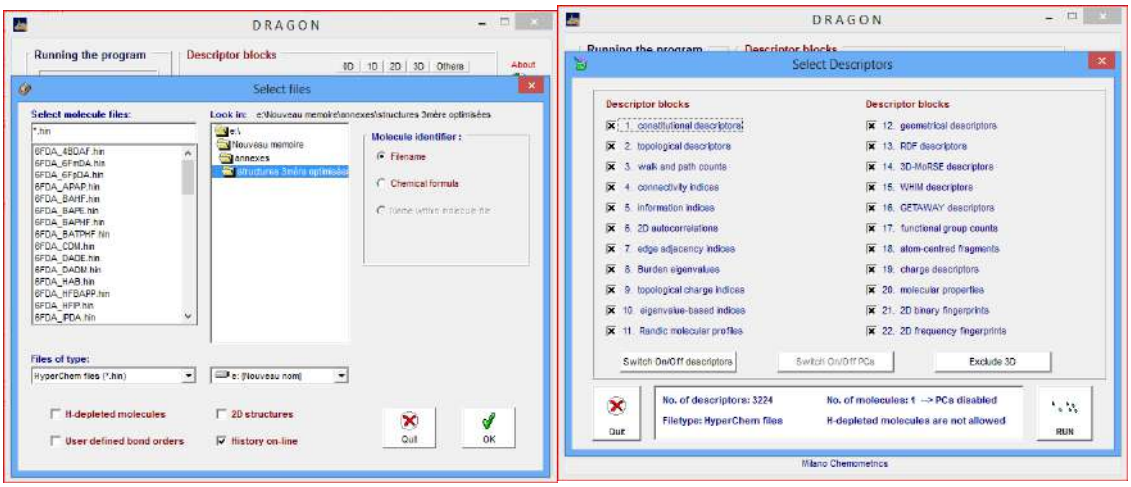
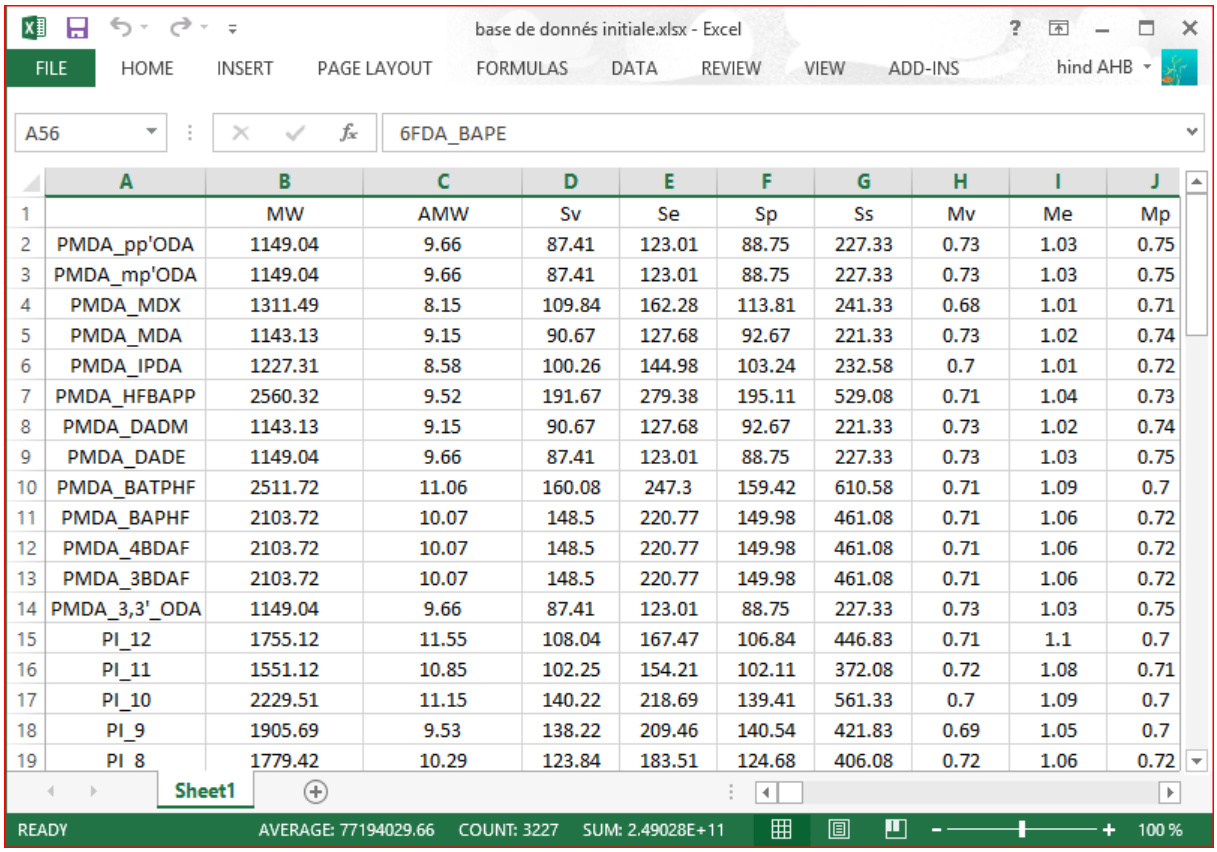


Figure III-6 : étapes de calcul des descripteurs par DRAGON TALETE

Nous avons calculé tous les descripteurs parce que dans notre étude nous avons besoin du maximum d'information structurale pour pouvoir prédire la relation quantité structure des polymères et leur perméabilité aux gaz N_2 (QSPR).

III.7. Collection des données dans un fichier EXCEL

Nos descripteurs sont calculés dans des fichiers texte contenant les noms de chaque descripteur et la valeur attribuée ainsi que le nom du polymère et le nombre des descripteurs qui a atteint dans notre étude 3224 descripteurs. On copie les noms des descripteurs et on les fait rentré dans un fichier EXCEL dans la 1ere ligne, puis dessous on réinjecte soigneusement et sans faute les valeurs numériques correspondants (Figure III-7).



base de donnés initiale.xlsx - Excel

FILE HOME INSERT PAGE LAYOUT FORMULAS DATA REVIEW VIEW ADD-INS hind AHB

A56 | 6FDA_BAPE

	A	B	C	D	E	F	G	H	I	J
1		MW	AMW	Sv	Se	Sp	Ss	Mv	Me	Mp
2	PMDA_pp'ODA	1149.04	9.66	87.41	123.01	88.75	227.33	0.73	1.03	0.75
3	PMDA_mp'ODA	1149.04	9.66	87.41	123.01	88.75	227.33	0.73	1.03	0.75
4	PMDA_MDX	1311.49	8.15	109.84	162.28	113.81	241.33	0.68	1.01	0.71
5	PMDA_MDA	1143.13	9.15	90.67	127.68	92.67	221.33	0.73	1.02	0.74
6	PMDA_IPDA	1227.31	8.58	100.26	144.98	103.24	232.58	0.7	1.01	0.72
7	PMDA_HFBAPP	2560.32	9.52	191.67	279.38	195.11	529.08	0.71	1.04	0.73
8	PMDA_DADM	1143.13	9.15	90.67	127.68	92.67	221.33	0.73	1.02	0.74
9	PMDA_DADE	1149.04	9.66	87.41	123.01	88.75	227.33	0.73	1.03	0.75
10	PMDA_BATPHF	2511.72	11.06	160.08	247.3	159.42	610.58	0.71	1.09	0.7
11	PMDA_BAPHF	2103.72	10.07	148.5	220.77	149.98	461.08	0.71	1.06	0.72
12	PMDA_4BDAF	2103.72	10.07	148.5	220.77	149.98	461.08	0.71	1.06	0.72
13	PMDA_3BDAF	2103.72	10.07	148.5	220.77	149.98	461.08	0.71	1.06	0.72
14	PMDA_3,3'_ODA	1149.04	9.66	87.41	123.01	88.75	227.33	0.73	1.03	0.75
15	PI_12	1755.12	11.55	108.04	167.47	106.84	446.83	0.71	1.1	0.7
16	PI_11	1551.12	10.85	102.25	154.21	102.11	372.08	0.72	1.08	0.71
17	PI_10	2229.51	11.15	140.22	218.69	139.41	561.33	0.7	1.09	0.7
18	PI_9	1905.69	9.53	138.22	209.46	140.54	421.83	0.69	1.05	0.7
19	PI_8	1779.42	10.29	123.84	183.51	124.68	406.08	0.72	1.06	0.72

Sheet1

READY AVERAGE: 77194029.66 COUNT: 3227 SUM: 2.49028E+11 100 %

Figure III-7 : La base de données des descripteurs finale

III.8.Réduction du nombre de variables

Après optimisation des structures tridimensionnelles, les descripteurs sont déterminés dont certains peuvent être corrélés à la propriété qui nous intéresse qui est la perméabilité (P) selon la relation suivante :

$$\boldsymbol{P = P_0 + \alpha_1 D_1 + \alpha_2 D_2 + \alpha_3 D_3 + \cdots} \qquad \text{III-1}$$

Où P est la propriété étudiée (variable dépendante), D_i est le descripteur i (variable indépendante) et α_i mesure le poids de chaque descripteur qui entre dans la relation [3].

Notre base de données qu'on appellera « base de données initiale » est un tableau de 60*3225 descripteurs, car on a pris en considération le descripteur T° température qui influence la perméabilité dont les valeurs occupent la dernière colonne du tableau. Or puisque nous avons des valeurs redondantes nous avons procéder à une élimination préliminaire des valeurs insignifiantes selon [13] comme suit [14]:

- les descripteurs invariants ont été supprimés manuellement.

- tout descripteur ayant plus de 75% de valeurs identiques est supprimé

-tout descripteur ayant un RSD (Relative standard déviation) <0.05 est supprimé [15].

D'où une pré-réduction de plus de 60% des descripteurs peux ou non signifiants. Cette base de données appelée « nouvelle base de données » est devenu 60*1184 va subir une autre réduction par les méthodes proposées qui sont :

- la PCA ;
- la PLS avec ses quatre variantes : PLS analyse discriminante ainsi que la PLS avec élimination de variable de Monte-carlo MCVEPLS, PLS avec validation croisée et PLS sans validation croisée ; qui ont le même principe de base de la régression PLS, mais qui diffèrent dans les critères de sélection des variable [16];
- l'AG.

Nous avons appliqué toutes les méthodes à travers l'interface du logiciel de calcul mathématique Matlab R2012b version 8.0.0.783, 64 bit. Les algorithmes de la PLS et ses variantes ainsi que GA-PLS sont implémentés dans la bibliothèque du logiciel selon le lien (http://www.mathworks.com). Les résultats sont présentés en deux formes :

- Résultats sous forme des tableaux.
- Résultats sous forme des figures.

2EME PARTIE :

ANALYSE & DISCUSSION

III.9. Application de L'analyse en composantes principales

Signalons que notre étude de l'ACP n'est pas en tant que méthode exploratoire, mais en tant qu'outil de réduction de notre base de données.

La méthode PCA calcul le nombre de composantes principales qui représentes le maximum de variance entre les variables et dans notre cas les deux facteurs F1 et F2 représente plus de 55% de variance, mais en fait le choix d'une variance de plus de 80% nécessite le choix de 6 facteurs principaux selon la première cassure représentée dans le graphe de la Figure III-8.

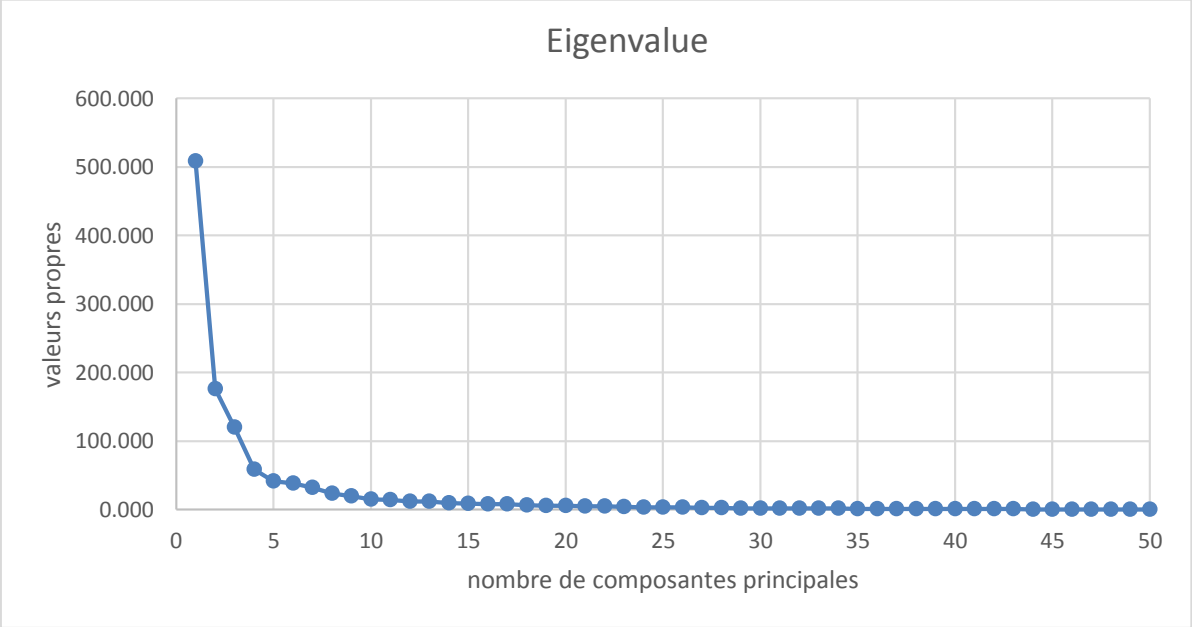


Figure III-8 : représentation des valeurs propres en fonctions du nombre de composantes principales

La projection des variables (descripteurs) sur les nouveaux axes (F1 et F2) indépendants nous permet de choisir les variables les moins corrélées situées dans la périphérique du cercle (rayon=1). Cependant, en raison du nombre important des variables représentées le graphe est devenu inexploitable (Figure III-9).ce qui nous a poussé à résoudre le problème analytiquement à travers un programme Matlab qui consiste à faire entrer notre base de donnée finale de 60*1184 dans le programme qui a calculé la matrice de corrélation 1184*1184 .Tout descripteur ayant un coefficient de corrélation linéaire de Bravais-Pearson plus de 0.99 a été supprimé.

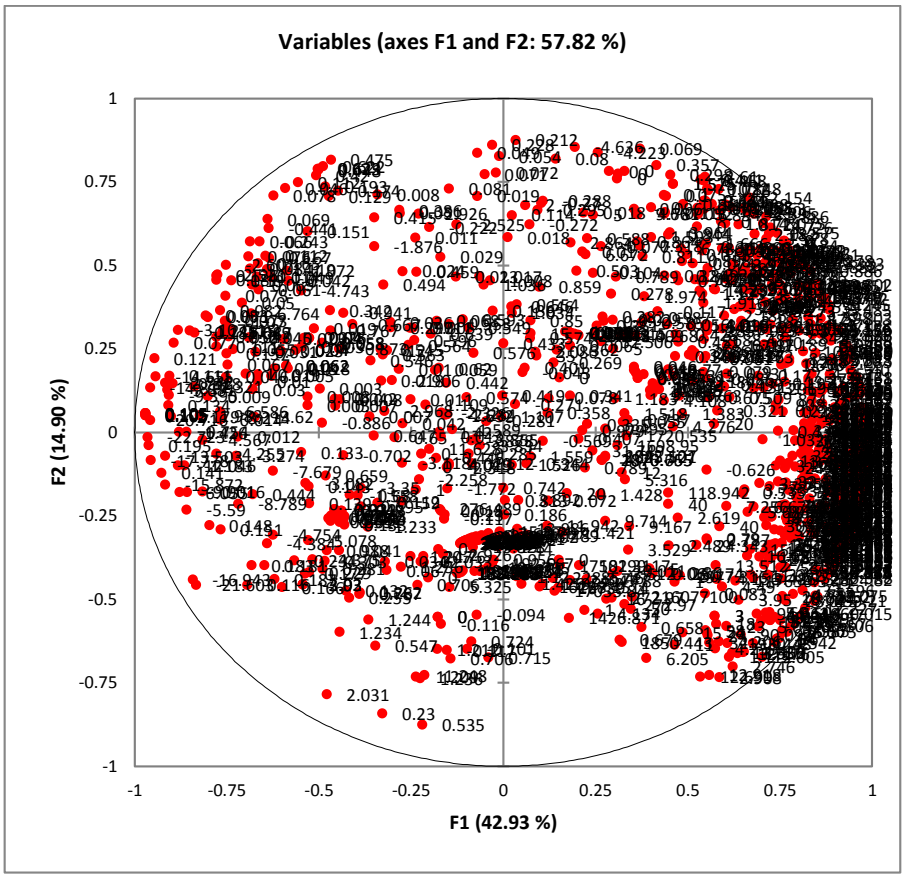


Figure III-9 : Représentation des variables en fonction des premiers principaux facteurs

La matrice des variables réduites retenues est constituée de 81 descripteurs qui sont :

Descripteurs 0D: ARR

Descripteurs2D:
HNar/MAXDN/Lop/X3Av/TIC1/ATS6e/MATS2m/MATS7e/GATS7m/GATS8m/GATS1v/ GATS7v/GATS1e/GATS5e/GGI5/GGI9/JGI3/VEA2/ F03[C-O]/

Descripteurs 3D:
SP01/ADDD/SPAM/DISPv/RDF035m/RDF055m/RDF080m/RDF100m/RDF145m/RDF150 m/RDF050e/RDF040p/Mor04u/Mor22u/Mor26u/Mor05m/Mor10m/Mor14m/Mor25m/Mor26 m/Mor30m/Mor06p/G1u/E3m/E2v/G2p/Gm/Gs/Dp/Ds/H3u/H8u/HATS0u/H0m/H0v/H7v/H Tv/HATS4v/H4e/HATS1e/HATS5e/H1p/H3p/H7p/HATSp/R4u/R4m/R6m/R8m/R2m+/R3v/ RTv/R6e/RTe/R4e+/R7e+/R8e+/R3p/R2p+/

Descripteurs différents : MLOGP2 qui appartient au bloc des propriétés moléculaires ainsi que la T°(C°).

On constate qu'aucun descripteur 1D n'est présent dans la base réduite par PCA.

III.10. Application de la méthode des moindres carrées partielles PLS

Selon le schéma des différentes variantes de la PLS ; nous avons choisis d'enrichir notre étude comparatrice pour mieux comprendre l'effet de chaque méthodes sur la réduction du nombre de descripteurs.

Les différentes méthodes issues de la PLS sont classées dans la Figure III-10 selon leur principe et plus de détails sur les différentes variantes de choix des variables sont données dans la Figure III-11. Celles coloriées en jaune ont étés appliquées à notre base finale de données.

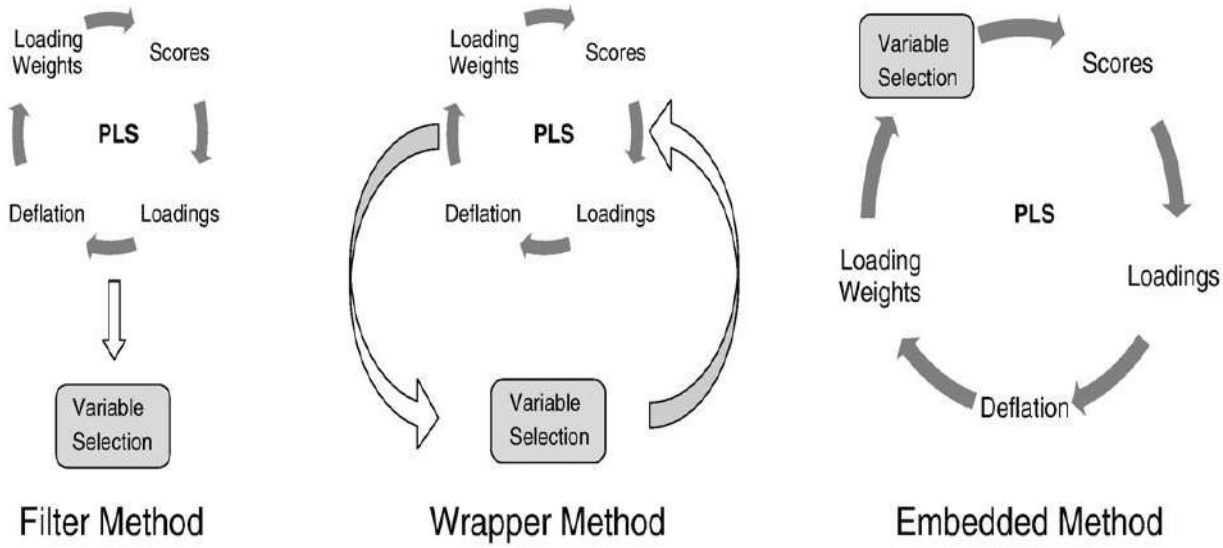


Figure III-10 : illustration des méthodes de PLS : Filtrantes, Enveloppantes et intégrés

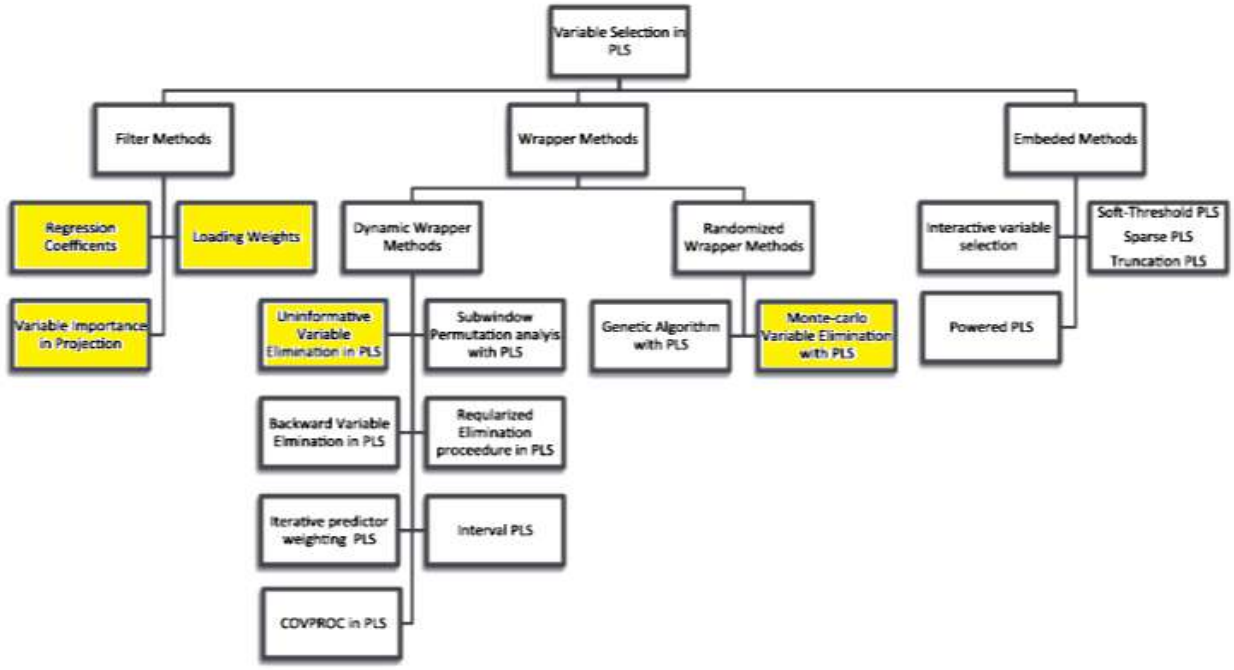


Figure III-11 : Différentes variantes de PLS pour la sélection des variables

On a fait entrer notre base de donnée finale de 60*1185 en tenant compte de la colonne de la perméabilité du gaz N_2dans le programme de la PLS qui a calculé la régression de la sortie -Perméabilité- sur les entrés –descripteurs- selon les différentes variantes suivantes :

III.10.1. Selon la méthode de PLS avec validation croisée

La réduction par PLS est sensible aux nombre des facteurs principaux imposés. Une fois que le changement de la variance exprimé par le Y prédit ne change pas plus de 2% en fonction du nombre de composantes principales on fixe le nombre de facteurs principales (Figure III-12).

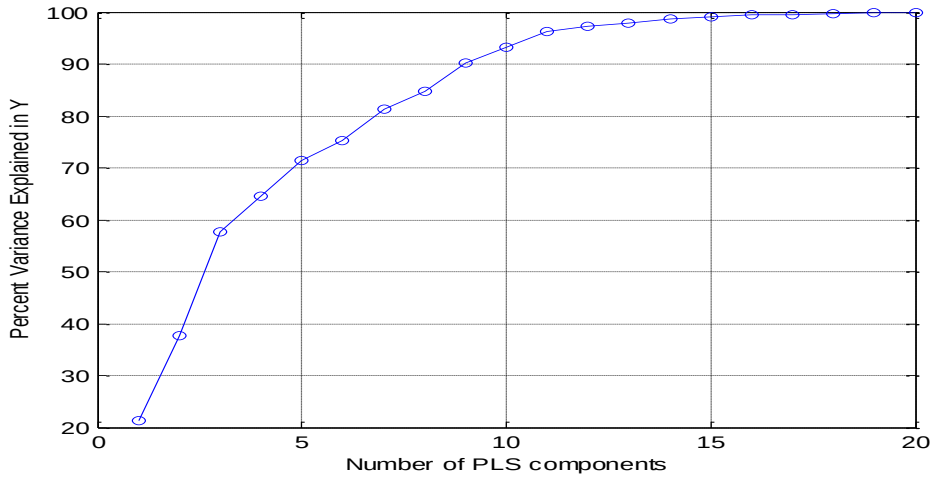


Figure III-12 : effet du nombre de composantes principales sur la variance exprimé des Y

Le choix des variables à éliminer se fait soit sur la base du coefficient de régression Beta ou le poids de chaque descripteur.

Pour le premier cas, toute variable ayant un Beta en valeur absolue située dans la bonde de la limite « cutoff » (voir Figure III-13) est supprimée et considérée comme bruit.

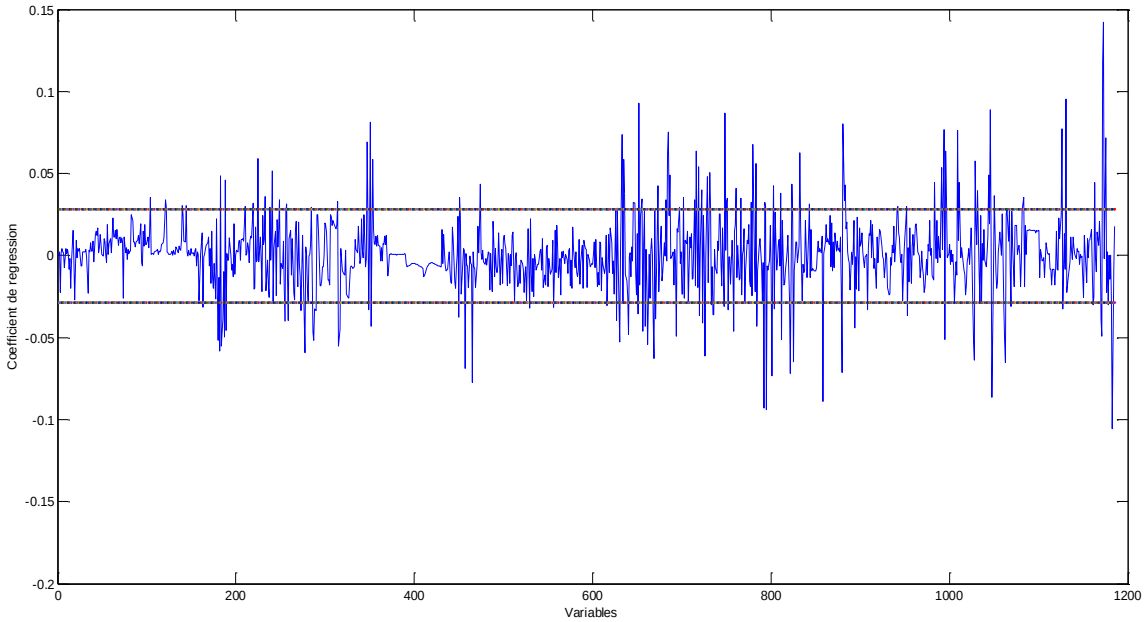


Figure III-13 : choix des variables selon le coefficient de régression

Dans le deuxième cas, la réduction de fait en se basant sur La réduction se fait en se basant sur le poids (W) de chaque descripteur en choisissant un seuil (Figure III-14) qui limite le nombre de variables à une vingtaine comme point de départ pour le bon fonctionnement de la partie prédiction par les réseaux de neurones.

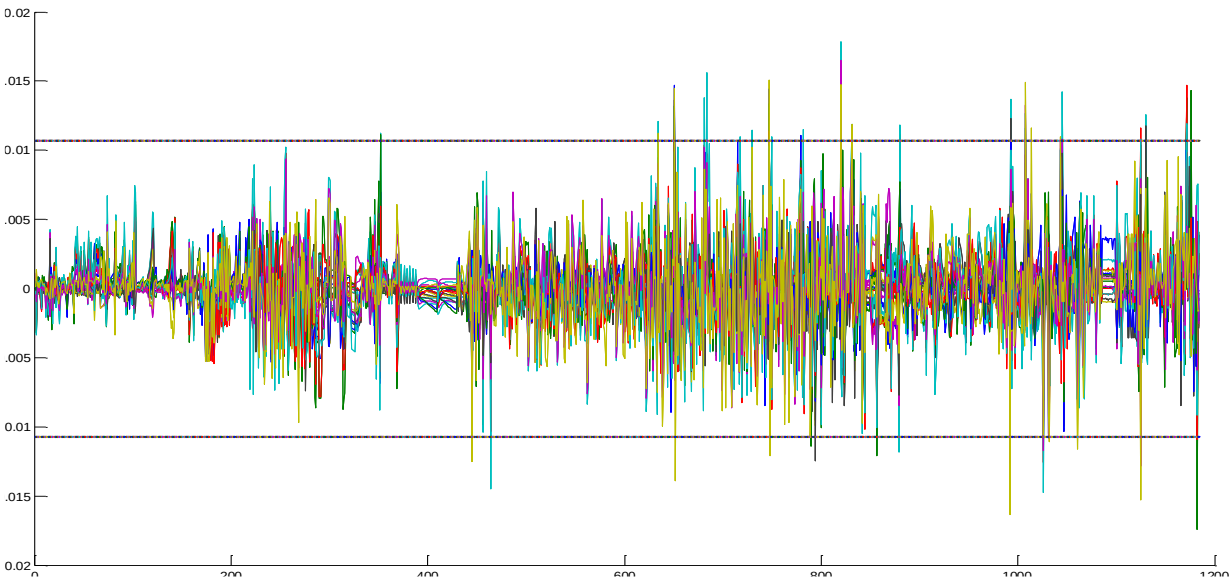


Figure III-14 : choix des variables selon le poids des descripteurs

Cette méthode a réduit le nombre des descripteurs jusqu'à 23 descripteurs qui sont :

-Descripteurs 2D :'MATS5p"B07 [O-O]"F05[C-N]"F09[C-N]"F09[O-O]"F10[N-O]'

-Descripteurs3D :'DISPp"Mor29u"Mor22v"Mor22e"Mor22p"G2v"G2e"HGM"R4u+"R2m+' 'R8m+' 'R3v+"R8v+"R2e+"R7e+' 'R8e+"R8p+'

On remarque l'élimination des descripteurs de 0D et 1D et les descripteurs liés à la charge et aux propriétés moléculaires ainsi que la température par cette PLSCV.

III.10.2. Selon la méthode de PLS sans validation croisée

Cette méthode a réduit le nombre des descripteurs jusqu'à 37 descripteurs qui sont :

-Descripteurs 2D : 'JGI10"PJI3'

-Descripteurs3D:'DISPp"Mor11u"Mor28u"Mor29u"Mor26m"Mor29m"Mor31v"Mor11e' 'Mor28e"Mor29e"Mor31e"Mor28p"Mor31p"G2u"E3u"G2e"G3p"Gu"ISH"HGM"R4u+"R5u+" R2m+"R8m+"R2v+"R7v+"R8v+"R2e+"R4e+"R2p+"B07[O-O]"B09[N-N]"F09[C-O]"F09[O-O]' 'F10[N-O]'

On remarque l'élimination des descripteurs de 0D et 1D et les descripteurs liés à la charge et aux propriétés moléculaires ainsi que la température par cette PLSsans CV comme la précédente.

III.10.3. Selon la méthode de PLS d'élimination de la variable non informative (PLSUVE)

Cette méthode a réduit le nombre des descripteurs jusqu'à 21 descripteurs qui sont :

-Descripteurs 2D : 'F07[O-F]"F08[C-C]"F08[C-N]"F08[C-O]"F08[C-F]"F08[O-F]' 'F09[C-C]' 'F09[C-N]"F09[C-O]' 'F09[C-F]"F09[N-N]"F09[N-O]"F09[O-O]"F10[C-C]"F10[C-N]"F10[C-O]"F10[C-F]' 'F10[N-N]"F10[N-O]"F10[O-O]'

-Autre descripteurs :'T° (C°)'

On remarque l'élimination des descripteurs de 0D, 1D, 3D et les descripteurs liés à la charge et aux propriétés moléculaires ainsi que la température par cette PLSUVE.

III.10.4. Selon la méthode de PLS analyse discriminante

-Descripteurs 2D : 'F08[C-C]"F08[C-N]"F08[C-O]' 'F08[C-F]"F08[O-F]"F09[C-C]"F09[C-N]"F09[C-O]' 'F09[C-F]"F09[N-N]"F09[N-O]"F09[O-O]"F10[C-C]"F10[C-N]"F10[C-O]' 'F10[C-F]"F10[N-N]"F10[N-O]"F10[O-O]'

-Autres descripteurs : 'T° (C°)'

On remarque l'élimination des descripteurs de 0D, 1D, 3D et les descripteurs liés à la charge et aux propriétés moléculaires ainsi que la température par cette PLS analyse discriminante. Elle est pratiquement identique à la précédente dans ses résultats.

III.11. Application de la méthode des Algorithme génétiques

Nous avons entré la base complète de 60*1185dansun algorithme génétique à base de PLS en prenant la valeur du coefficient de validation croisé LOO (leave one out) comme fonction d'objectif (fitness function) [17] avec les caractéristiques suivants : volume de population égale à 30 chromosomes; group délétion est de 5 ; nombre moyen des variables sélectionnés dans le chromosome de la population du départ est 3 ; pas de jumeaux. La représentation de la racine carrée de l'erreur relative moyenne en prédiction donne le graphe suivant qui explique que lorsque -ce RMSEP devient constant on prend les descripteurs qui sont avant cette constante Figure III-15.

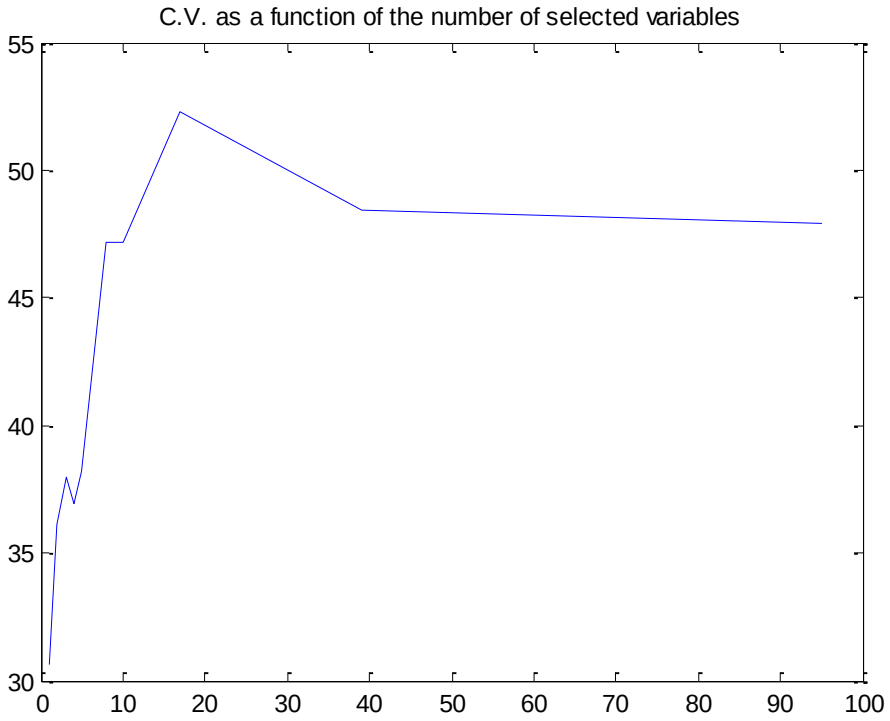


Figure III-15 : représentation du RMSEP en fonction du nombre de variables sélectionnées

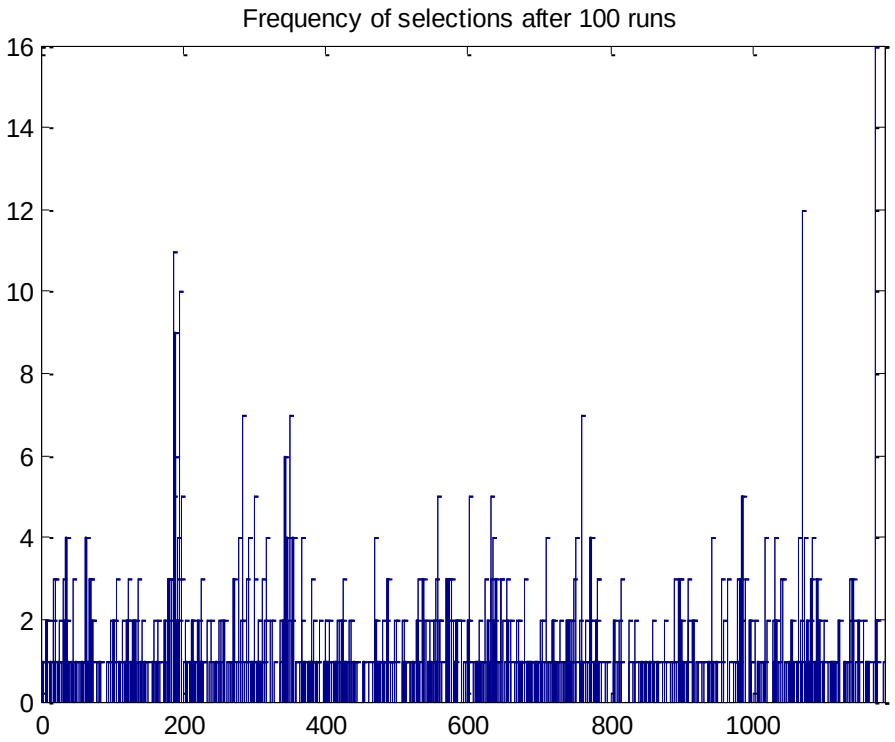


Figure III-16 : représentation des fréquences de sélection après 100 itérations

Les descripteurs retenus par cet algorithme sont 17 :

-Descripteurs 2D: SIC3/CIC3/BIC3/ IC4/SIC5/BIC5/EEig11x/EEig15d/JGI1/JGI8/F09[C-N]

-Descripteurs 3D :RDF130v/RDF045p/Mor09u/Mor09p/R6u/

-Descripteurs 1D : nCp

On remarque l'absence des descripteurs de 0D et les descripteurs liés à la charge et aux propriétés moléculaires ainsi que la température.

III.12. Conclusion

En comparant les résultats de toutes ces méthodes : de la PCA, PLS, et AG on a remarqué qu'ils ne réduisent pas de la même manière et pas toujours les même descripteurs. Cependant les descripteurs 3D et 2D sont toujours présent dans la majorité des cas.

IV. Application des descripteurs retenus à la prédiction de la perméabilité

Nous avons utilisé un réseau de neurones MLP (multi layer perceptron) [17 1] avec une fonction de transfert [logsig purline] pour choisir la méthode de réduction des descripteurs optimaux dans la prédiction de la perméabilité du N_2.Ce choix s'effectue selon l'erreur relative absolue moyen la plus basse sachant que :

$$ERAM = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_{exp}-y_{cal}}{y_{exp}}\right) \quad \text{IV-1}$$

Ou y_{exp} la valeur expérimental, y_{cal} la valeur calculé. Et le R est le coefficient de régression du modèle neuronale obtenu par la fonction « postreg » sur Matlab défini par

$$R = 1 - \frac{\sum_1^n (y_i - y_i^t)^t}{\sum_1^n (y_i - y_0^t)^2} \quad \text{IV-2}$$

Avec

$$y_0 = \left[\frac{1}{n}\sum (y_i - y_i^t)^2\right]^{1/2} \quad \text{IV-3}$$

Où $y_i$la valeur expérimental, y_i^t la valeur prédite. Et tant que R tend vers 1 mieux le modèle est bien.

Le tableau suivant donne les ERAM et R de chaque méthode.

Tableau IV-1 : résultats des ERAM et R du modèle neuronale pour chaque méthode

Méthode	PCA	PLSsansCV	PLS avec cv	PLSUVE	PLSDisc	AGPLS
ERAM	0.52	1.3	0.56	0.16	0.08	0.03
R	0.9	0.94	0.86	0.83	0.9	0.92

On remarque que la plus basse valeur de ERAM correspond à celui de la méthode d'algorithme génétique avec PLS ainsi que le R correspondant est le plus près de 1.

En comparant les graphes représentant l'Y_{predi} en fonction du Y_{exp} pour chaque méthode, en remarque que la meilleure méthode de réduction des variables entre PCA et PLS et AG -où le principe des méthode est complètement différent- est celle des Algorithmes génétiques avec une très belle courbe ajustée et avec un R de 0.92 très proche du 1.

Mais entre les 4 variantes de la PLS on trouve que la PLS sans cross-validation est la meilleure avec un R de 0.94 puis la PLS discriminante avec un R de 0.9 puis la PLS avec cross-validation avec R égale à 0.86 et enfin la PLS Uninformative variable est la mauvaise parmi toutes avec un R de 0.83 ; mais toutes les 4 méthodes ne dérivent pas trop du modèle parfait.

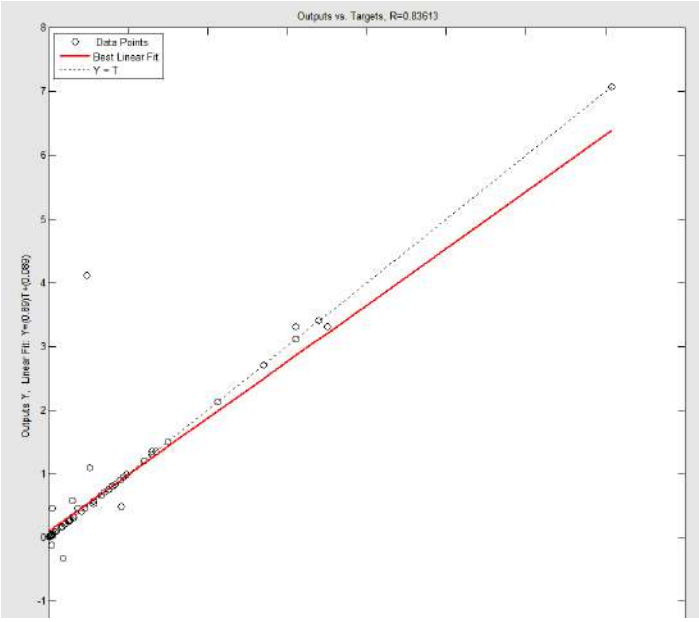


Figure IV-1 : représentation de l'écart entre le modèle prédit et l'expérimental de la PLSUVE

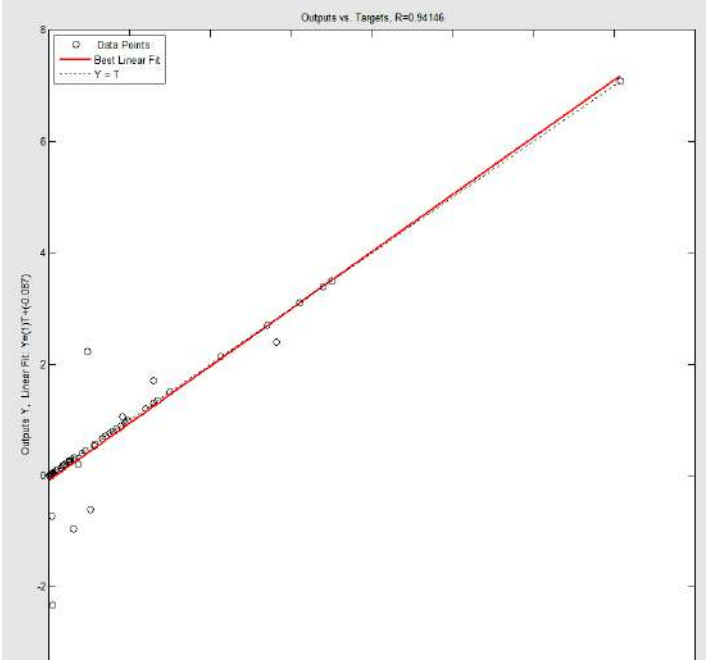


Figure IV-2 : représentation de l'écart entre le modèle prédit et l'expérimental de la PLS sans CV

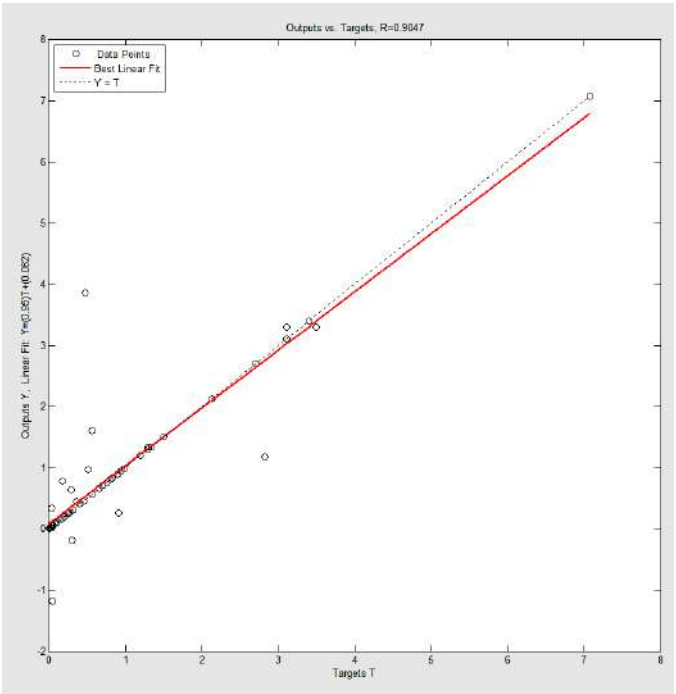


Figure IV-3 : représentation de l'écart entre le modèle prédit et l'expérimental de la PLSdisc

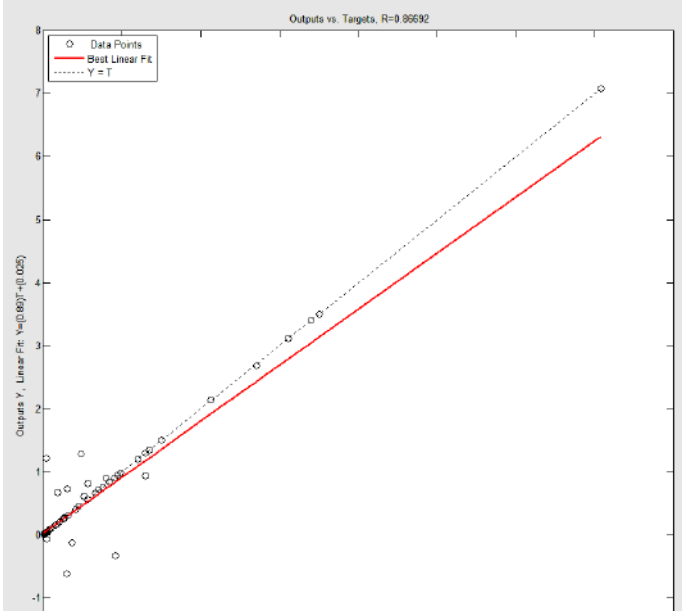


Figure IV-4 : représentation de l'écart entre le modèle prédit et l'expérimental de la PLSCV

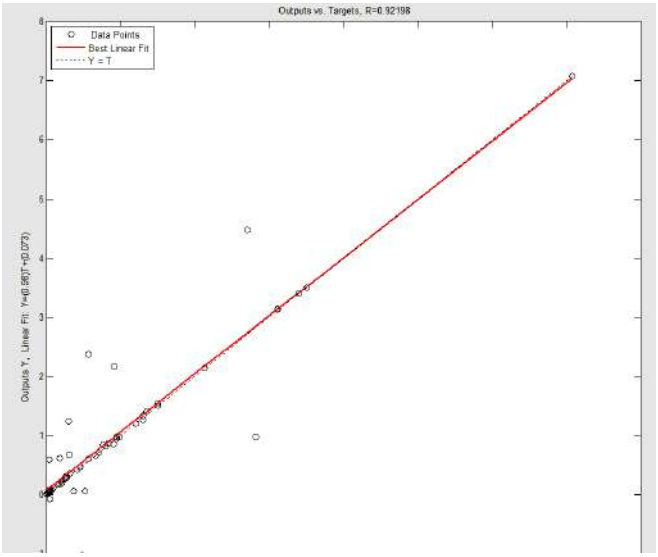


Figure IV-5 : représentation de l'écart entre le modèle prédit et l'expérimental de la PLSAG

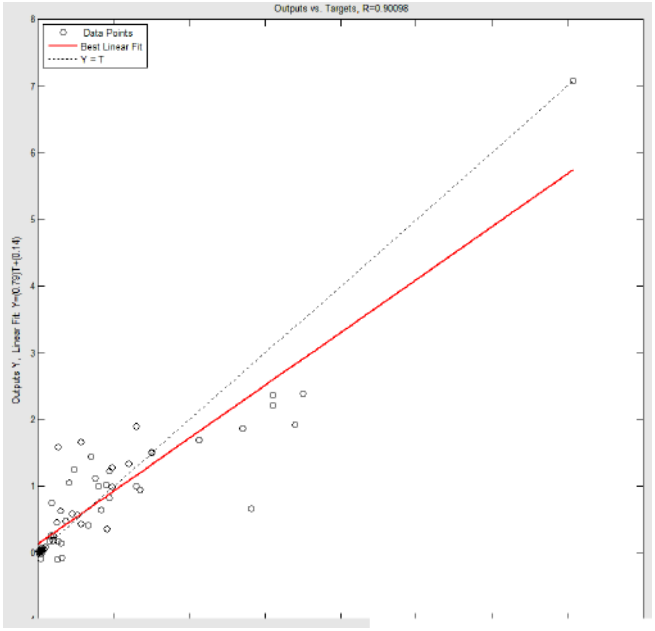


Figure IV-6 : représentation de l'écart entre le modèle prédit et l'expérimental de la PCA

Conclusion Générale

V. Conclusion générale

Lors de cette étude nous avons montré qu'il existe une influence de la méthode choisie pour la réduction des descripteurs moléculaires des polyimides.

Nous avons choisi trois méthodes d'un principe complètement différent et qui correspondent à notre cas de prédiction qui est un nombre d'individus (polymères) très petit par rapport aux nombre d'observations (descripteurs moléculaires) ; qu'on nomme PCA ; PLS ;AG.

Les deux premières méthodes sont anciennes et différentes car la PCA ne prend pas en considération l'influence de la sortie (perméabilité) par contre la PLS intègre la sortie dans toutes ses étapes de calculs ce qui fait sa puissance.

La troisième méthode est complètement innovatrice et très prometteuse, car inspirée du cerveau humain donne des résultats intéressants.

Nous avons remarqué que pour la réduction de notre base de donné l'algorithme génétique est le plus rapide en terme de convergence par rapport aux deux autre méthodes, avec un R de 0.92 et un ERAM presque nul de 0.03.

Par contre entre les différentes variantes de PLS c'est la PLS sans cross-validation qui est la mieux adaptée avec un R de 0.94 et ERAM 1.3. Mais toutes les méthodes ne dérivent pas énormément du modèle parfait et ceci est justifiable, car l'étude est faite sur une seule et même famille de polymères qui est les polyimides, ce qui serait très intéressant si l'étude explore différentes familles de polymères.

VI. Références Bibliographiques

[1] Hassnaoui, H.; « Prédiction de la perméabilité des polymères aux gaz par les réseaux de neurones », Médéa, 2011.

[2] Felder, R.M., «Estimation of Gas Transport Coefficients from Differential Permeation, Integral Permeation, and Sorption Rate Data", Journal of Membrane Science, V. 3, n°1, 1978, p15-27.

[3] Julie Anthoni, Synthèse enzymatique, modélisation moléculaire et caractérisation d'oligomères de flavonoïdes, INPT de Lorraine, soutenue 2007.

[4] Todeschini et Consonni, version 5.5, 2007

[5] http://www.Cours-ACP.pdf

[6] Méthodes et applications sur la Sélection de caractéristiques par Hassan CHOUAIB 2011

[7] A Yu.Alentiev, K.A.Loza, Development of the methods for prediction of gas parameters of glassy polymers: polyimides as alternating co-polymers; journal of Membrane Science 167 (2000) 91-106.

[8] Stern*, S. Alexander; Polymers for gas separations: the next decade; Journal of Membrane Science, Elsevier Science B.V., Amsterdam; 1994; l-65.

[9] Martha L. Ceocopieri-Gomez et al, On the limits of gas separation in CO_2/CH_4, N_2/CH_4 and CO_2/N_2 binary mixtures using polyimide membranes, Februry 2007, journal of membrane Science 293 (2007) 53-65.

[10] D. Palomba, G.E. Vazquez, M.F. Diaz; Novel Descriptors from Main and Side Chains of high-molecular-weight Polymers applied to Prediction of Glass Transition Temperatures; Journal of Molecular Graphics and Modelling; 2012; 1-53.

[11] valeriu bivol; Modelling of the 3d-structure of cam:oma photopolymers by using of computational chemistry program,moldova republic, received 2004

[12] Hermann Dugas, chapitre 3, introduction aux méthodes de minimisation d'énergie ; modelisation.

[13] Konovalov,D .A. ,Sim,N.,Deconinck, E., Heyden,Y.V.,and Coomans,D.(2008). Statistical confidence for variable selection in QSAR models.J.Chem. inf .Model .48, 370.

[14] Srinivas Reddy, A.,Kumar, S., and Garg, R. (2010). Hybrid-genetic algoritnm based descriptor optimization and QSAR models for predicting the biological activity of tipranavir analogs for HIV protease inhibition. J. Mol. Graph. Model.28, 852.

[15] Konovalov, D.A.,Sim,N., Deconinck, E.,Heyden,Y.V., and Coomans, D. (2008). Statistical confidence for variable selection in QSAR models. J. Chem. Inf. Model. 48,370.

[16] Tahir Mehmood, A review of variable selection methods in Partial Least Squares Regression, Biostatistics, Department of Chemistry, Biotechnology and Food Sciences, Norwegian University of Life Sciences, Norway the advances in technology. Chemometrics and Intelligent Laboratory Systems 118 (2012) 62–692012

[17] Eslam Pourbasheer et al, QSAR stydy of ACK1 inhibitor by genetic algorithm-multiple linear regression (GA-MLR); accepted manuscript Athens, Greece.

NOMENC LATURE

&

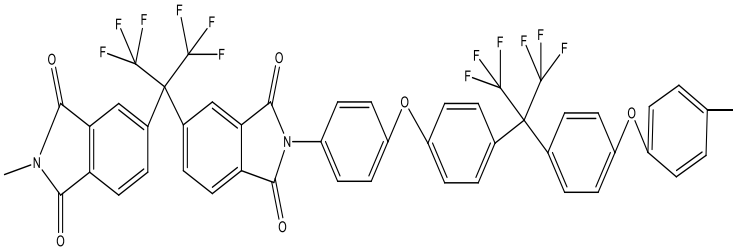
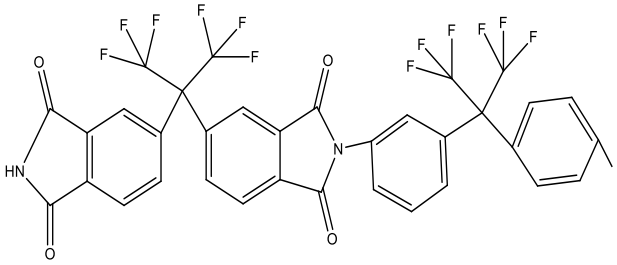
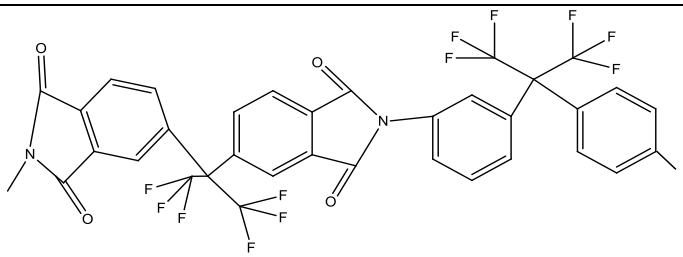
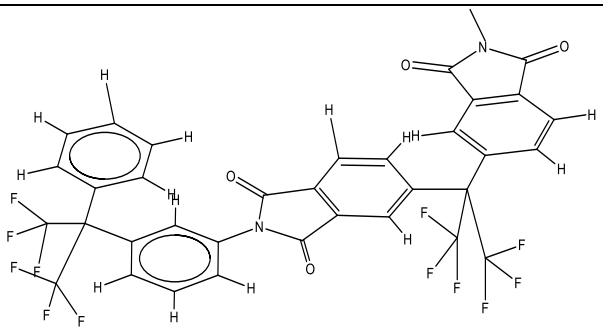
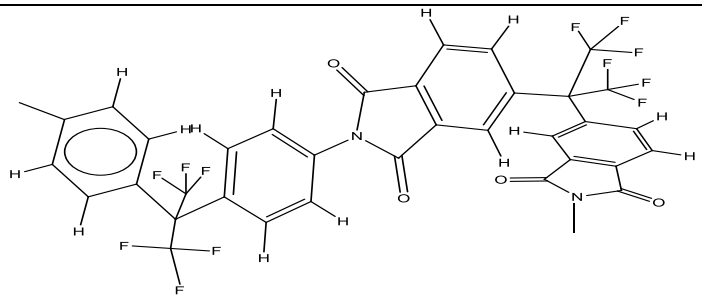
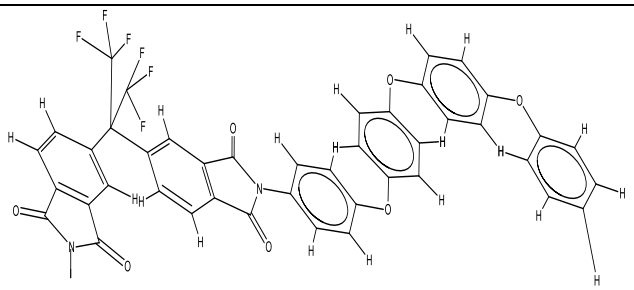
SYMBOLE

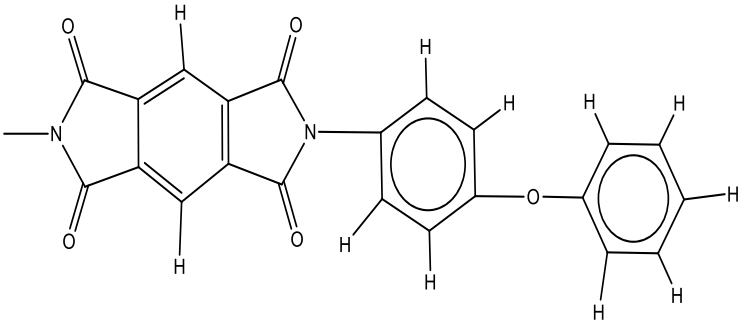
MMM	Map min max
T_g	Température de transition vitreuse
RCP	la régression sur composantes principales
PCA	Principal component analysis (i.e : Analyse en composantes principales)
ERAM	Erreur relative absolue moyenne
R	Coefficient de corrélation
N_2	Gaz d'azote
COV	Composé organique volatil
RDF	Radial destribution function
0D	Zero dimension
1D	Une dimension
2D	Deux dimensions
3D	Trois dimensions
MoRSE	Molecule Representation of Structures based on Electron diffraction
ACP	Analyse en composantes principales
PLS	L'analyse des moindres carrés partiels
AG	Les algorithmes génétiques.
QSAR	Relation quantitative structure-activité
QSPR	Relation quantitative structure-propriété
LME	Laboratoire matériaux et environnement
PRESS	Prédiction Sum of Squares
T	Température
P	Pression
p'	Pressions constantes amont
p''	Pressions constantes aval
e	Epaisseur de la membrane
$S_i(p_i)$	Le coefficient de solubilité de Henry
p_i	Pression partielle du gaz i
c_i	Concentrations limites
J_i	Flux de perméation
D_i	Diffusivité du gaz i
S_i	Solubilité du gaz i
$\overline{S_i}\overline{D_i}$	Coefficient moyen de perméabilité ($\overline{P_i}$)
n	Nombre d'individus
p	Nombre de variables

z_i	Valeur centrée de l'individu i
z_j	Valeur centrée de la variable j
v_k^j	la j-ème coordonnée du vecteur v_k
z^j	la variable initiale
c^k	la nouvelle variable «composante principale »
$r(c^k, z^j)$	le coefficient de corrélation linéaire
b_i	les coefficients de régression
$\hat{\beta}_A^j$	la matrice des coefficients du modèle construit avec les individus restants
MCVEPLS	Monte carlo variable elimination partial least square
PLSUVE	PLS d'élimination de la variable non informative
RMSEP	La racine carrée de l'erreur relative moyenne en prédiction
RSD	Relative standard déviation (déviation relative standard)

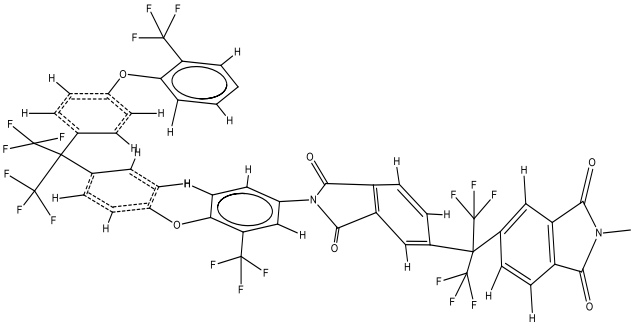
VII. Annexe

Tableau VII.1 : Monomères structurales des PI et PEI étudiés

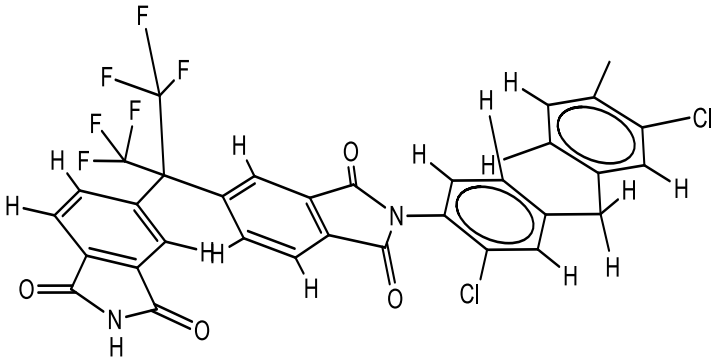
 6FDA_4BDAF	 6FDA_6FmDA
 6FDA_6FpDA	 6FDA_APAP
 6FDA_BAHF	 6FDA_BAPE



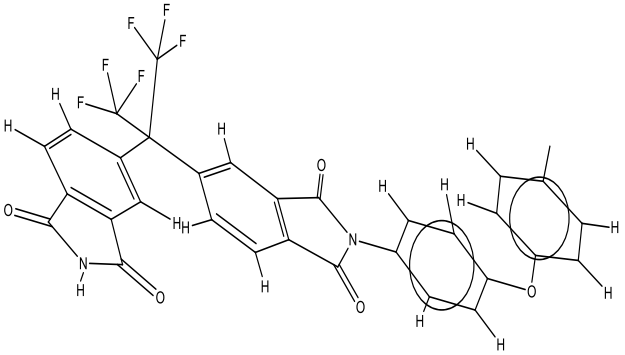
PMDA_pp'_ODA



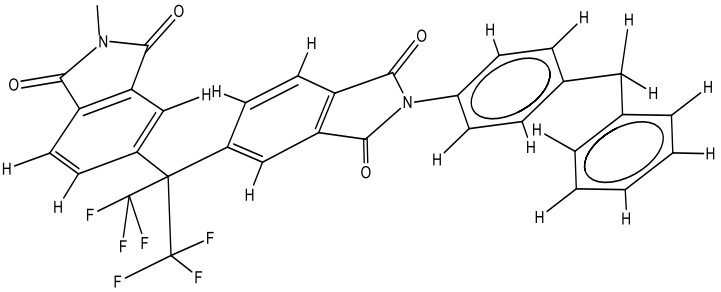
6FDA_BATPHF



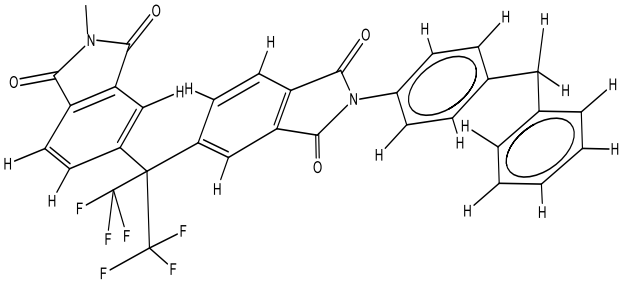
6FDA_CDM



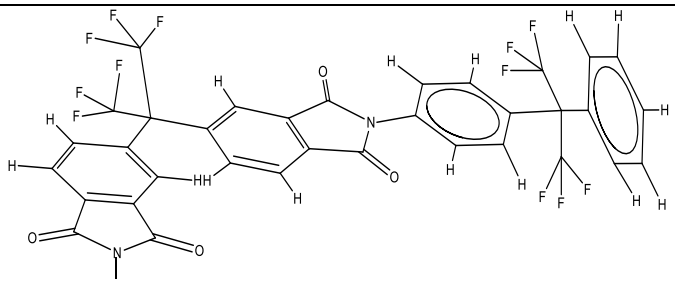
6FDA_DADE



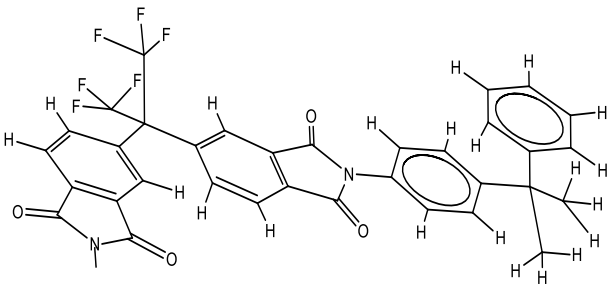
6FDA_DADM



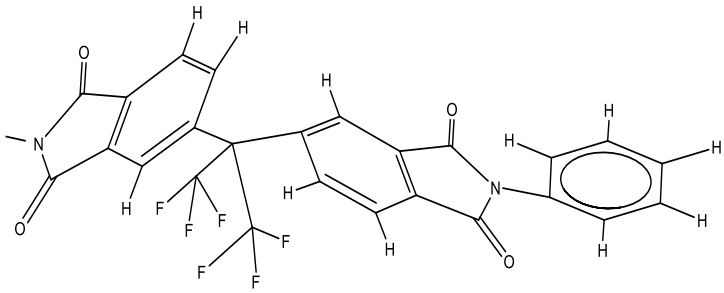
6FDA_HAB



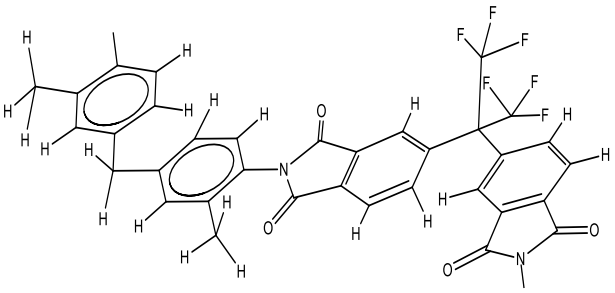
6FDA_HFIP



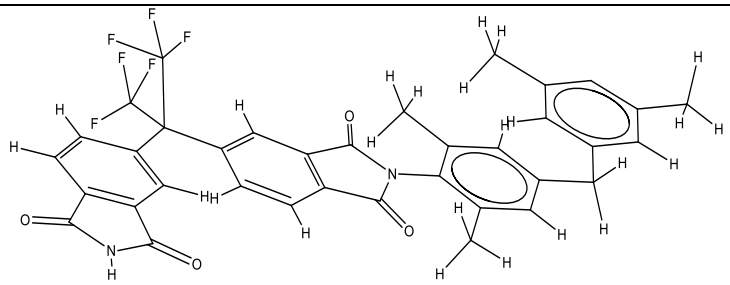
6FDA_IPDA



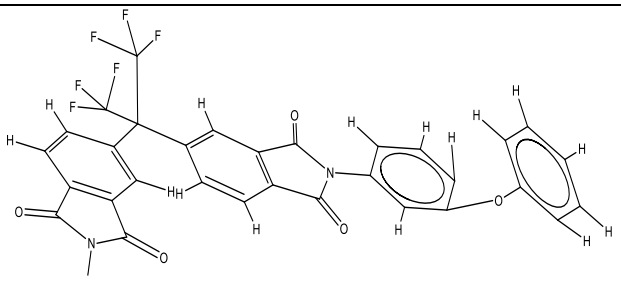
6FDA_m_PDA



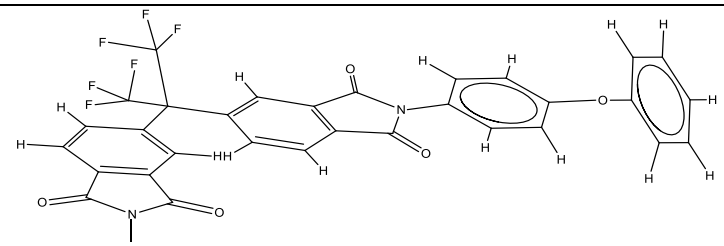
6FDA_MDT



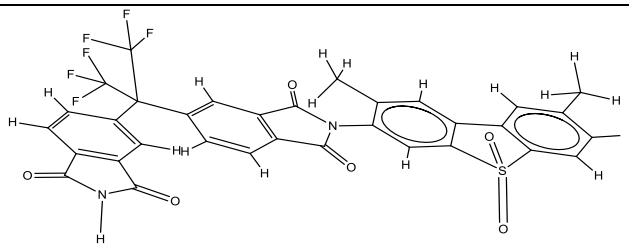
6FDA_MDX



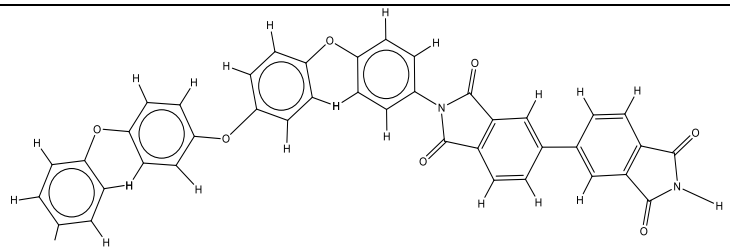
6FDA_mp'ODA



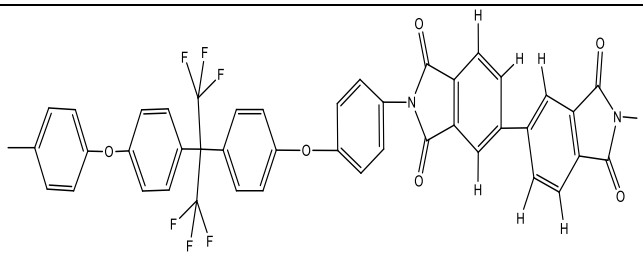
6FDA_ODA



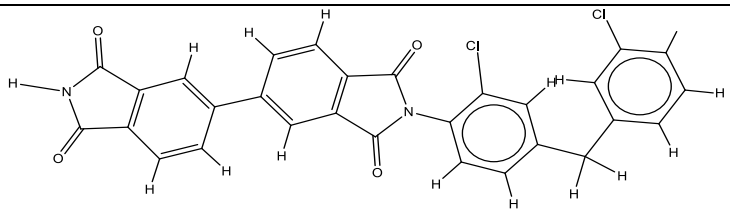
6FDA_TSN



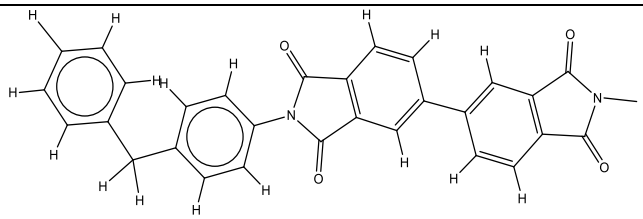
BPDA_BAPE



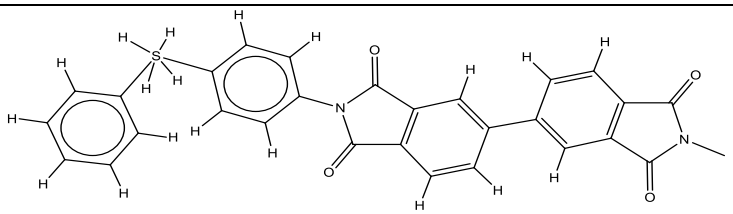
BPDA_BAPHF



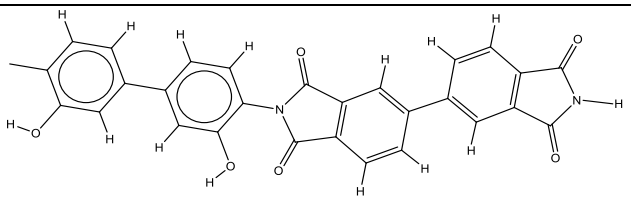
BPDA_CDM



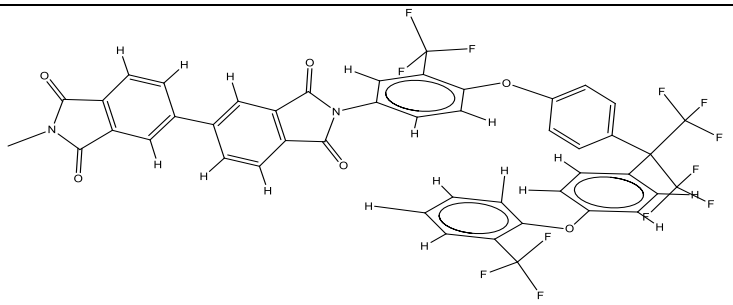
BPDA_DADM



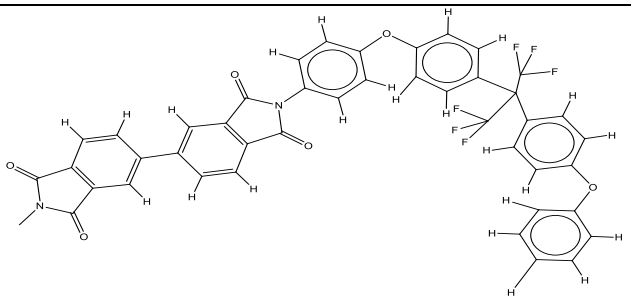
BPDA_DADS



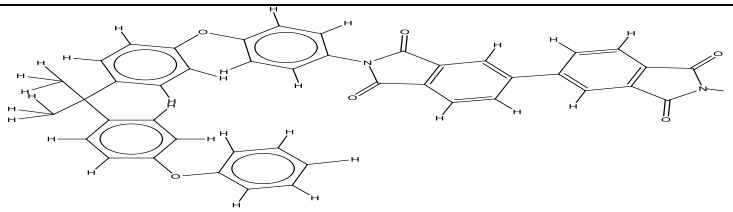
BPDA_HAB



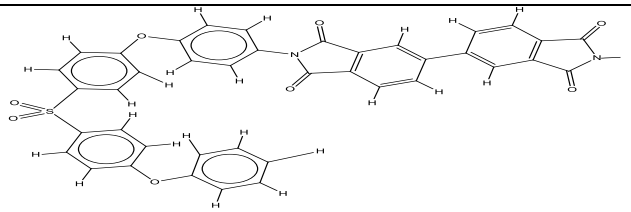
BPDA_BATPHF



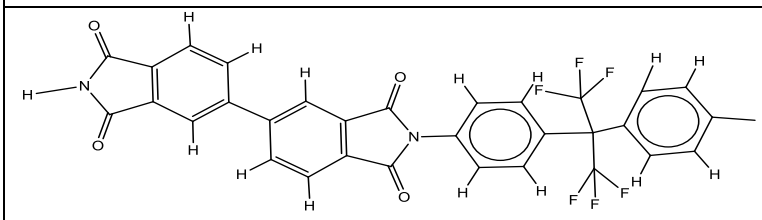
BPDA_HFBAPP



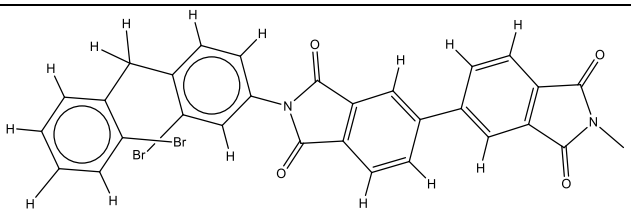
BPDA_BAPP



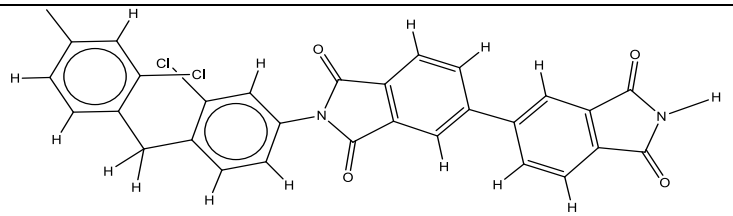
BPDA_BAPS



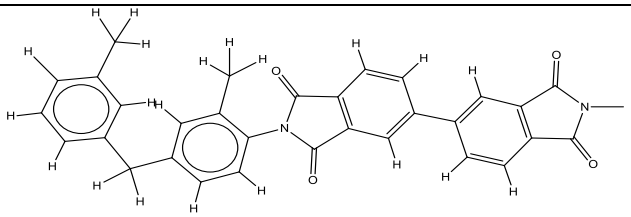
BPDA_HFIP



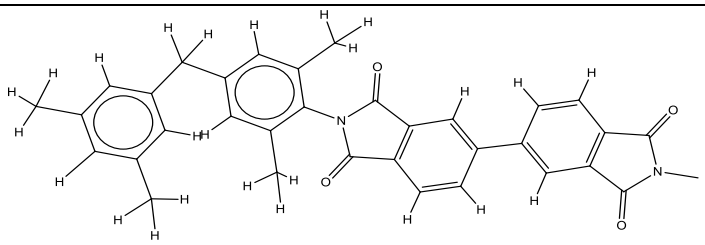
BPDA_MBA



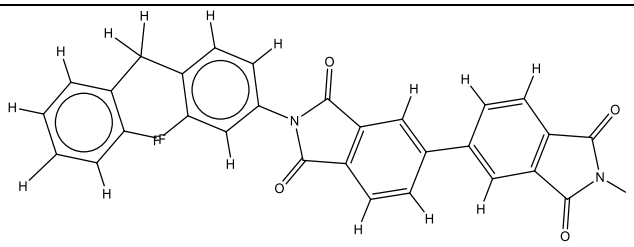
BPDA_MCA



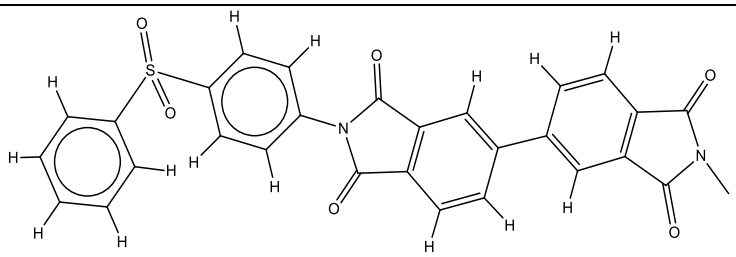
BPDA_MDT



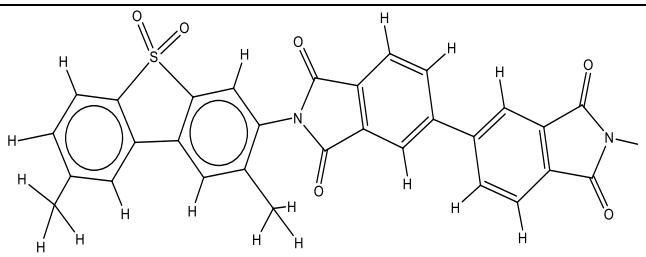
BPDA_MDX



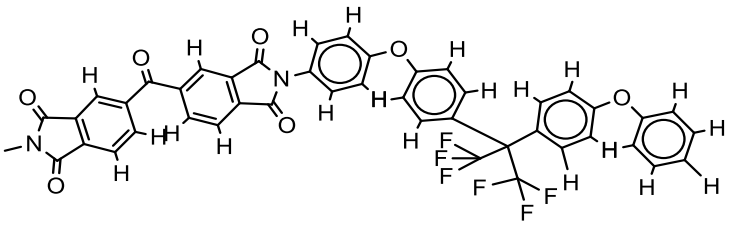
BPDA_MFA



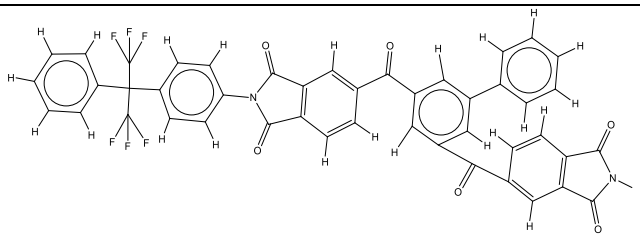
BPDA_PASN



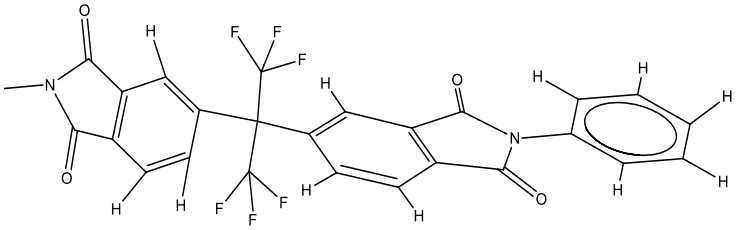
BPDA_TSN



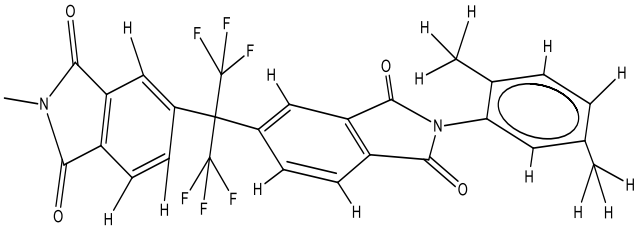
BTDA_BAPHF



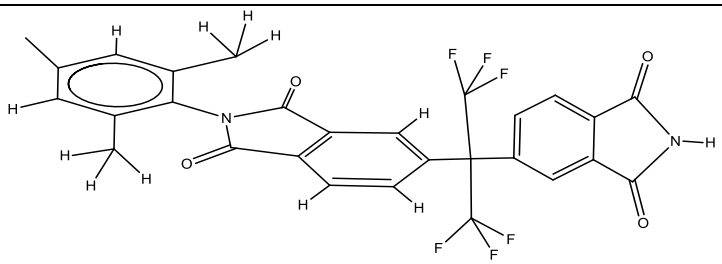
PDCDA_6F



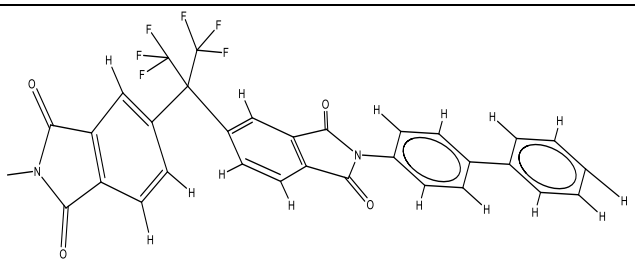
PI_1



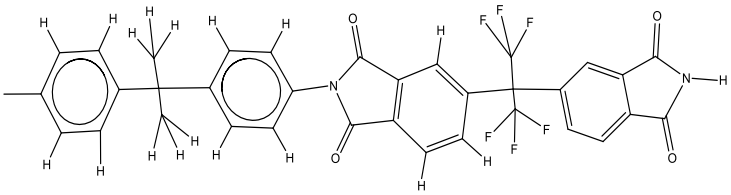
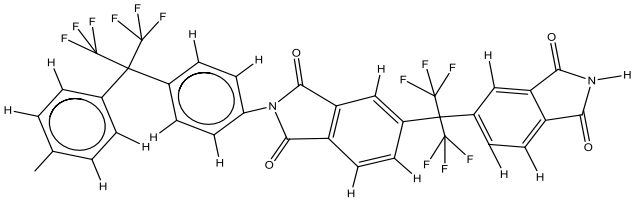
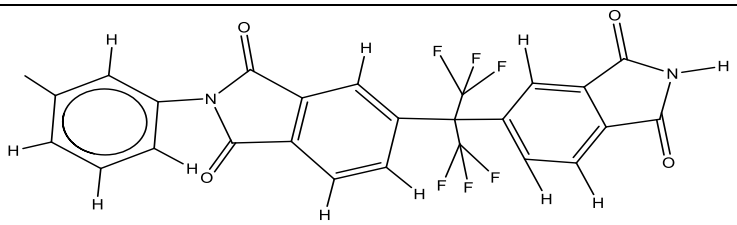
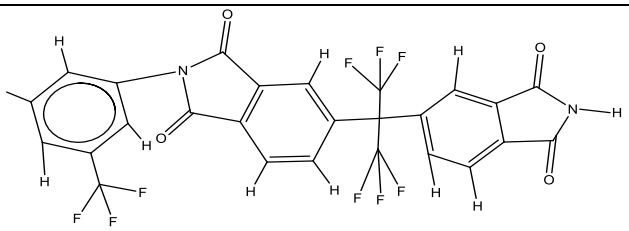
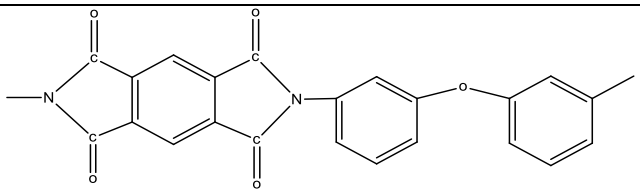
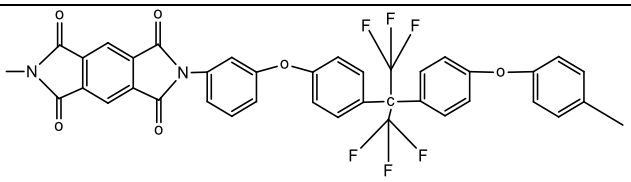
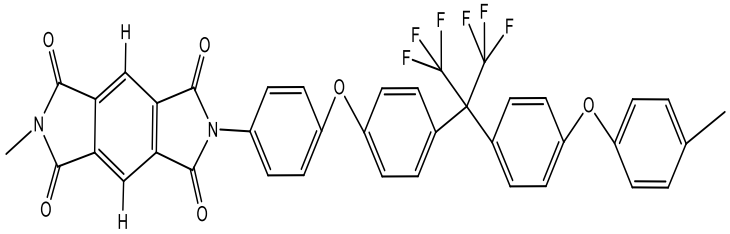
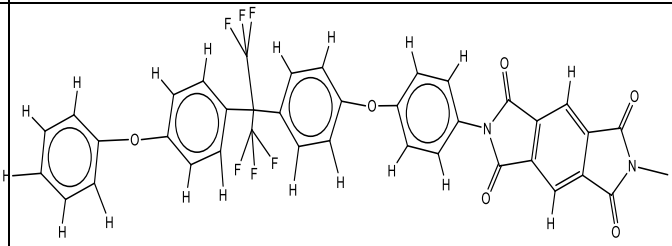
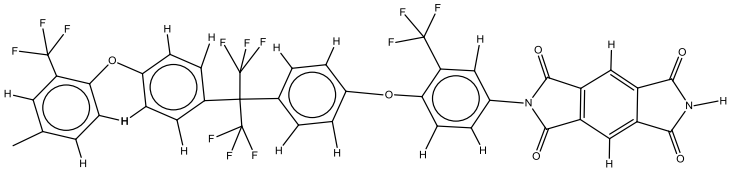
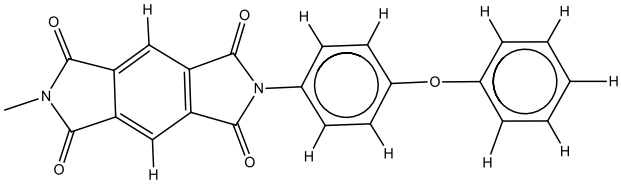
PI_2

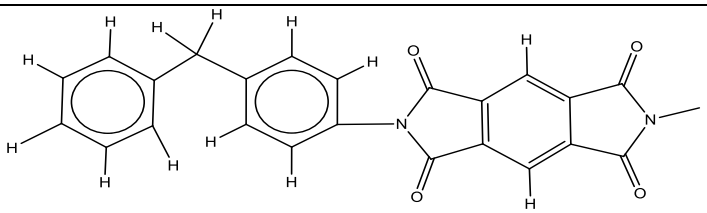


PI_4

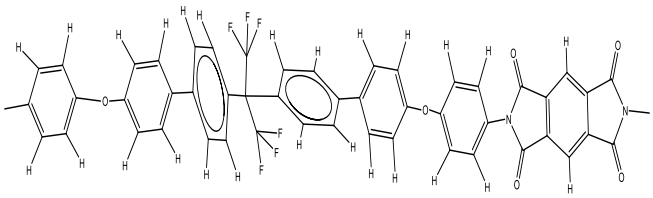


PI_8

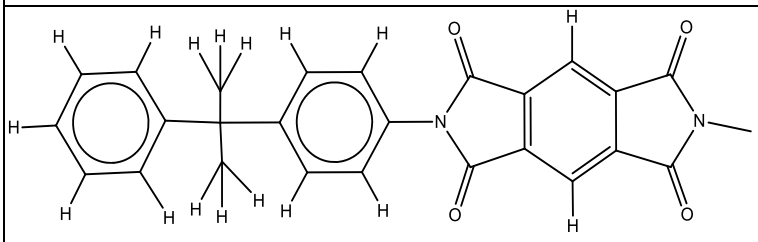
 PI_9	 PI_10
 PI_11	 PI_12
 PMDA_3,3'_ODA	 PMDA_3BDAF
 PMDA_4BDAF	 PMDA_BAPHF
 PMDA_BATPHF	 PMDA_DADE



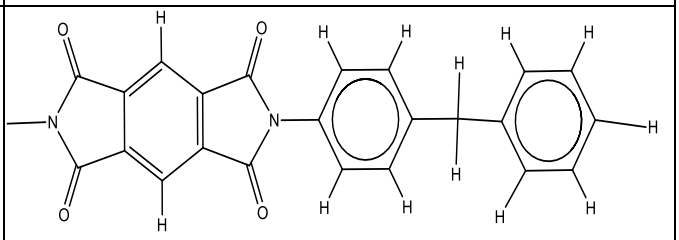
PMDA_DADM



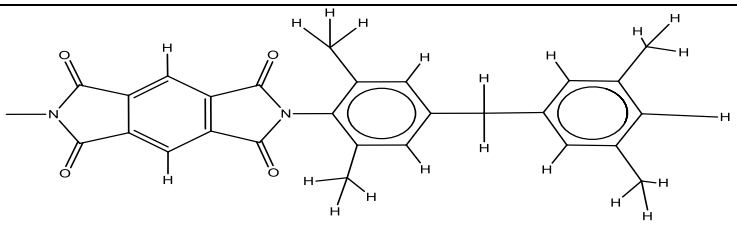
PMDA_HFBAPP



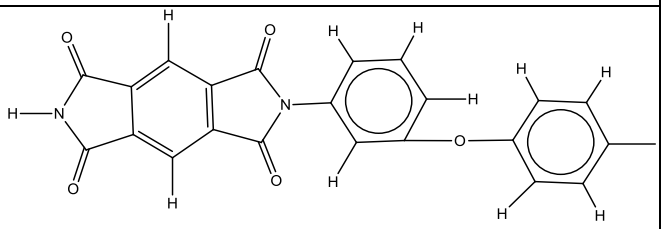
PMDA_IPDA



PMDA_MDA



PMDA_MDX



PMDA_mp'ODA